COMITÊ DE MÁQUINAS

Prof. Dr. Clodoaldo A M Lima

Sumário

- Definições e Motivação
- Ensembles
  - Aspectos Gerais
  - A Questão da Diversidade
  - Etapas de construção
    - Seleção de componentes
    - Combinação de componentes
    - Treinamento e inserção de diversidade
  - Mistura de Especialistas
  - Referências Bibliográficas

Definições Motivações

- Quando decisões importantes devem ser tomadas sobre assuntos de grande impacto em nossas vidas, geralmente recorremos a opiniões e considerações de um *grupo de pessoas que dominem o assunto em questão, visando maximizar a chance de tomarmos uma decisão correta.*

- Essa tendência a organizar comitês para tomar decisões pode ser observada em diversos níveis na sociedade humana, desde em pequenas reuniões familiares para decidir qual o melhor colégio para os filhos, até nas sessões do Congresso Nacional para discutir e votar algum Projeto de Lei

Definições Motivações

- O objetivo principal por trás da formação desses *comitês está em reunir pessoas que tenham um certo domínio do assunto em questão, mas que, ao mesmo tempo, tenham opiniões diversas*.

- Isso permite levantar discussões que contribuam para:
  - A identificação dos *principais pontos positivos e negativos que possam estar associados a cada uma das opções de decisão;*
  - A escolha da *melhor solução para o problema em questão.*

Definições Motivações

- Na área de *aprendizado de máquina*, a *idéia de formação de comitês de indivíduos que tenham um bom conhecimento sobre um problema e ao mesmo tempo tenham "opiniões", em certo grau, distintas dos demais indivíduos no comitê foi adotada nos chamados:*

  - Comitê de Máquinas

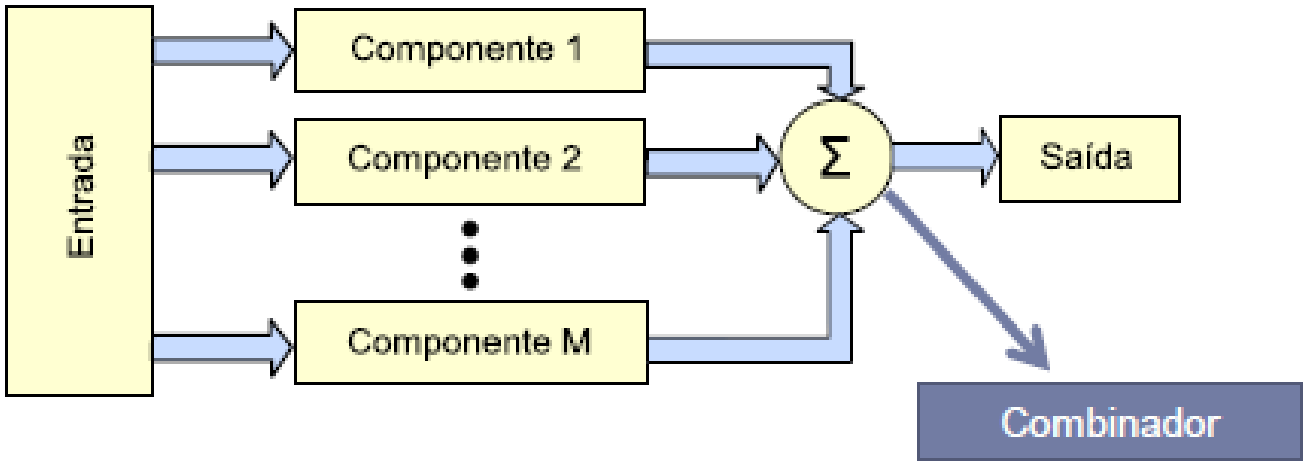
Comitê de Máquinas

- Comitê de máquinas ou Agrupamento de Máquinas
  - Método de aprendizado supervisionado ou não-supervisionado cujo objetivo é aumentar a capacidade de generalização de estimadores (aproximadores de função/regressores, classificadores, etc.)

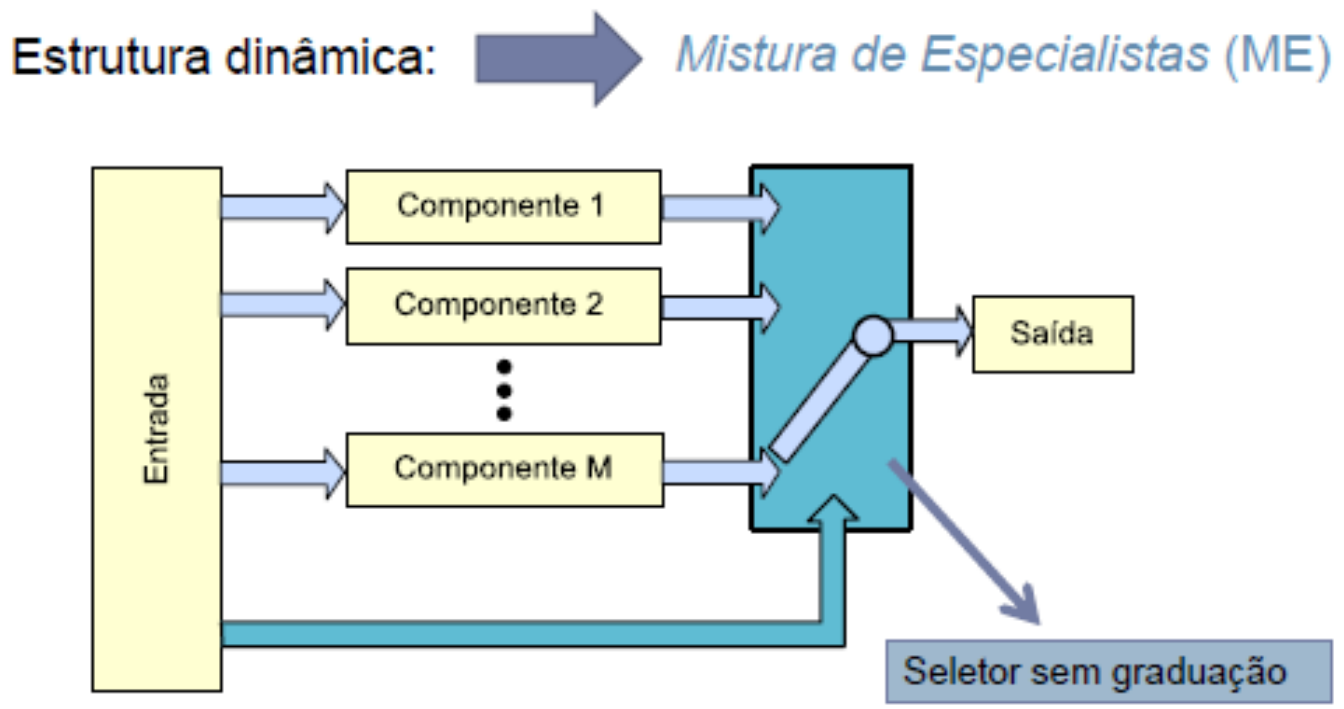
- Categorias de Comitê de Máquinas
  - Estrutura estática
    - ensemble
  - Estrutura dinâmica
    - Mistura de especialistas (ME)
    - Mistura hierárquica de especialistas (HME)

Tipos de Comitês de Máquinas

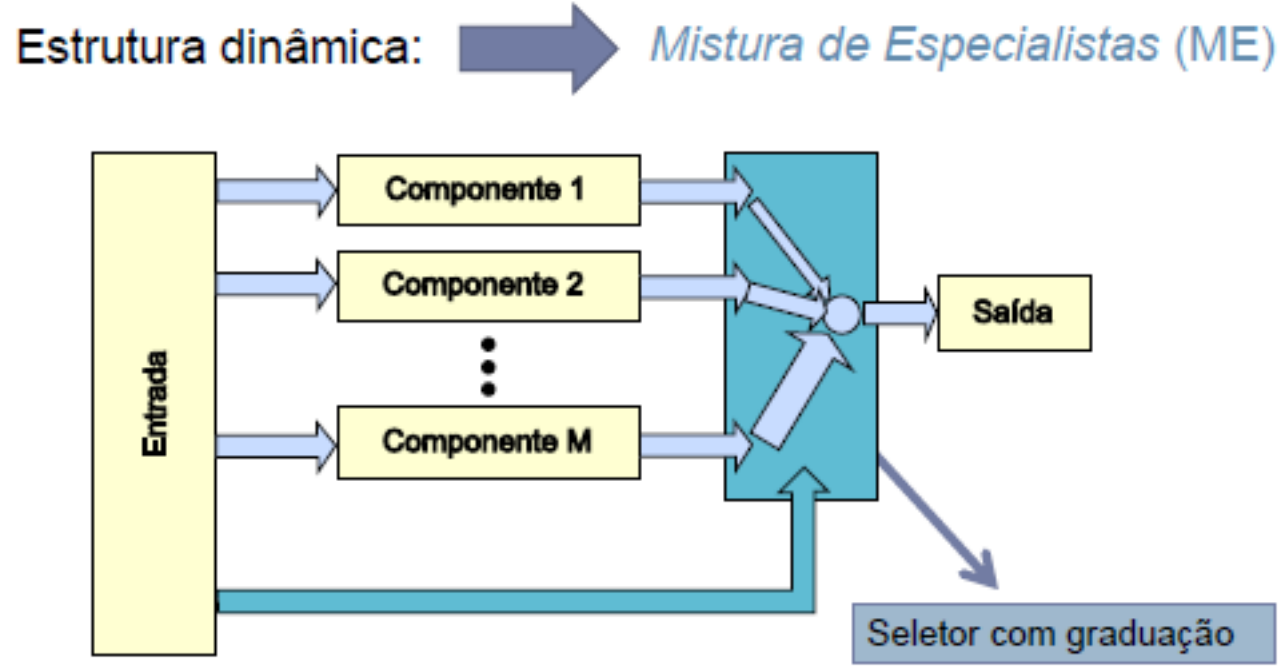
Estrutura estática: → *Ensembles*



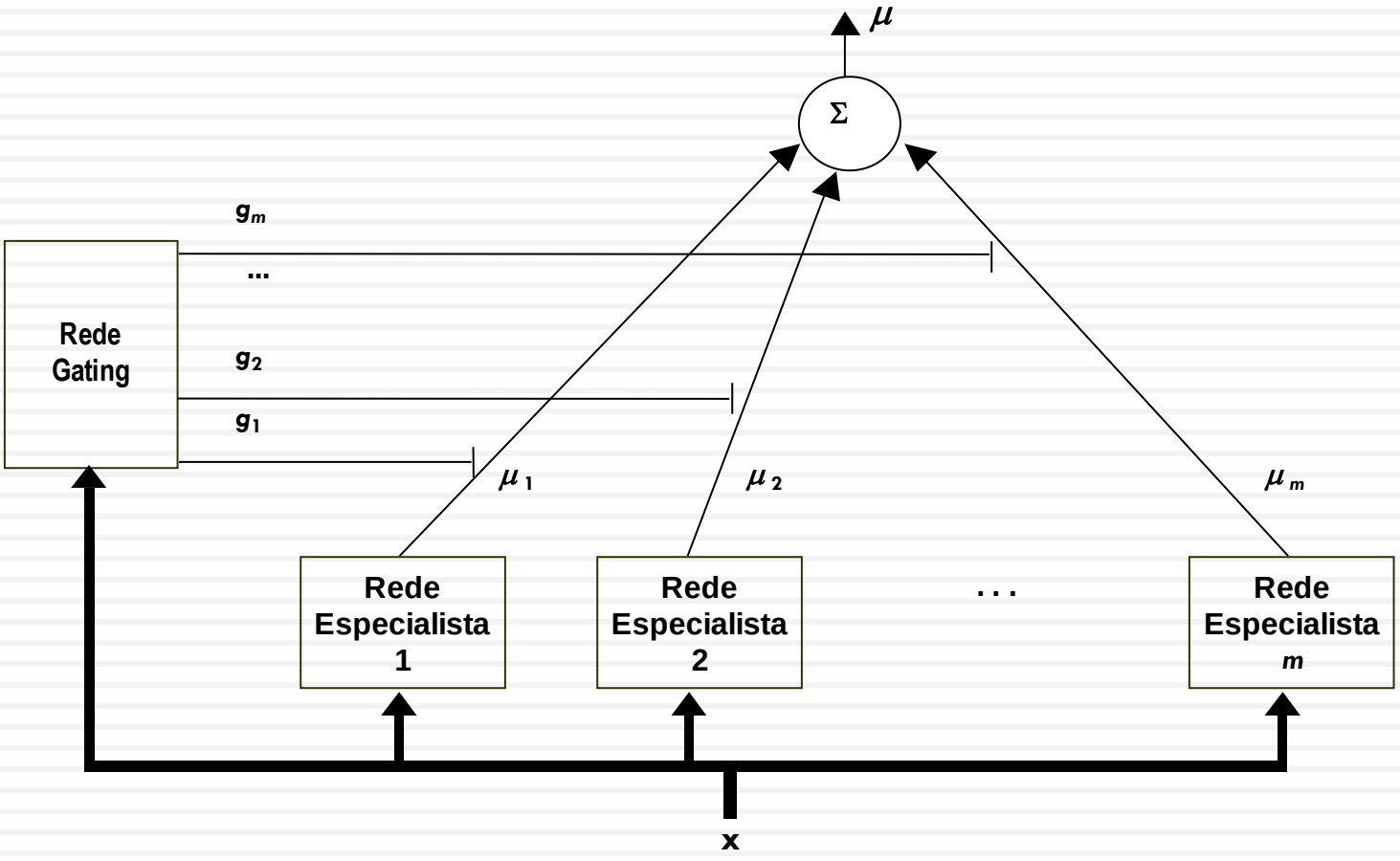
Tipos de Comitês de Máquinas



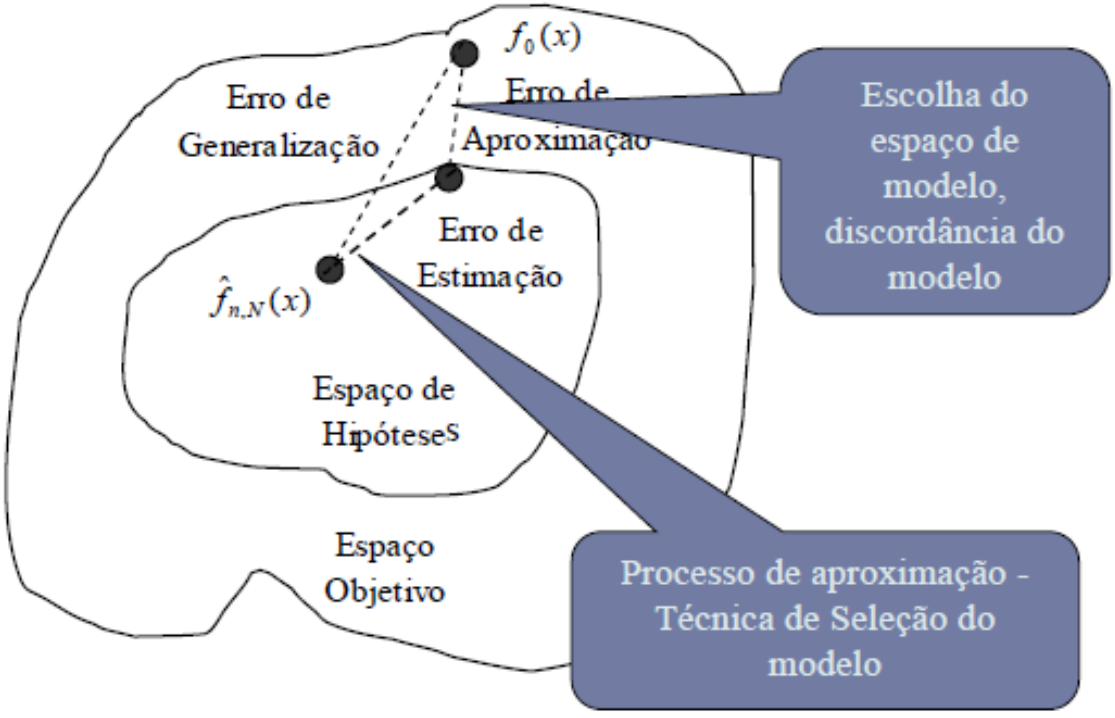
Tipos de Comitês de Máquinas



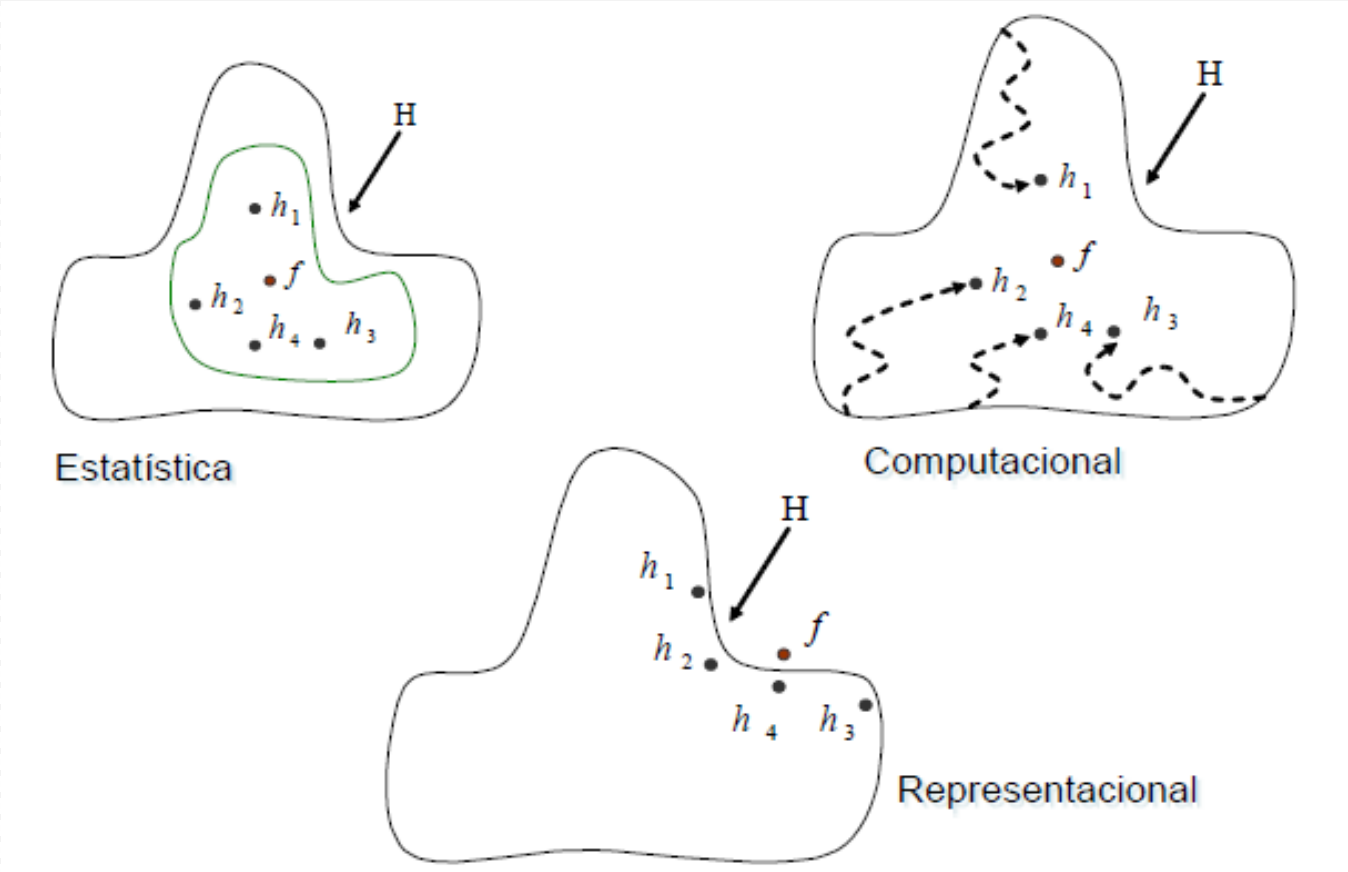
Estrutura de uma Mistura de Especialista



Erro de Generalização



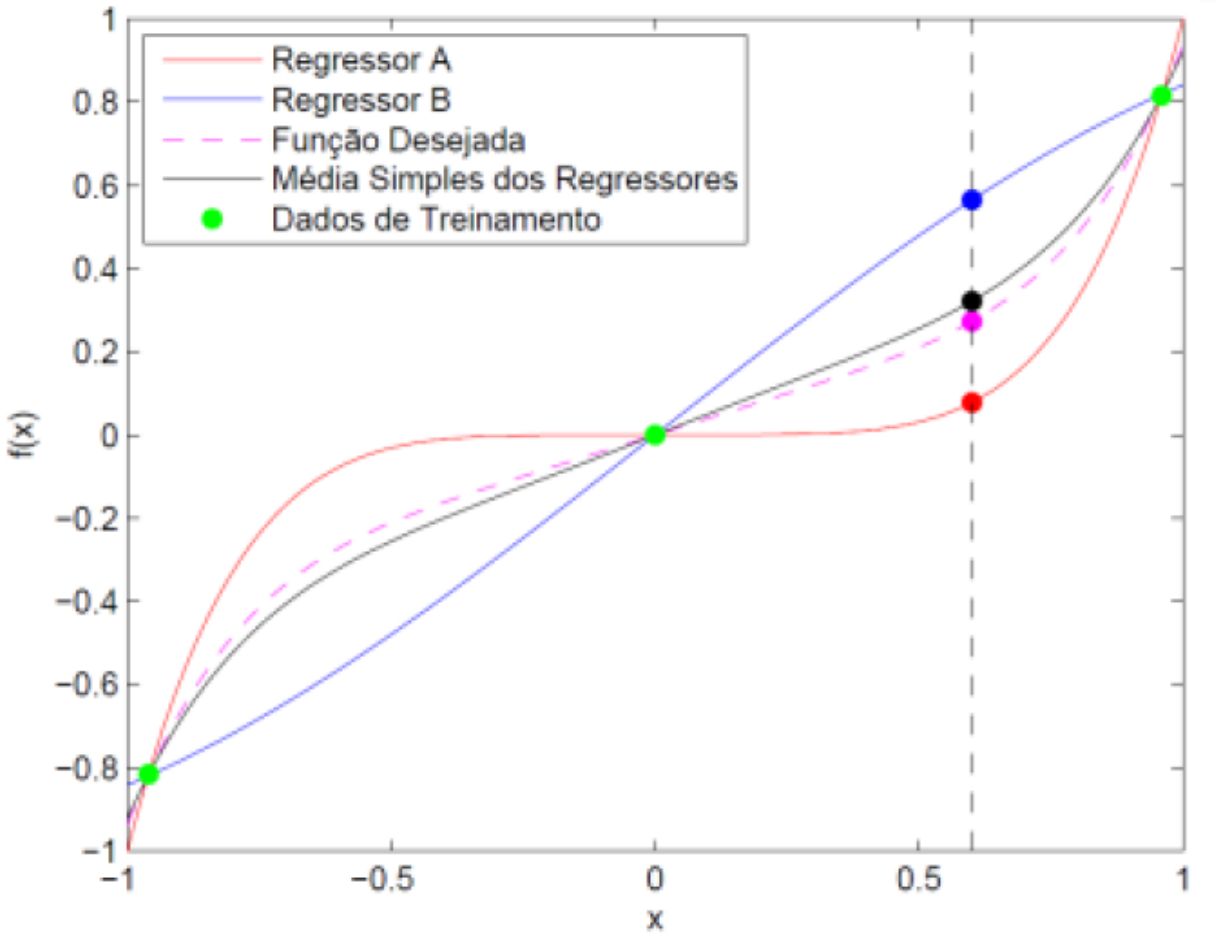
Três formas de reduzir o erro de generalização



Exemplo: Aproximação de funções

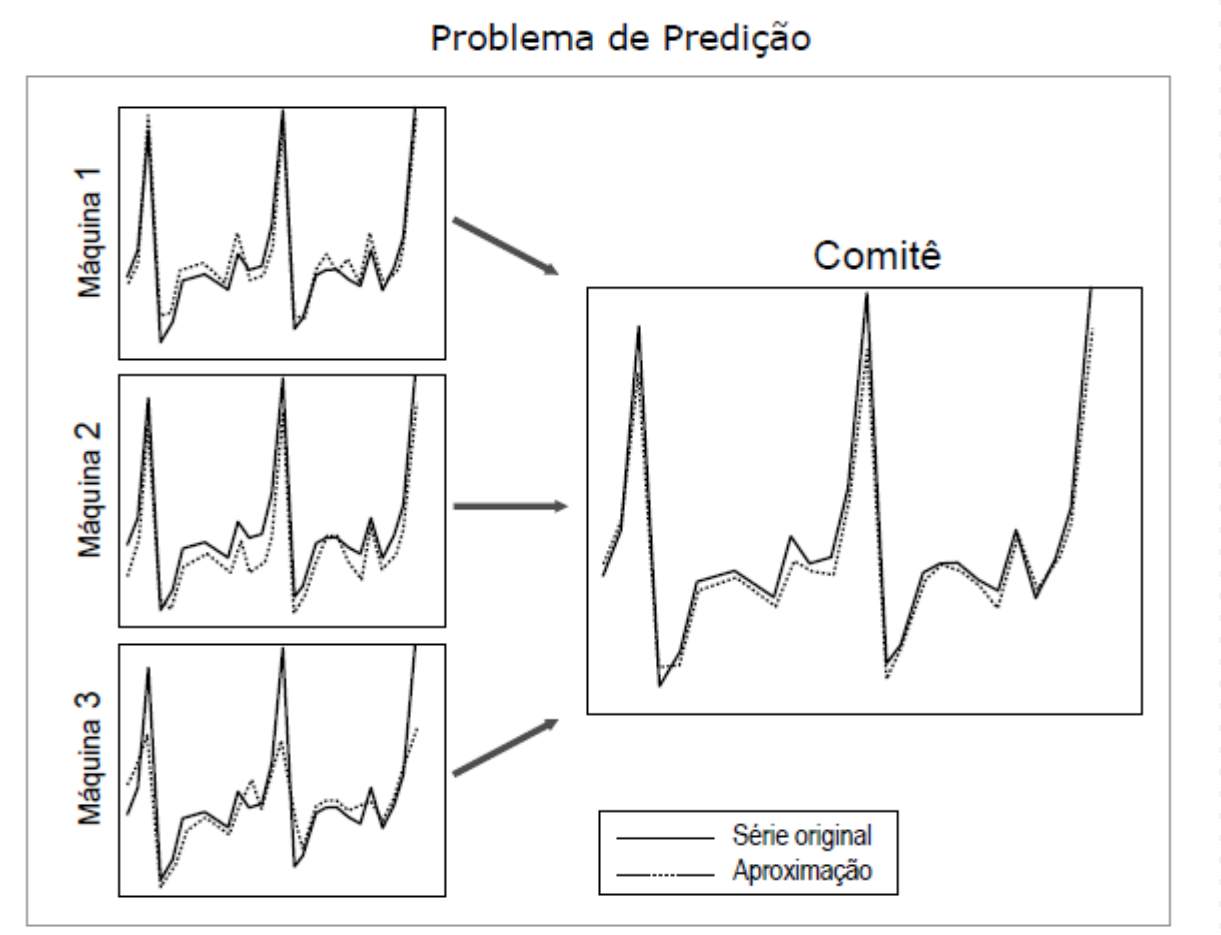
- Combinação por média simples

Ensembles:



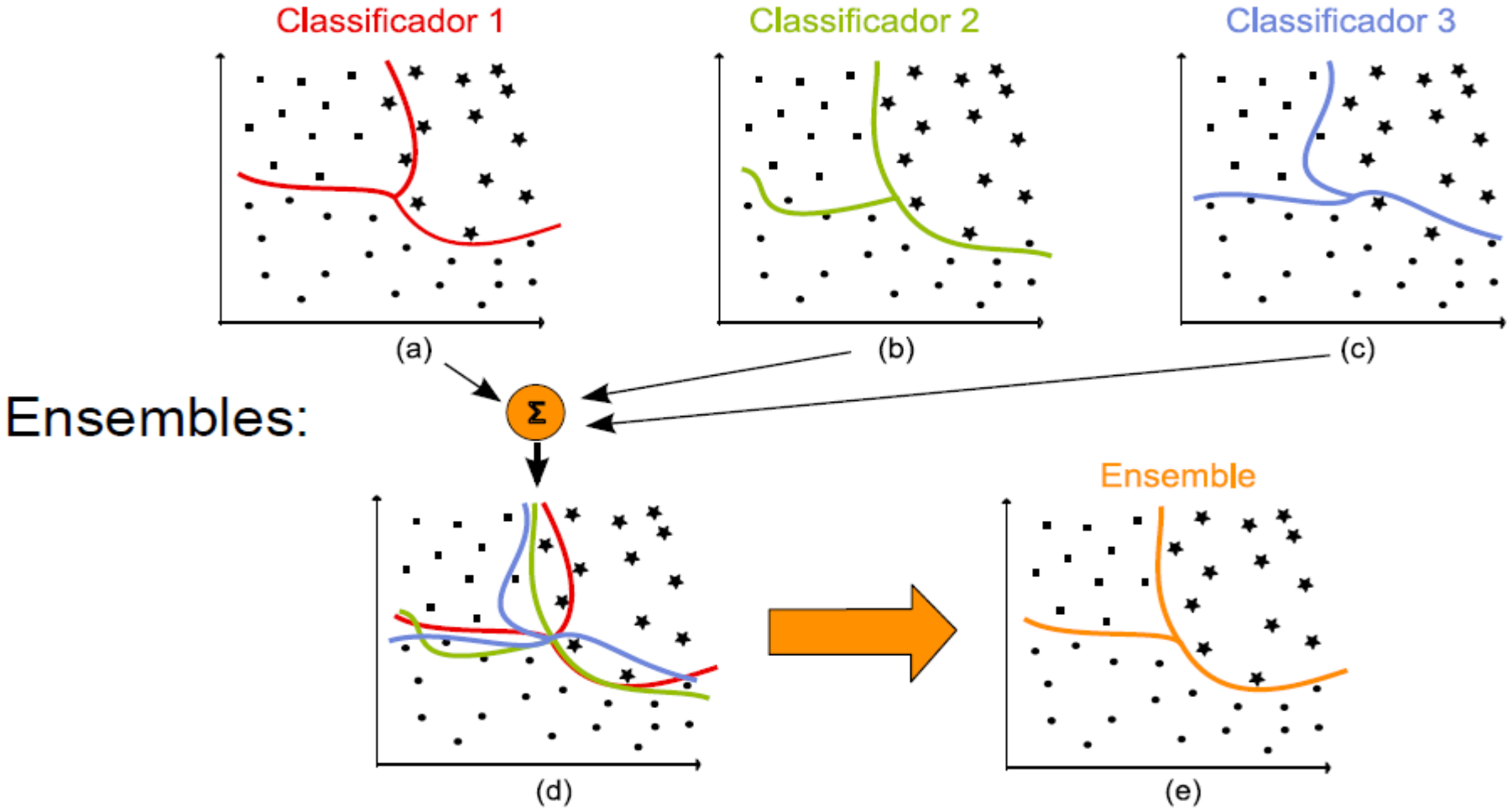
Exemplo: Predição de Séries Temporais

- Combinação por média simples

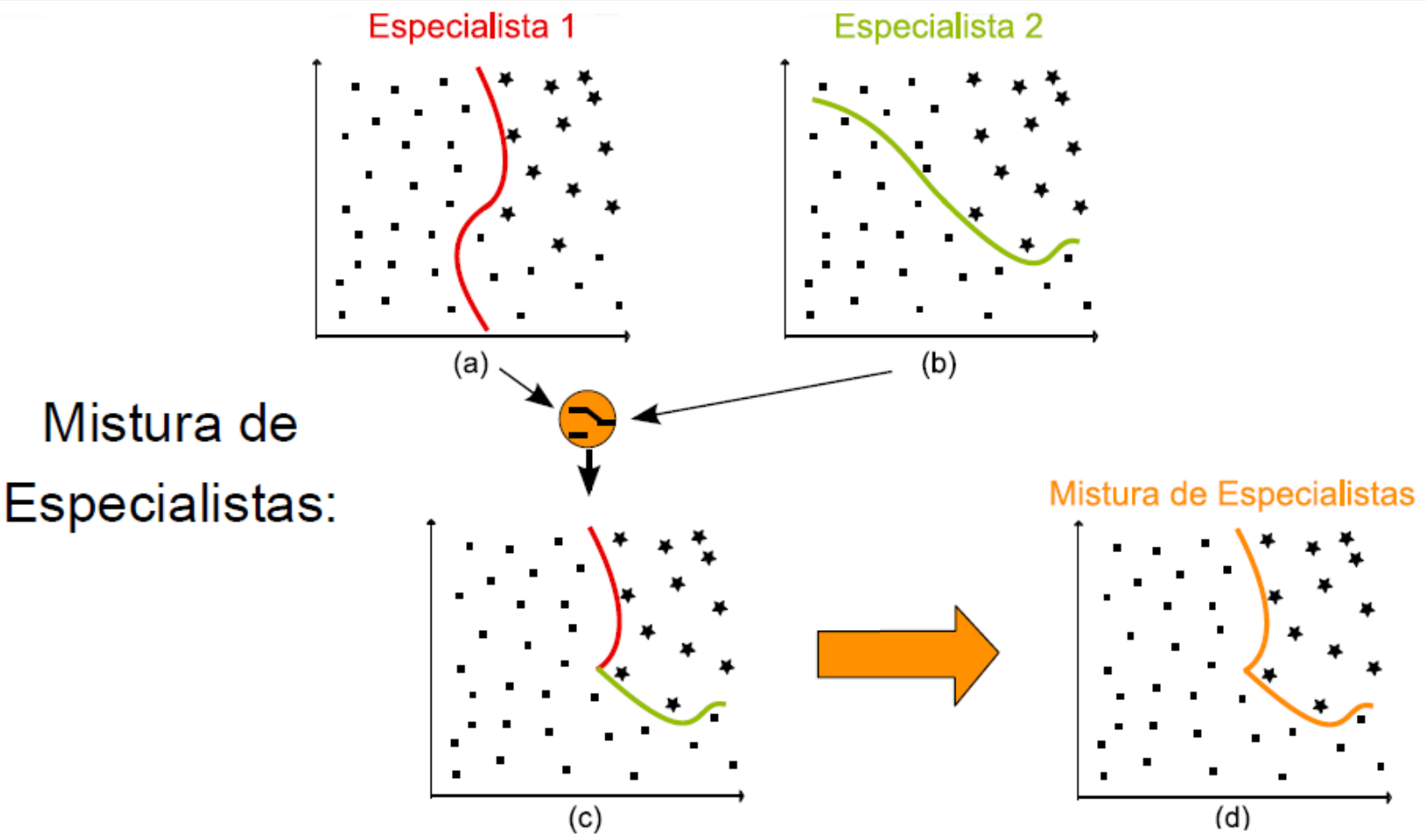


Exemplo: Classificação de padrões

- Combinação por voto majoritário

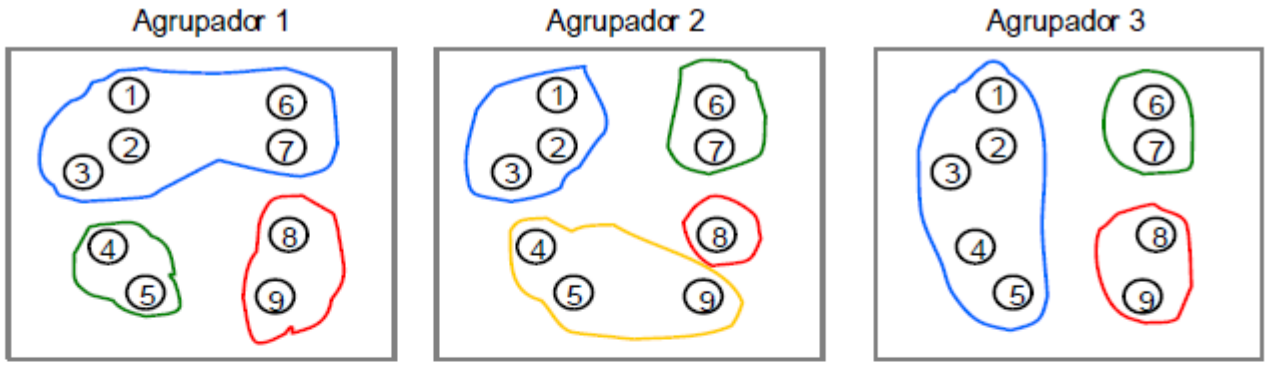


Exemplo: Classificação de padrões

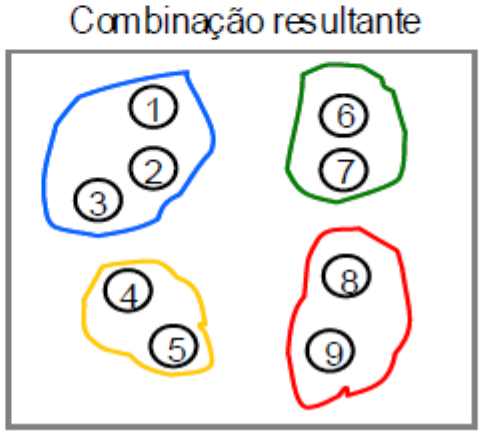


- ME também podem ser aplicadas a problemas de regressão.

Exemplo: Agrupamento de dados (clusterização)



Ensembles:

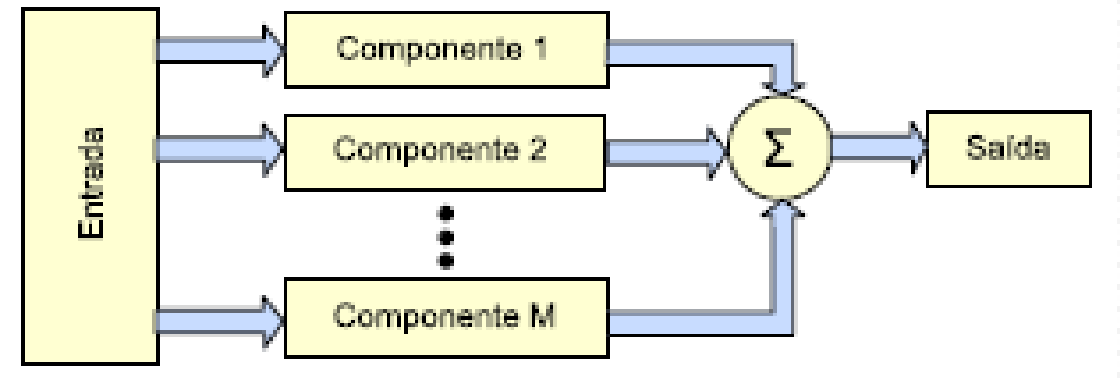


Comitê de Máquina

Ensemble

Ensembles

- Ensemble é um paradigma de aprendizado em que propostas alternativas de solução para um problema, denominadas componentes, têm suas saídas individuais combinadas na obtenção de uma solução final.



- Em um ensemble, os *M componentes recebem os mesmos dados de entrada e geram resultados individuais, que são combinados (Σ) em uma única saída.*

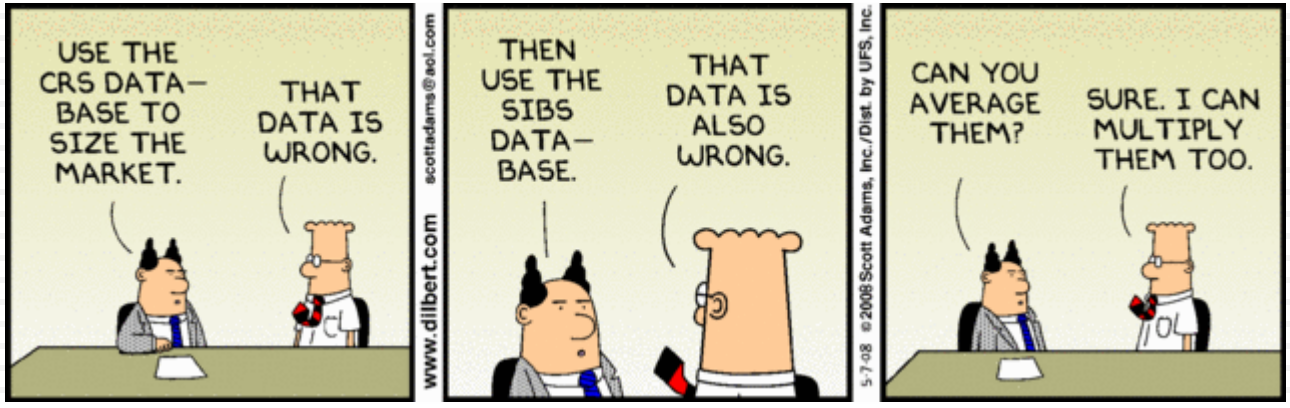
Ensembles

- Intuitivamente, a combinação de múltiplos componentes é vantajosa, uma vez que componentes diferentes podem implicitamente representar aspectos distintos e, ao mesmo tempo, relevantes para a solução de um dado problema.

- A abordagem de ensembles tem sido amplamente utilizada nas últimas duas décadas, tanto para problemas de regressão quanto para problemas de classificação de padrões, já que os ensembles são comprovadamente capazes de aumentar a capacidade de generalização e, consequentemente, o desempenho geral do sistema (Hansen & Salamon, 1990; Hashem et al., 1994).

Ensembles

- No entanto, tal melhora na capacidade de generalização se apóia na *qualidade de seus componentes* e na *diversidade do erro apresentada por eles (Perrone & Cooper, 1993):*
  - Cada um dos componentes em um ensemble deve apresentar um *bom desempenho quando aplicado isoladamente ao problema* e, ao mesmo tempo, deve *"cometer erros" distintos quando comparados aos demais componentes.*



Ensemble

- Essa necessidade de diversidade do erro dos componentes é, de certa forma, intuitiva;
- Se forem combinados componentes que apresentam um mesmo padrão de erro, claramente não haverá nenhum incremento de desempenho:
  - Erros iguais para um mesmo subconjunto de estímulos de entrada implica em acertos também coincidentes;
- Esta combinação trará apenas um aumento no custo computacional, sem resultados práticos de desempenho.

Ensembles: a Questão da Diversidade

- Os estudos iniciais sobre a combinação de componentes para problemas de *regressão foram feitos paralelamente por Perrone (1993) e Hashem (1993; 1997), e tornaram-se um tópico intensamente investigado nos anos subsequentes.*
- Esse interesse acabou por contribuir muito para o amadurecimento do conceito de *diversidade de erros em regressores, levando ao desenvolvimento das seguintes teorias (Brown et al., 2005):*
  - Ambiguity Decomposition, proposta por Krogh & Vedelsby (1995),
  - Bias-Variance-Covariance Decomposition, proposta por Ueda & Nakano (1996).

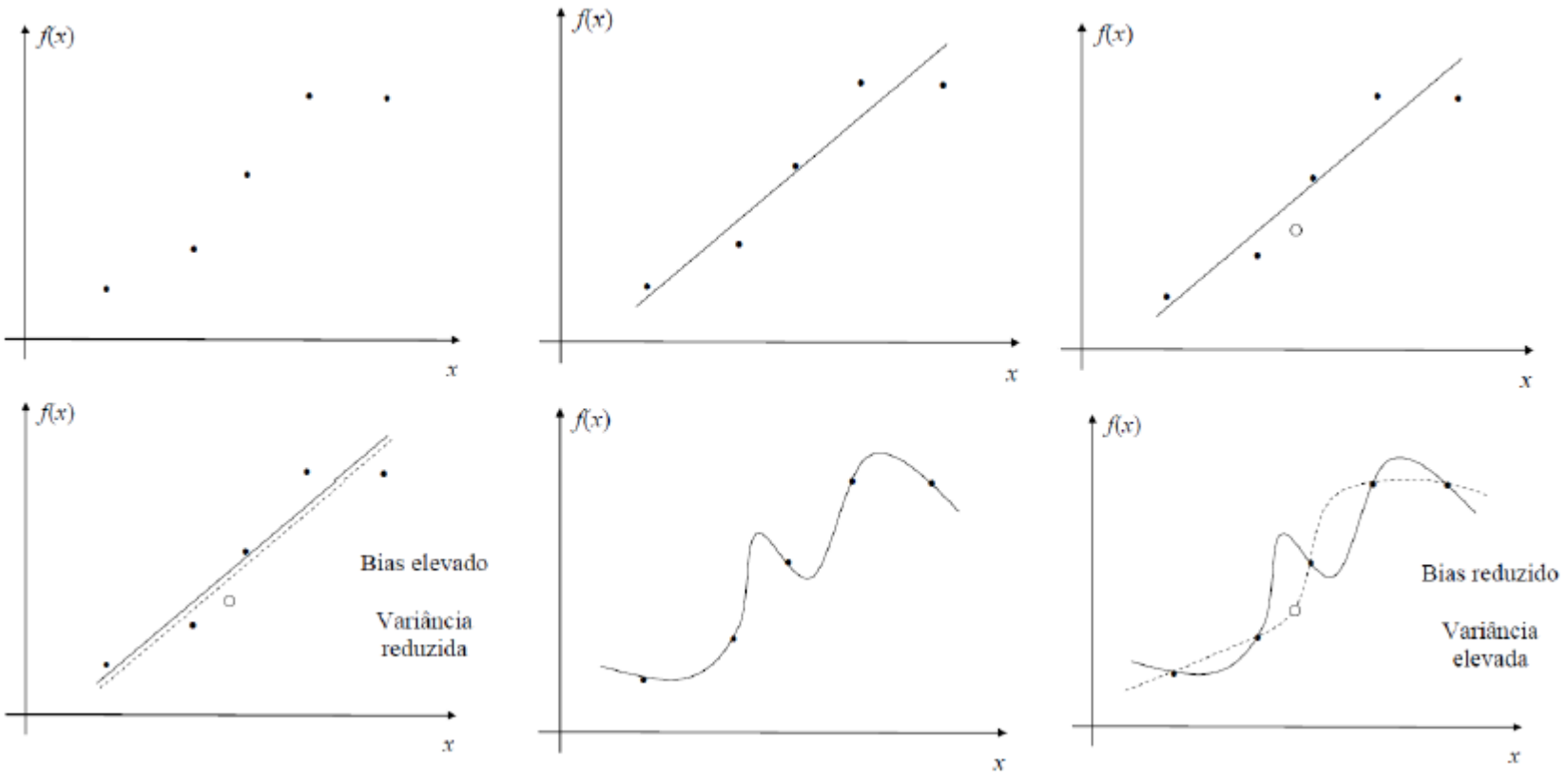
Ensembles: a Questão da Diversidade

- De modo geral, estas teorias mostram que o *erro quadrático médio de um ensemble é criticamente dependente da correlação entre os erros de cada componente*;
- **Decomposição *Bias-Variância de estimadores individuais*:**
  - Propõe que o erro de generalização de um estimador pode ser dividido em dois componentes: *bias* e *variância*;
  - **Bias:** pode ser definido como uma medida do quão perto um dado estimador está do seu alvo (em média, tomada em diferentes conjuntos de treinamento);
  - **Variância**: é uma medida do quão estável uma dada solução é.

Ensembles: a Questão da Diversidade

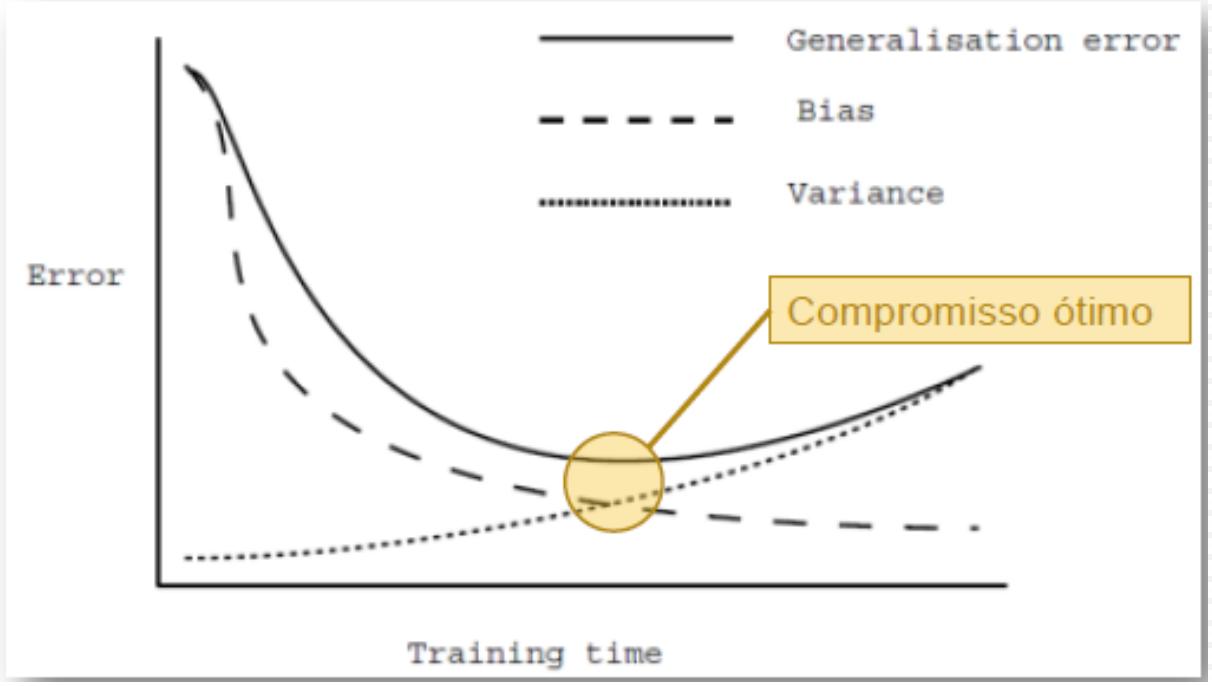
- Relembrando (Aula 1)

$$erro_{true}(h) \leq \underbrace{erro_D(h)}_{bias} + \underbrace{\sqrt{\frac{\ln|\mathcal{H}| + \ln(1/\delta)}{2m}}}_{variance}$$



Ensembles: a Questão da Diversidade

- Estes dois componentes são conflitantes:
  - Tentativas de redução de *bias levam a aumentos na variância e vice-versa;*



Ensembles: a Questão da Diversidade

- **Ambiguity Decomposition (Krogh & Vedelsby, 1995):**
  - Considere um ensemble com combinação convexa de seus componentes, ou seja:

$$f_{ens} = \sum_i w_i f_i$$

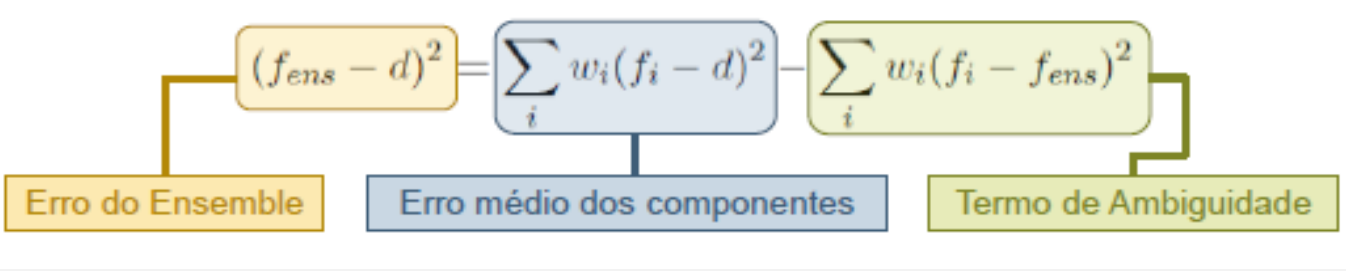
    - onde fi é a saída do i-ésimo componente e fens a saída do ensemble
  - Para uma dada amostra dos dados, a teoria de *Ambiguity Decomposition mostra que (d é o valor desejado):*

$$(f_{ens} - d)^2 = \sum_i w_i (f_i - d)^2 - \sum_i w_i (f_i - f_{ens})^2$$

  - ou seja, o erro quadrático do ensemble é garantidamente menor ou igual ao erro quadrático médio de seus componentes.

Ensembles: a Questão da Diversidade

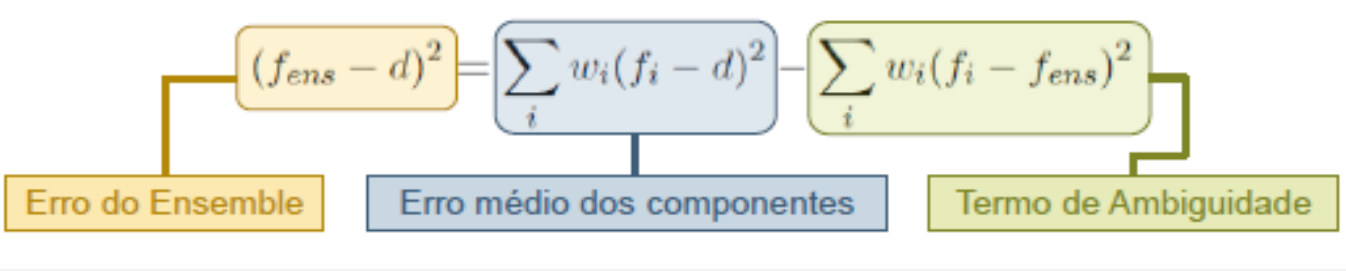
- **Ambiguity Decomposition (Krogh & Vedelsby, 1995):**



- Termo de Ambiguidade:
  - Mede a variabilidade das saídas dos componentes do ensemble, para a amostra em questão;
  - É sempre não-negativo;
  - Quanto maior o seu valor, maior a redução do erro do ensemble;

Ensembles: a Questão da Diversidade

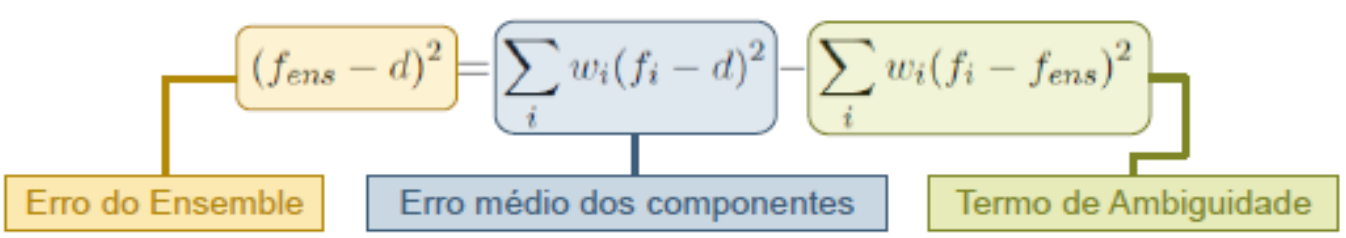
- **Ambiguity Decomposition (Krogh & Vedelsby, 1995):**



- Termo de Ambiguidade:
  - No entanto, o aumento da variabilidade dos indivíduos também implica no aumento do seu erro médio (decomposição bias-variância);
  - A diversidade sozinha não é suficiente: é preciso encontrar o equilíbrio ótimo entre diversidade e acurácia individual para que se tenha o menor erro no ensemble.

Ensembles: a Questão da Diversidade

- **Ambiguity Decomposition (Krogh & Vedelsby, 1995):**

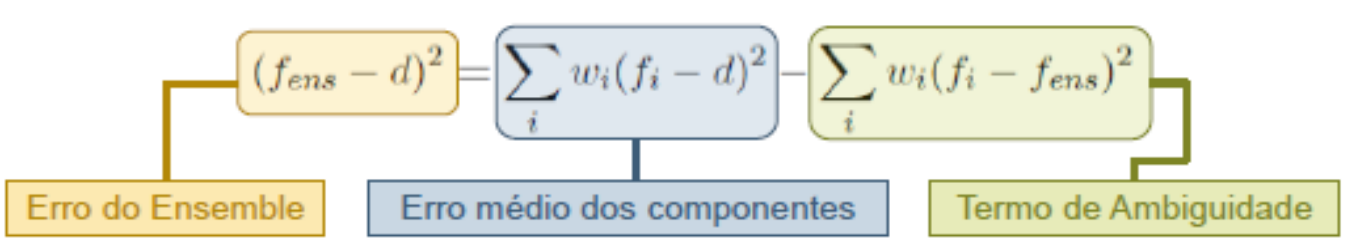


- Pergunta: Certamente eu tenho componentes que, individualmente, apresentam erros menores que a média para alguma amostra! E eles podem até mesmo ser melhores que o ensemble! Por que não usar estes componentes?

- Mas como selecionar estes componentes?
- **Não é possível escolhê-los de antemão.**

Ensembles: a Questão da Diversidade

- **Ambiguity Decomposition (Krogh & Vedelsby, 1995):**



- A teoria de *Ambiguity Decomposition* nos diz que a *combinação de múltiplos preditores é, na média, melhor que a seleção aleatória de preditores individuais*

Ensembles: a Questão da Diversidade

- A teoria de *Ambiguity Decomposition é válida para combinações convexas de componentes em ensembles treinados em um único conjunto de dados;*

- Não leva em conta as possíveis distribuições de probabilidade dos conjuntos de treinamento e nem das possíveis inicializações de pesos (no caso de redes neurais);

- Para atender a estes aspectos, Ueda & Nakano (1996) propuseram a teoria de decomposição em *Bias-Variância-Covariância:*
  - Uma análise detalhada desta teoria pode ser encontrada em Brown (2004), bem como uma comparação com a teoria de *Ambiguity Decomposition*.

Ensembles: a Questão da Diversidade

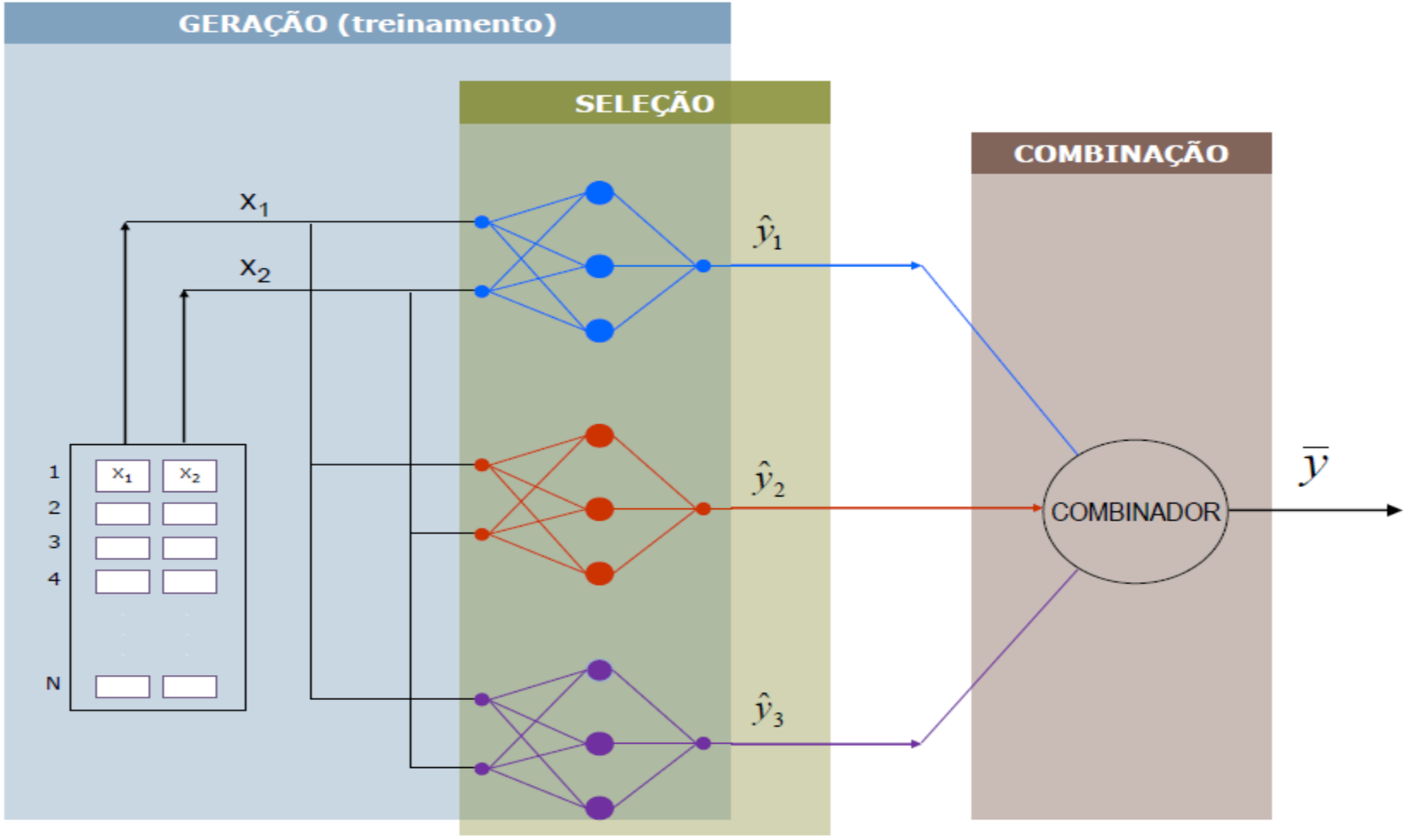
- No contexto de classificação de padrões, apesar da diversidade ser reconhecidamente importante, ainda não existe uma definição rigorosa sobre como as diferenças entre as saídas de cada componente contribuem para sua acurácia;
- Não existe uma definição fechada para o conceito de diversidade entre classificadores;
- Para o caso de dois classificadores, é possível derivar diversas expressões heurísticas a partir da literatura de estatística mas, diante da falta de uma definição exata, não existem análises formais e rigorosas sobre o tema;
- A situação é ainda pior quando se tem mais de dois classificadores.

Ensembles: a Questão da Diversidade

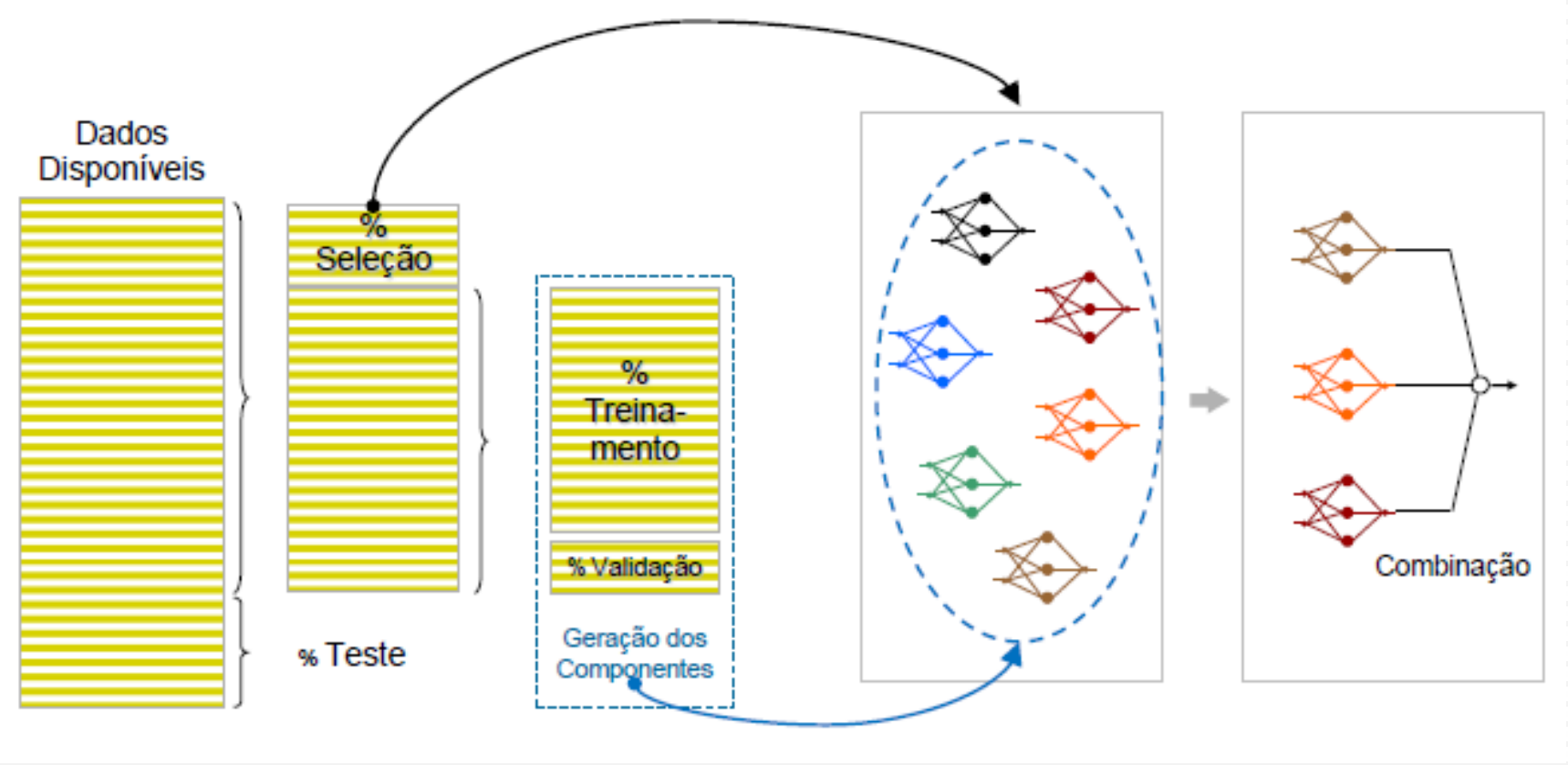
- Tais *expressões heurísticas podem ser divididas em duas categorias (Kuncheva & Whitaker, 2003):*
  - Aquelas que consistem em tomar a média de uma dada métrica de distância entre todos os classificadores do ensemble (medidas do tipo *par-a-par*):
    - Estatística Q;
    - Coeficiente de Correlação (ρ);
    - Métrica de não-concordância (*disagreement metric*);
    - Medida de dupla-falta; etc...
  - Métricas que se baseiam em *entropia ou na correlação de cada classificador com a saída média dos classificadores*:
    - Entropia;
    - Diversidade Generalizada; etc...

Ensembles: Etapas de Construção

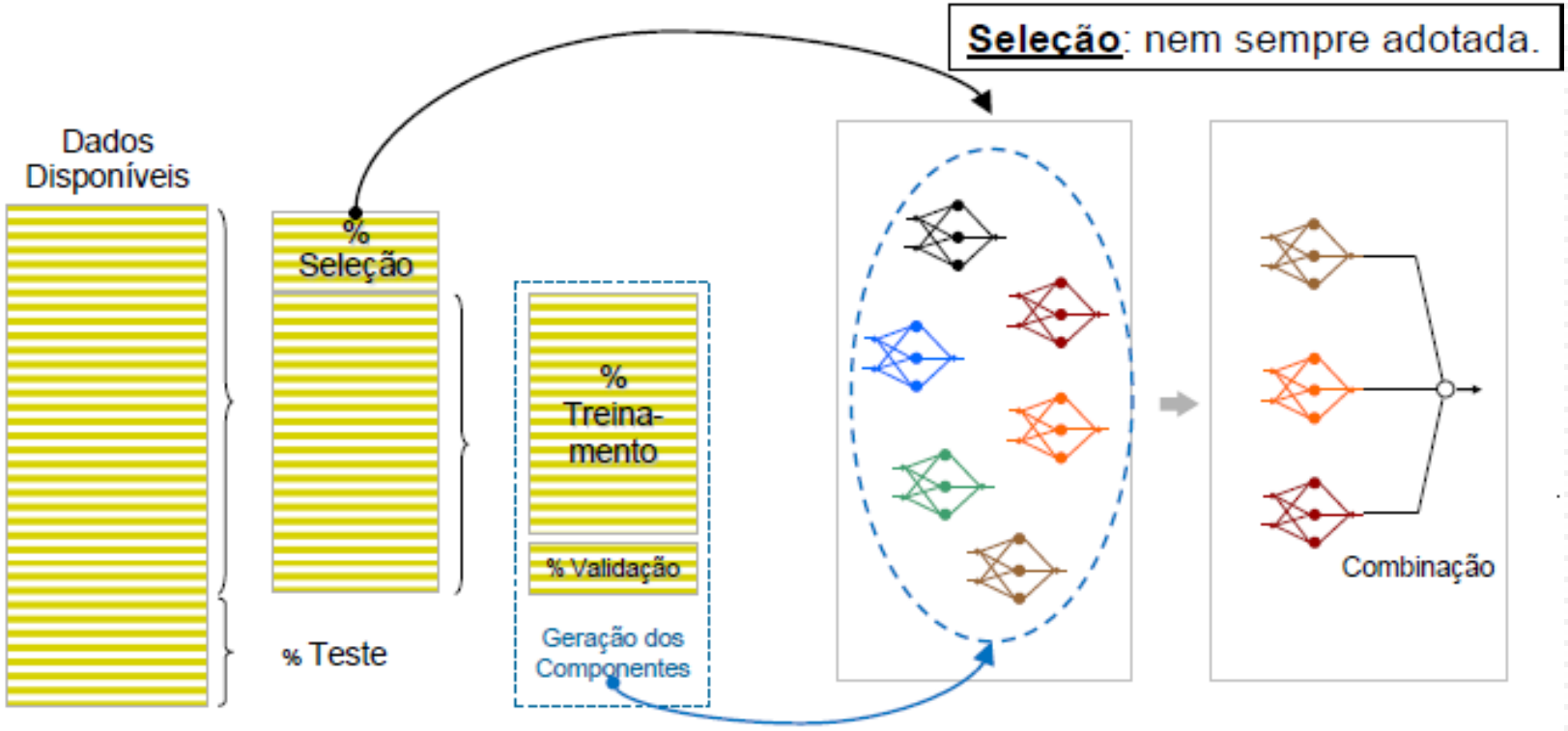
Ensembles: Etapas de Construção



Ensembles: Etapas de Construção



Ensembles: Etapas de Construção



Geração de componentes: principal etapa em que atuam os métodos de introdução de diversidade.

Ensembles: Etapas de Construção

- Separação ideal dos dados:
  - Treinamento;
  - Validação;
  - Seleção;
  - Teste;
- Nem sempre é possível fazer tal divisão, quando se tem um pequeno número de amostras;
- Divide-se então o conjunto de dados apenas em:
  - Treinamento, Validação (usado também na seleção) e Teste; ou
  - Treinamento (usado no treinamento e seleção) e Teste.

Ensembles: Etapas de Construção

- Seleção
  - Exemplos de técnicas de seleção de componentes:
    - Construtiva sem exploração;
    - Construtiva com exploração;
    - Poda sem exploração;
    - Poda com exploração;
    - Uso de alguma meta-heurística (GA, Estratégia Evolutiva, ACO,...).

Ensembles: Etapas de Construção

- **Seleção: Construtiva sem Exploração**

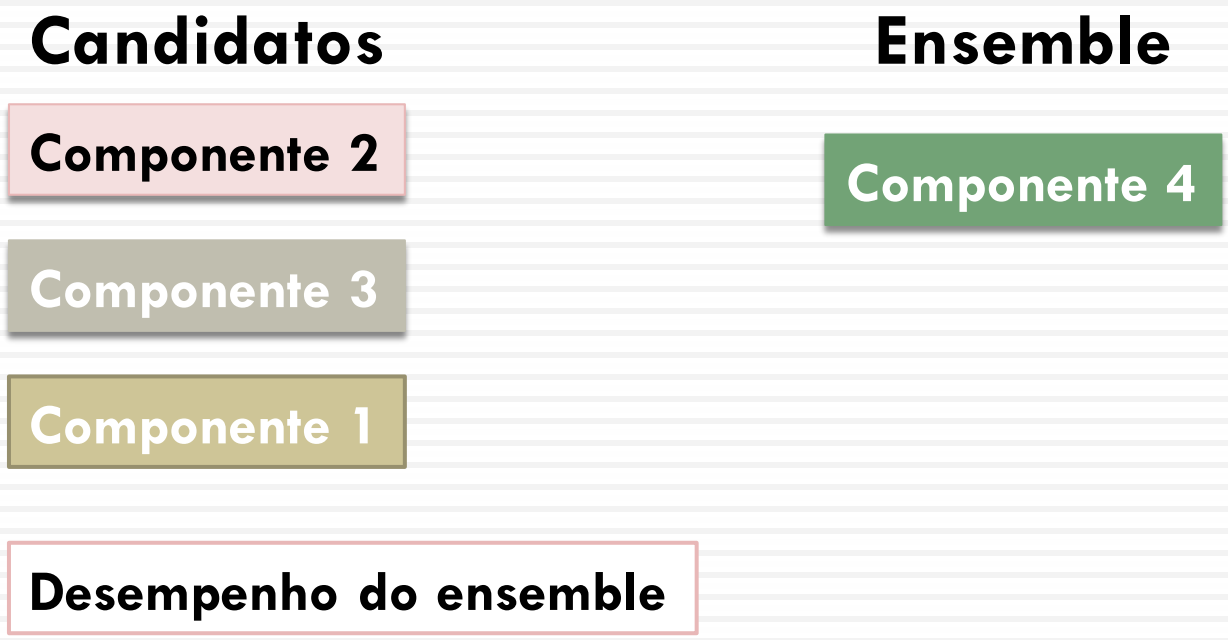
Ensembles: Etapas de Construção

- **Seleção: Construtiva sem Exploração**

- **Ordena os candidatos de acordo com o seu desempenho individual**

Ensembles: Etapas de Construção

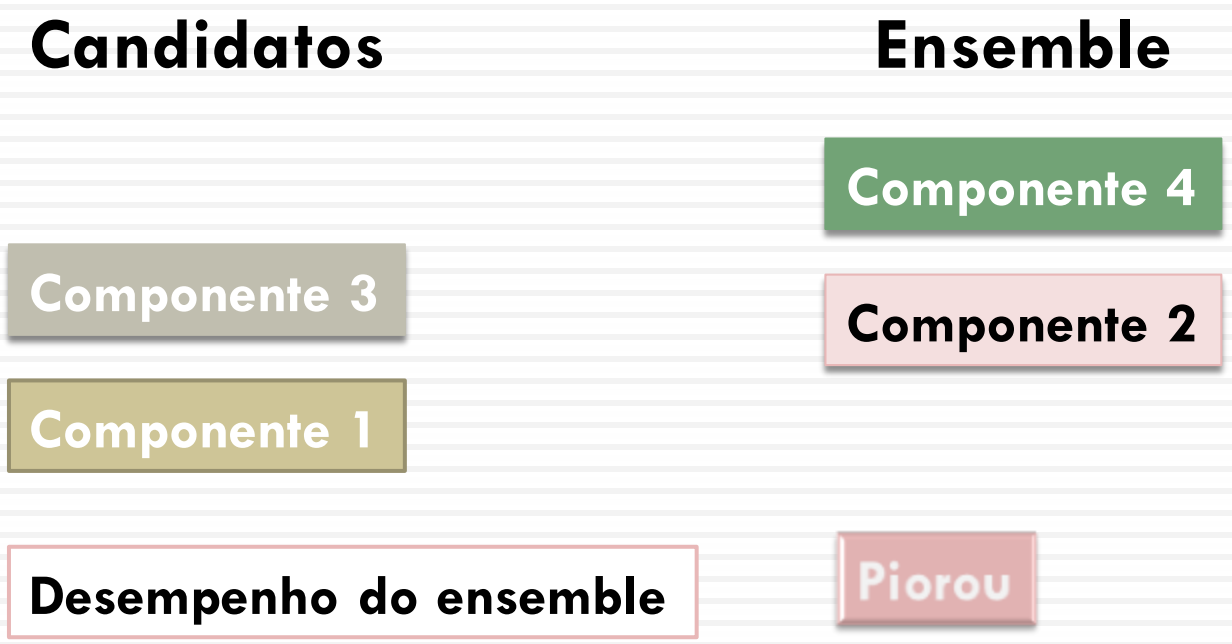
- **Seleção: Construtiva sem Exploração**



- **Insere o melhor candidato no ensemble e avalia seu desempenho**

Ensembles: Etapas de Construção

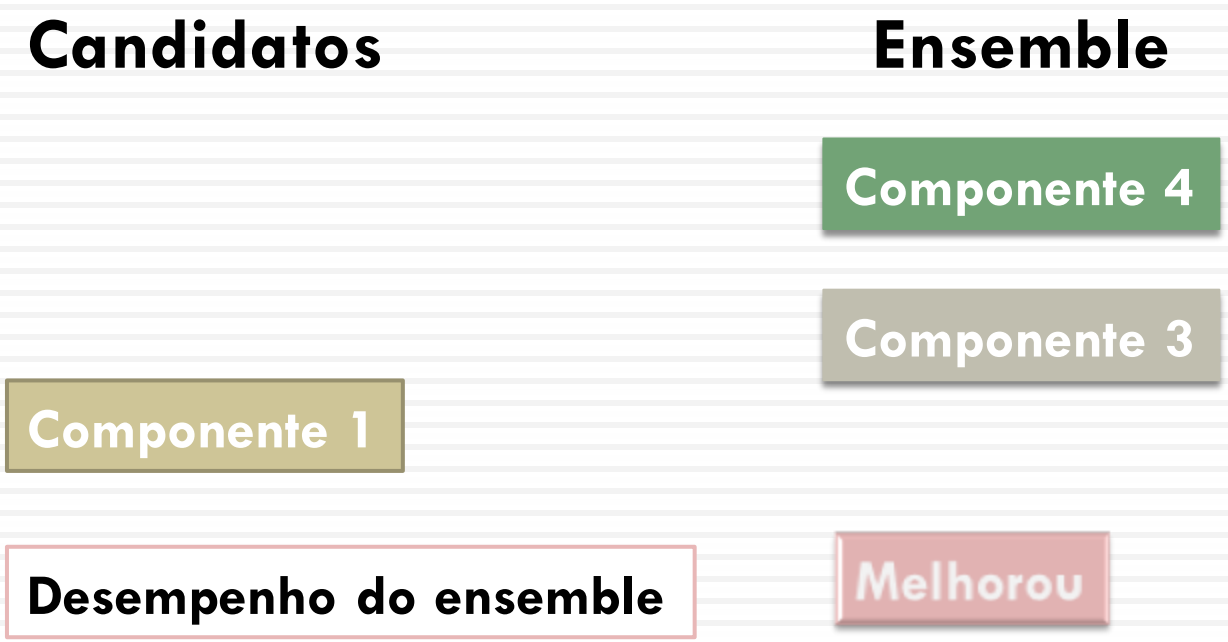
- **Seleção: Construtiva sem Exploração**



- **Insere o segundo candidato de melhor desempenho individual e verifica se o desempenho do ensemble melhorou. Caso tenha melhorado, o ensemble passa a ter dois componentes. Caso contrário, o candidato inserido é removido.**

Ensembles: Etapas de Construção

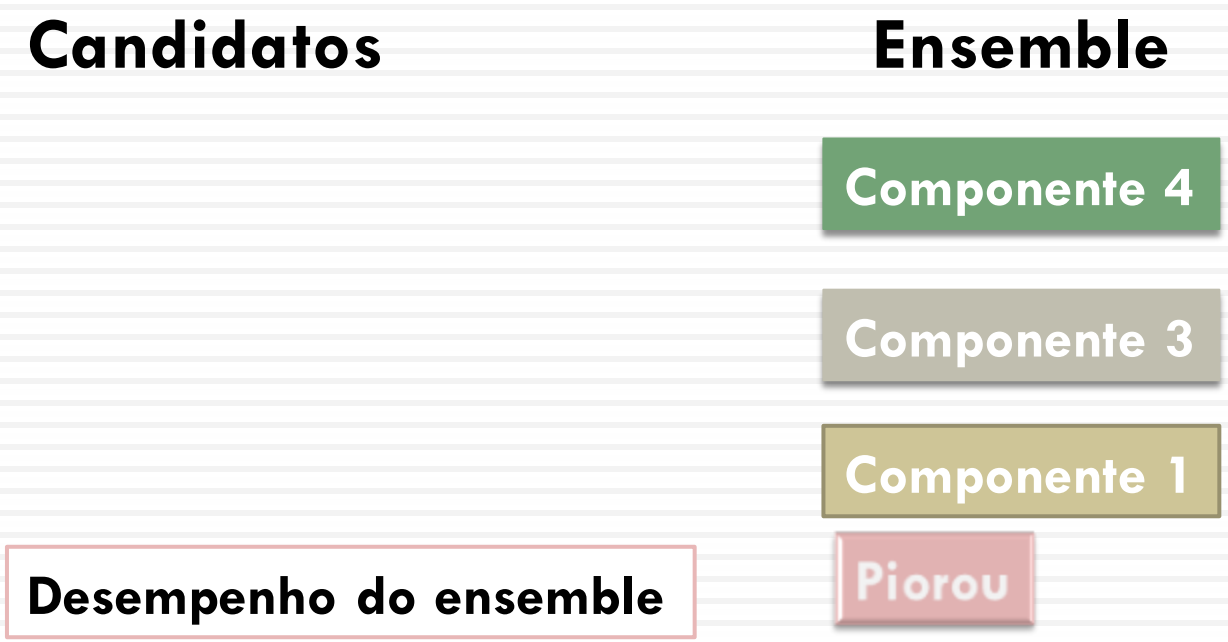
- **Seleção: Construtiva sem Exploração**



- **Repita o processo para os demais componente.**

Ensembles: Etapas de Construção

- **Seleção: Construtiva sem Exploração**



- **Repita o processo para os demais componente.**

Ensembles: Etapas de Construção

- **Seleção: Construtiva sem Exploração**

Candidatos **Ensemble**

Componente 4

Componente 3

- **Ensemble com dois componentes.**

Ensembles: Etapas de Construção

- **Seleção: Construtiva com Exploração**
  - Esta técnica também inicia com a ordenação dos *M candidatos* e *inserção do de melhor desempenho no ensemble;*
  - No entanto, ao invés de inserir o segundo candidato de melhor desempenho individual, todos os *M-1 restantes são considerados* e *aquele que gerar o aumento de performance mais significativo no ensemble é inserido;*
  - Caso nenhum candidato melhore o desempenho do ensemble, o processo é encerrado;
  - Caso haja melhoras, o processo continua e se repete para os demais M-2 candidatos;

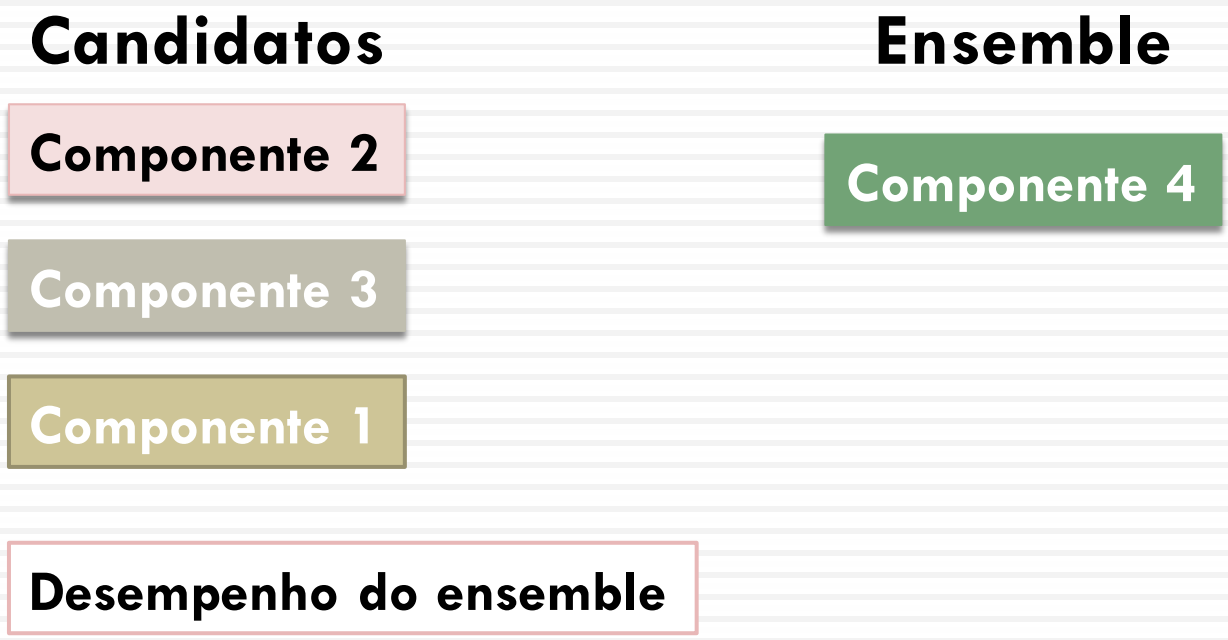
Ensembles: Etapas de Construção

- **Seleção: Construtiva sem Exploração**

- **Ordena os candidatos de acordo com o seu desempenho individual**

Ensembles: Etapas de Construção

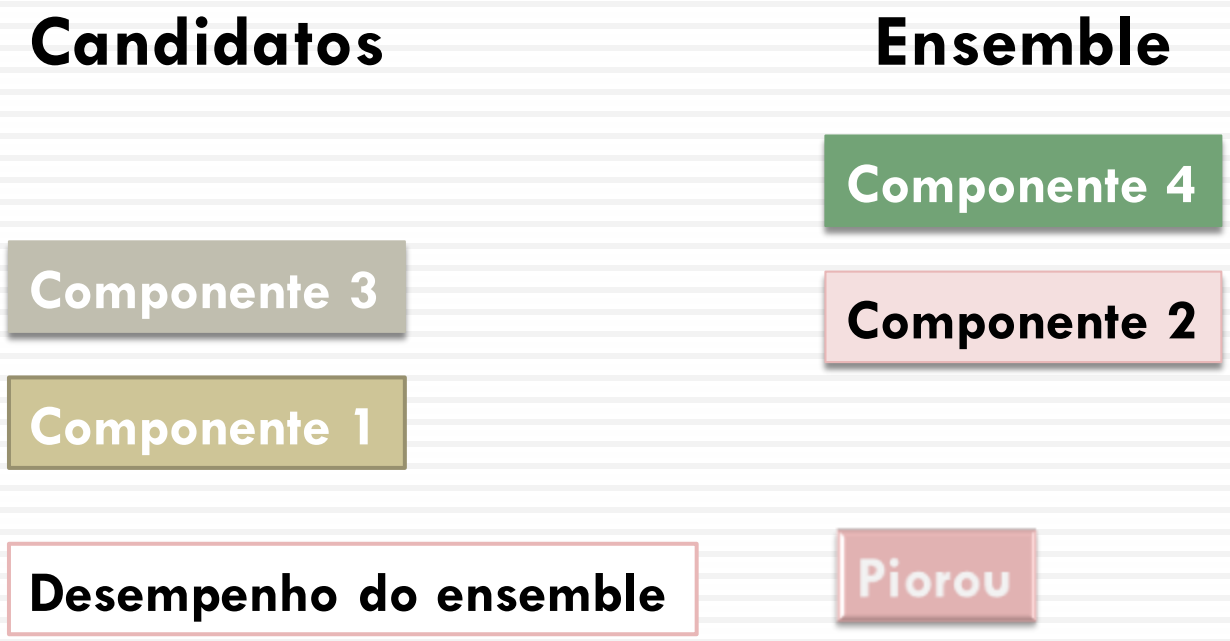
- **Seleção: Construtiva sem Exploração**



- **Insere o melhor candidato no ensemble e avalia seu desempenho**

Ensembles: Etapas de Construção

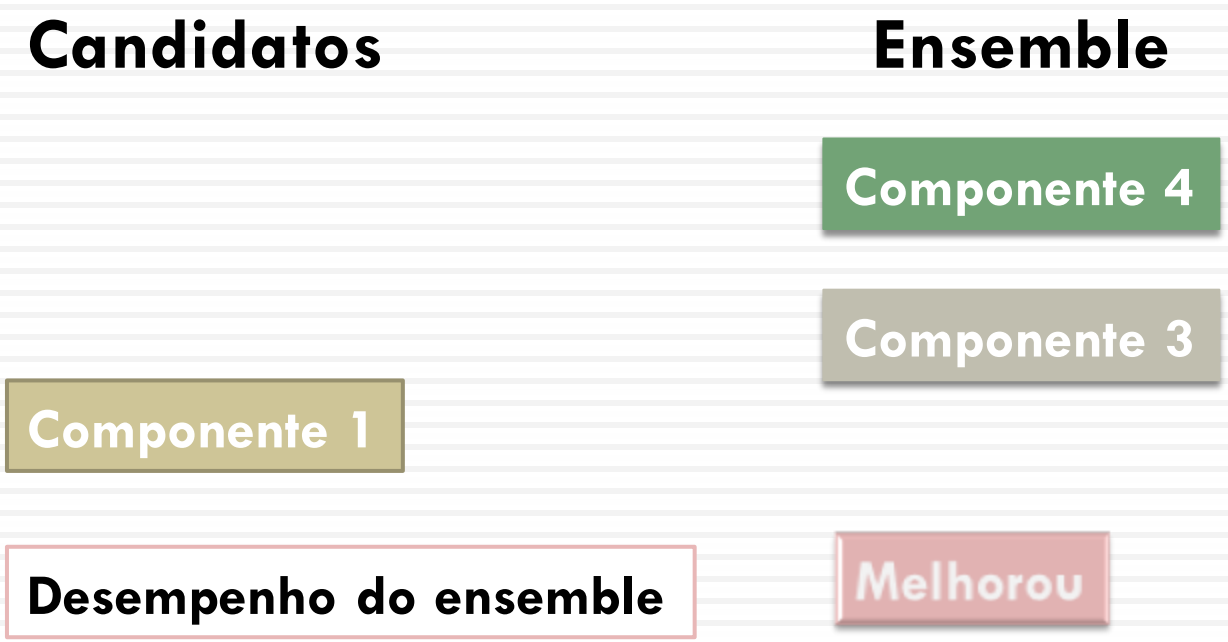
- **Seleção: Construtiva sem Exploração**



- **Insere o segundo candidato de melhor desempenho individual e verifica se o desempenho do ensemble melhorou.**

Ensembles: Etapas de Construção

- **Seleção: Construtiva sem Exploração**



- **Repita o processo para os demais componente.**

Ensembles: Etapas de Construção

- **Seleção: Construtiva sem Exploração**

- **Repita o processo para os demais componente.**

Ensembles: Etapas de Construção

- **Seleção: Construtiva sem Exploração**

- **Mantém no ensemble aquele que proporcionou a maior melhora.**

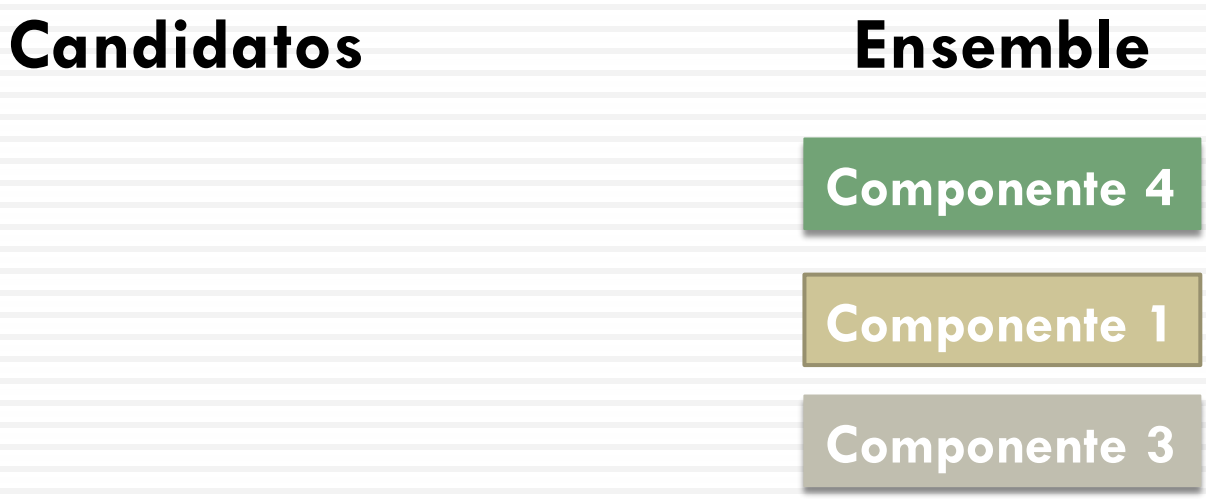
Ensembles: Etapas de Construção

- **Seleção: Construtiva sem Exploração**

- **Repita o processo novamente.**

Ensembles: Etapas de Construção

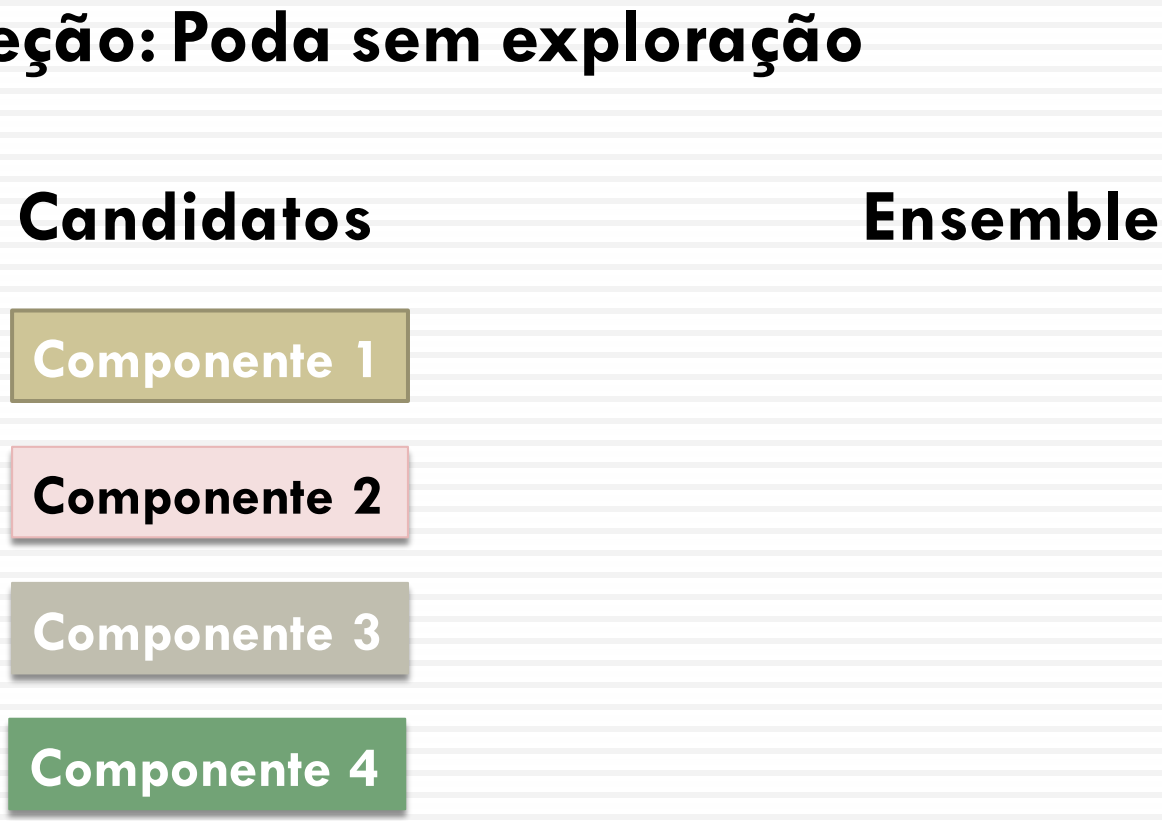
- **Seleção: Construtiva sem Exploração**



- **Ensemble com três componentes.**

Ensembles: Etapas de Construção

- **Seleção: Poda sem exploração**



Ensembles: Etapas de Construção

- **Seleção: Poda sem exploração**

- **Ordena os candidatos de acordo com o seu desempenho individual**

Ensembles: Etapas de Construção

- **Seleção: Poda sem exploração**

- **Insira todos os candidatos no ensemble e avalia seu desempenho**

Ensembles: Etapas de Construção

- **Seleção: Poda sem exploração**

- **Remova o candidato de pior desempenho individual e verifica se o desempenho do ensemble melhorou. Caso tenha melhorado, o ensemble passa a ter M-1 componentes. Caso contrário, o candidato removido é inserido novamente.**

Ensembles: Etapas de Construção

- **Seleção: Poda sem exploração**

- **Repita o processo para os demais componentes, exceto para o de melhor desempenho que nunca é removido**

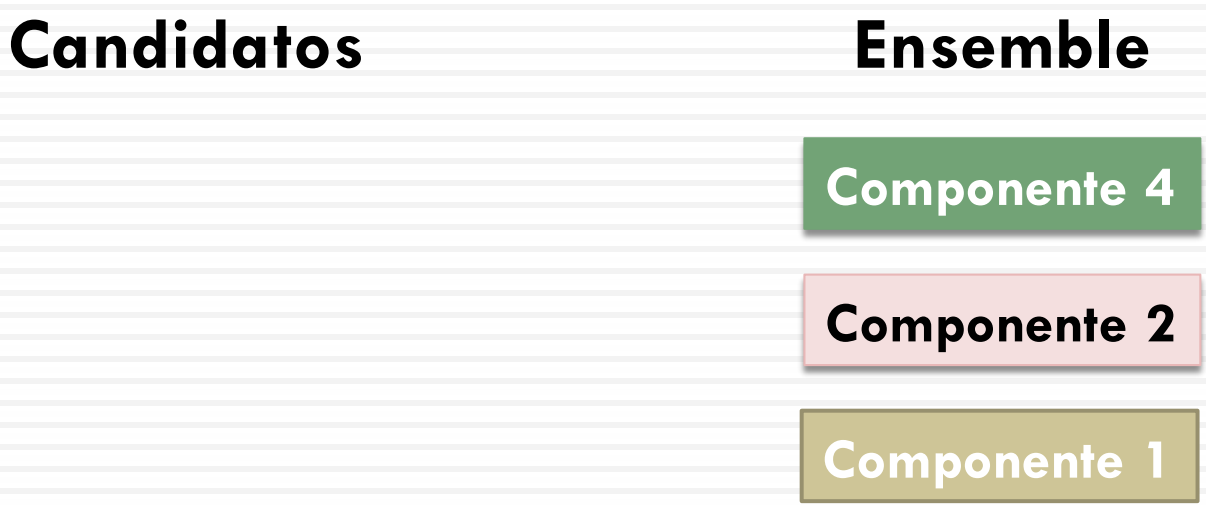
Ensembles: Etapas de Construção

- **Seleção: Poda sem exploração**

- **Repita o processo para os demais componentes, exceto para o de melhor desempenho que nunca é removido**

Ensembles: Etapas de Construção

- **Seleção: Poda sem exploração**



- **Ensemble com três componentes.**

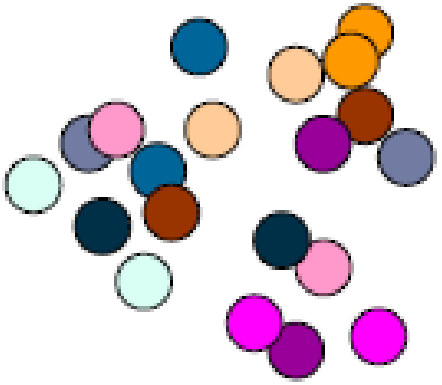
Ensembles: Etapas de Construção

- **Seleção: Poda com exploração**
  - Nesta técnica os candidatos são novamente ordenados e inseridos no ensemble;
  - No entanto, ao invés de se remover o de pior desempenho individual, todos os candidatos são considerados exceto o melhor deles;
  - Aquele candidato que ao ser removido promove o maior aumento de desempenho do ensemble é definitivamente excluído e o processo se repete para os demais;
  - Caso nenhuma remoção produza um aumento na performance do ensemble, o procedimento é encerrado.

Ensembles: Etapas de Construção

- **Seleção: Meta-heuristica**

Candidatos



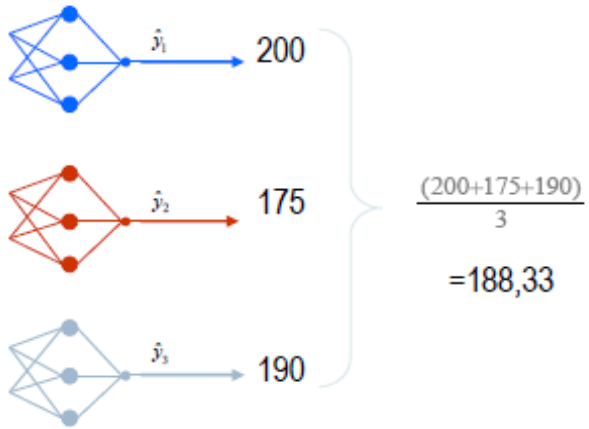
Ensemble

Ensemble com 3 componentes.

Ensembles: Etapas de Construção

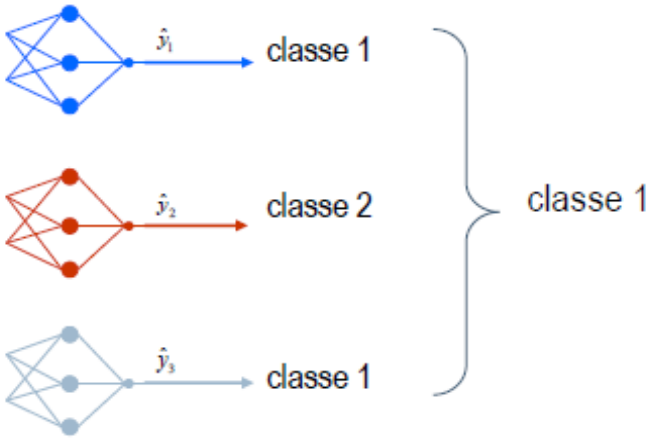
- Combinação

- Regressão:
  - Média Simples:



  - Média Ponderada sem bias;
  - Média Ponderada com bias;
  - ...

- Classificação:
  - Voto Majoritário:



  - Winner-takes-all;
  - ...

Ensembles: Etapas de Construção

- Combinação: Regressão

- Média ponderada sem bias:

$$y^k = p_1 \times y_1^k + p_2 \times y_2^k + \cdots + p_M \times y_M^k$$

- Média ponderada com bias:

$$y^k = p_0^k + p_1^k \times y_1^k + p_2^k \times y_2^k + \cdots + p_M^k \times y_M^k$$

- onde M é o número de componentes no ensemble, y^k é a k-ésima saída do ensemble, y_i^k é a k-ésima saída do i-ésimo componente e p_i é o peso atribuído ao componente i;

Ensembles: Etapas de Construção

- Combinação: Regressão

- Para obter os pesos p_i, basta resolver o seguinte problema:

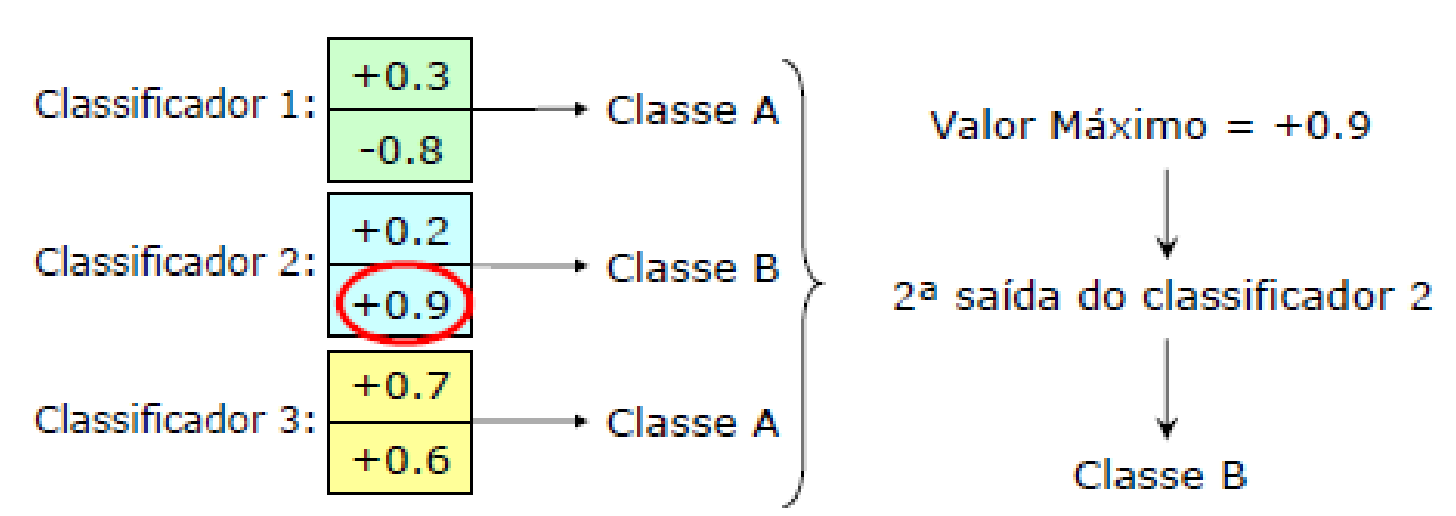
$$\min_{p^k} \frac{1}{2} \| \mathsf{Y}\, \vec{p}^{k} - \vec{y}_d^{k} \|_2^2$$

sujeito a $\sum_{i=1}^{M} p_i^k = 1$ e $p_i^k \geq 0, \forall i \in [1, \ldots, M]$

Esta restrição geralmente é adotada para evitar inversões das saídas dos componentes (pesos negativos).

Ensembles: Etapas de Construção

- Combinação: Regressão
  - Winner-takes-all:
    - É uma técnica de combinação não-linear e elitista, onde a saída do ensemble corresponde à saída do componente que possuir maior "*certeza*":



Ensembles: Etapas de Construção

- **Treinamento: inserção de diversidade**
  - Na etapa de construção de ensembles, existem técnicas que tentam explicitamente otimizar uma dada métrica de diversidade, enquanto que outras não.
  - Isto nos permite fazer uma distinção entre essas duas abordagens, classificando cada um dos métodos como:
    - Métodos implícitos; e
    - Métodos explícitos.
  - Os métodos *implícitos se baseiam em aleatoriedades presentes na etapa de treinamento para gerar diversidade.*

Ensembles: Etapas de Construção

- **Treinamento: inserção de diversidade**
  - Uma outra taxonomia, também relacionada à inserção de diversidade, pode ser adotada para os métodos de treinamento de componentes.
  - Esta taxonomia se baseia na maneira como cada técnica atua sobre o *espaço de hipóteses*:

    - ***Hipótese: é cada mapeamento entrada-saída feito por um componente do ensemble.***
    - ***Espaço de Hipóteses: conjunto de todos os mapeamentos possíveis de serem representados pelos componentes em questão.***

Ensembles: Etapas de Construção

- **Treinamento: inserção de diversidade**
  - Métodos que atuam sobre o ponto de partida no espaço de hipóteses:
    - Os métodos incluídos neste grupo variam *os pontos de partida da busca no espaço de hipóteses, influenciando dessa forma o ponto de convergência.*
  - Métodos que atuam sobre os dados de treinamento:
    - Buscam gerar componentes que produzam mapeamentos diferentes através do fornecimento de *conjuntos de dados de treinamento diferentes para cada um dos componentes* do *ensemble (os estímulos de entrada serão distintos).*

Ensembles: Etapas de Construção

- **Treinamento: inserção de diversidade**
  - Métodos que manipulam a arquitetura de cada componente:
    - Estes métodos *variam a arquitetura de cada componente no ensemble, de maneira que diferentes conjuntos de hipóteses estejam acessíveis para cada componente.*
    - Ou seja, como os componentes do ensemble possuem arquiteturas diferentes, os conjuntos de hipóteses associados a esses componentes também serão distintos, o que pode contribuir para a diversidade.
    - **Ex.: utilizar redes neurais MLP com diferentes números de neurônios nas camadas intermediárias.**

Ensembles: Etapas de Construção

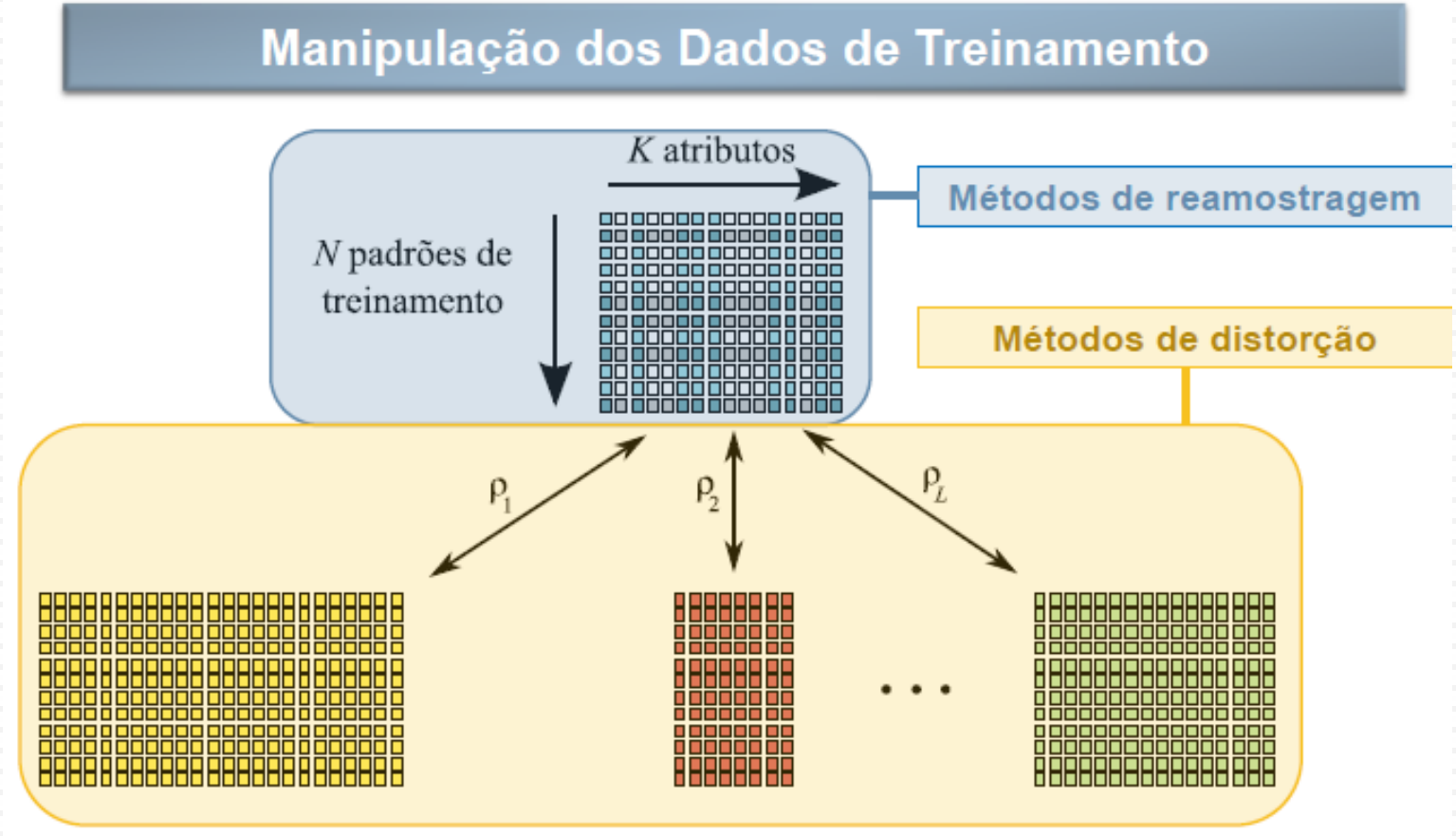
- **Treinamento: inserção de diversidade**
  - Métodos que atuam sobre a forma de exploração do espaço de hipóteses:
    - Alterando a forma de exploração do espaço de hipóteses, esses métodos estimulam os diferentes componentes a convergirem para diferentes hipóteses, mesmo tendo um mesmo ponto de partida.
  - Métodos Híbridos:
    - Formados por alguma combinação dos métodos anteriores.

Ensembles: Etapas de Construção

- Manipulação dos Dados de Treinamento
  - Estes métodos buscam produzir diversidade através do fornecimento de *conjuntos de treinamento diferentes para cada um dos componentes;*
  - É uma das formas de treinamento de ensembles mais pesquisadas;
  - Possibilidades:
    - Fornecer, a cada componente em treinamento, subconjuntos de amostras diferentes, com todos os atributos do conjunto de treinamento original;
    - Fornecer todas as amostras presentes mas formando-se subconjuntos com atributos diferentes;
    - Pré-processar os atributos de forma a obter uma representação diferente.

Ensembles: Etapas de Construção

- Manipulação dos Dados de Treinamento



Ensembles: Etapas de Construção

- **Manipulação dos Dados de Treinamento**
  - Métodos de reamostragem: Krogh & Vedelsby (1995)
    - Se baseia no k-fold cross-validation;
    - Para um ensemble com k componentes, divide aleatoriamente o conjunto de dados em k subconjuntos disjuntos;
    - Gera-se o conjunto de treinamento para cada membro do ensemble através da união de k 1 subconjuntos, sendo que para cada membro do ensemble um subconjunto distinto é deixado de fora.

Ensembles: Etapas de Construção

- **Manipulação dos Dados de Treinamento**
  - **Métodos de reamostragem: *Bagging (Breiman, 1996)***
    - Gera vários conjuntos de treinamento a partir de amostragem uniforme do conjunto de dados, com reposição;
    - Utiliza cada um desses conjuntos para treinar cada componente do ensemble;
    - Os conjuntos de treinamento possuem o mesmo número de amostras que o conjunto de dados original, mas algumas amostras podem ser selecionadas mais de uma vez, o que consequentemente implica que podem existir amostras que não são selecionadas;
    - Após a amostragem dos dados, permite o treinamento em paralelo dos componentes.

Ensembles: Etapas de Construção

- **Manipulação dos Dados de Treinamento**
  - **Métodos de reamostragem: *Boosting (Schapire, 1990)***
    - Foi aperfeiçoado por Freund & Schapire (1995) e Freund (1995);
    - Os conjuntos de treinamento *não são gerados via amostragem uniforme, como no algoritmo Bagging;*
    - A probabilidade de uma dada amostra ser escolhida depende da contribuição desta para o erro dos componentes já treinados;
    - Amostras que apresentam maior erro quando submetidas aos componentes já treinados têm maiores chances de comporem o conjunto de treinamento do próximo componente a ser treinado;
    - Exige que o treinamento dos componentes seja feito sequencialmente;

Ensembles: Etapas de Construção

- **Manipulação dos Dados de Treinamento**
  - **Métodos de reamostragem: *Boosting (Schapire, 1990)* – *continuação*.**
    - O método mais popular é o *AdaBoost (Freund & Schapire, 1996);*
    - Cada componente é treinado sequencialmente;
    - A amostragem dos dados de treinamento é feita levando-se em conta os erros do componente treinado *na etapa imediatamente anterior*.
  - Oza (2003) propôs uma variante do AdaBoost:
    - A distribuição das amostras depende dos erros de todos os componentes treinados até o momento e não apenas do componente treinado na etapa anterior.

Ensembles: Etapas de Construção

Original	1	2	3	4	5	6	7	8
Training set 1	2	7	8	3	7	6	3	1
Training set 2	7	8	5	6	4	2	7	1
Training set 3	3	6	2	7	5	6	2	2
Training set 4	4	5	1	4	6	4	3	8

Ensembles: Etapas de Construção

Original	1	2	3	4	5	6	7	8
Training set 1	2	7	8	3	7	6	3	1
Training set 2	1	4	5	4	1	5	6	4
Training set 3	7	1	5	8	1	8	1	4
Training set 4	1	1	6	1	1	3	1	5

Ensembles: Etapas de Construção

- **Manipulação dos Dados de Treinamento**
  - Métodos de distorção: *Sharkey et al. (1996)*
    - Adiciona novos atributos gerados através de uma transformação aleatória;
    - Nesta transformação, os atributos originais são passados por uma RNA não treinada.
  - Métodos de distorção: *Raviv & Intrator (1996)*
    - Aplicam uma amostragem como em *Bagging, mas adicionando ruído gaussiano aos dados de entrada.*

Ensembles: Etapas de Construção

- **Manipulação dos Dados de Treinamento**
  - Métodos de distorção: *Liao & Moody (2000)*
    - Agrupam os atributos de entrada de acordo com sua informação mútua (variáveis estatisticamente semelhantes são agrupadas);
    - Os conjuntos de treinamento são formados por atributos selecionados de grupos diferentes.
  - Métodos de distorção: Breiman (1998)
    - Propôs a adição de ruído à saída dos dados;
    - Os resultados foram superiores aos obtidos via *Bagging mas próximos aos obtidos via AdaBoost,*

Ensembles: Etapas de Construção

- **Manip. da Arquitetura dos Componentes**
  - No caso de ensembles de redes neurais, as duas maneiras mais diretas de variar a arquitetura de seus componentes são:
    - O uso de redes MLP com números distintos de camadas ocultas e de neurônios nestas camadas;
    - O uso de redes neurais de diferentes tipos (ex.: MLPs, RBFs,...);
  - Partridge & Yates (1996) exploraram estas duas abordagens e concluíram que (para um único conjunto de dados):
    - O uso de números diferentes de neurônios na camada oculta só não é pior que inicializações aleatórias das redes (no aspecto diversidade);
    - Mistura de MLPs e RBFs é mais eficiente que variar o número de neurônios.

Ensembles: Etapas de Construção

- **Manip. da Arquitetura dos Componentes**
  - Opitz & Shavlik (1996) utilizaram um algoritmo evolutivo para otimizar as topologias dos componentes :
    - Cada componente foi treinado via *backpropagation;*
    - O processo de seleção visou otimizar métricas de diversidade.
  - Islam et al. (2003) propuseram o algoritmo CNNE (*Cooperative Neural Network Ensembles):*
    - Gera ensembles construtivamente, monitorando a diversidade durante o processo;
    - É capaz de determinar automaticamente o número de redes neurais no ensemble e o número de neurônios na camada oculta.

Ensembles: Etapas de Construção

- **Manip. da Arquitetura dos Componentes**
  - Uma outra abordagem que se encaixa nesta categoria é o uso de *componentes de paradigmas distintos* → *ensembles heterogêneos*
    - **Ex.:** para um problema de classificação, um ensemble heterogêneo poderia ter redes neurais e classificadores baseados em regras.
  - Os trabalhos nessa área mostram que o uso de diferentes paradigmas leva a componentes com diferentes especialidades e precisões, que podem apresentar diferentes desempenhos e, com isso, diferentes padrões de generalização;
  - Esta especialização pode indicar que a seleção de um único componente ao invés da combinação de todos eles pode ser mais vantajosa (Brown et al., 2005) → *mistura de especialistas;*

Ensembles: Etapas de Construção

- **Manip. da Arquitetura dos Componentes**
  - Nesta linha, Langdon et al. (2002) utilizaram redes neurais e árvores de decisão como componentes:
    - Aplicaram Programação Genética para evoluir uma regra de combinação dos indivíduos;
  - Soares et al. (2006) utilizaram como componentes MLPs, RBFs, classificadores naïve Bayes, máquinas de vetores suporte (SVM) e classificadores baseados em regras;
    - Propuseram duas técnicas de seleção (baseadas em algoritmo de agrupamento e *k-nearest neighbors) que buscam não apenas reduzir o erro do ensemble, mas também aumentar a diversidade de seus componentes.*

Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - Os métodos de criação de diversidade que atuam sob a forma de exploração do espaço de hipóteses podem ser divididos em duas sub-categorias:
    - Métodos de *otimização convencional*, *como os chamados métodos de penalidades, que adicionam um termo de custo (por ausência de diversidade) ao erro de cada componente, buscando a criação de hipóteses diversas; e*
    - Métodos de *busca exploratória*, *onde estão os métodos de busca populacionais que encorajam a diversidade na população de candidatos (ex. algoritmos evolutivos).*

Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - Métodos de Otimização Convencional:
    - Os métodos de treinamento individual de componentes normalmente têm como objetivo minimizar o erro na saída de cada componente;
    - Geralmente se baseiam no gradiente da função de erro (ex.: *backpropagation em ensembles de RNAs);*
    - Originalmente não têm nenhuma preocupação com a diversidade;
    - No caso de ensembles, além de reduzir o erro do componente sendo treinado, *deve-se também estimular a diversidade deste componente em relação aos demais (já treinados ou em processo de treinamento simultâneo);*

Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - Métodos de Otimização Convencional (cont'd):
    - Diante disso, surgiram os chamados Métodos de Penalidade;
    - Nestes métodos, o erro de cada componente se torna algo como

$$e_i = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(f_i(n) - d(n))^2 + \frac{1}{N}\sum_{n=1}^{N}\lambda R_i(n)$$

    - onde *N é o número de amostras, f_i é a saída do componente i, d é a saída desejada e λ é um fator de ponderação do termo de penalidade R_i.*
    - O termo de penalidade *R_i está diretamente associado à diversidade do componente i, e sua importância no treinamento é controlada por λ.*

Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - Métodos de Penalidade:
    - Rosen (1996) usou o seguinte termo de penalidade em ensembles de redes neurais:

$$R_i = \sum_{j=1}^{i-1} c(i, j)\, p_i$$

    - onde c(j, i) *é uma função de indicação que especifica quais redes i e j devem ser descorrelacionadas entre si e pi é o produto das polarizações das i-ésima e j-ésima redes, dado por:*

$$p_i = (f_i - d)(f_j - d)$$

    - onde f_i é a saída da rede i, f_j é a saída da rede j e d é a saída desejada.

Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - **Métodos de Penalidade – Correlação Negativa:**
    - Já Liu (1998) propôs uma extensão para o trabalho de Rosen (1996) que permite o treinamento simultâneo das redes neurais;
    - Esta metodologia ficou conhecida como Correlação Negativa;
    - Nesta abordagem, o termo de penalidade é dado por:

$$R_i = (f_i - \bar{f}) \sum_{j \neq i} (f_j - \bar{f})$$

    - onde $\bar{f}$ é a saída média de todo o ensemble no passo anterior

Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - **Métodos de Penalidade – Correlação Negativa:**
    - O método de Correlação Negativa foi aplicado com sucesso em vários trabalhos, superando consistentemente o desempenho de outros ensembles;
    - Isto se dá pois a técnica de Correlação Negativa controla *diretamente o termo de covariância entre os componentes, ajustando assim a diversidade do ensemble (Brown, 2004);*
    - No entanto, esta técnica foi concebida especificamente para tratar problemas de regressão.

Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - Métodos de Busca Exploratória:
    - Dentre os métodos de busca exploratória, os algoritmos evolutivos (Algoritmos Genéticos, Estratégias Evolutivas, Programação Genética, etc.) exercem grande importância nas aplicações atuais;
    - No entanto, na literatura de computação evolutiva, o termo *diversidade possui um conceito diferente do utilizado na literatura de ensembles*;
    - Em computação evolutiva, a diversidade da população se refere à *presença de indivíduos que exploram regiões distintas do espaço de busca*;

Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - Métodos de Busca Exploratória:
    - Com isso, a manutenção de uma população diversa de indivíduos permite uma exploração maior do espaço de busca e, consequentemente, uma maior eficiência na localização de soluções melhores.
    - Apesar dessas diferenças conceituais, alguns autores exploraram os mecanismos de manutenção de diversidade, já desenvolvidos em computação evolutiva, junto à questão de ensembles;
    - No trabalho de Yao & Liu (1998), uma população de redes neurais é evoluída e, ao final, toda a população é combinada em um ensemble. Para estimular a diversidade, foi utilizada a técnica de *fitness sharing*

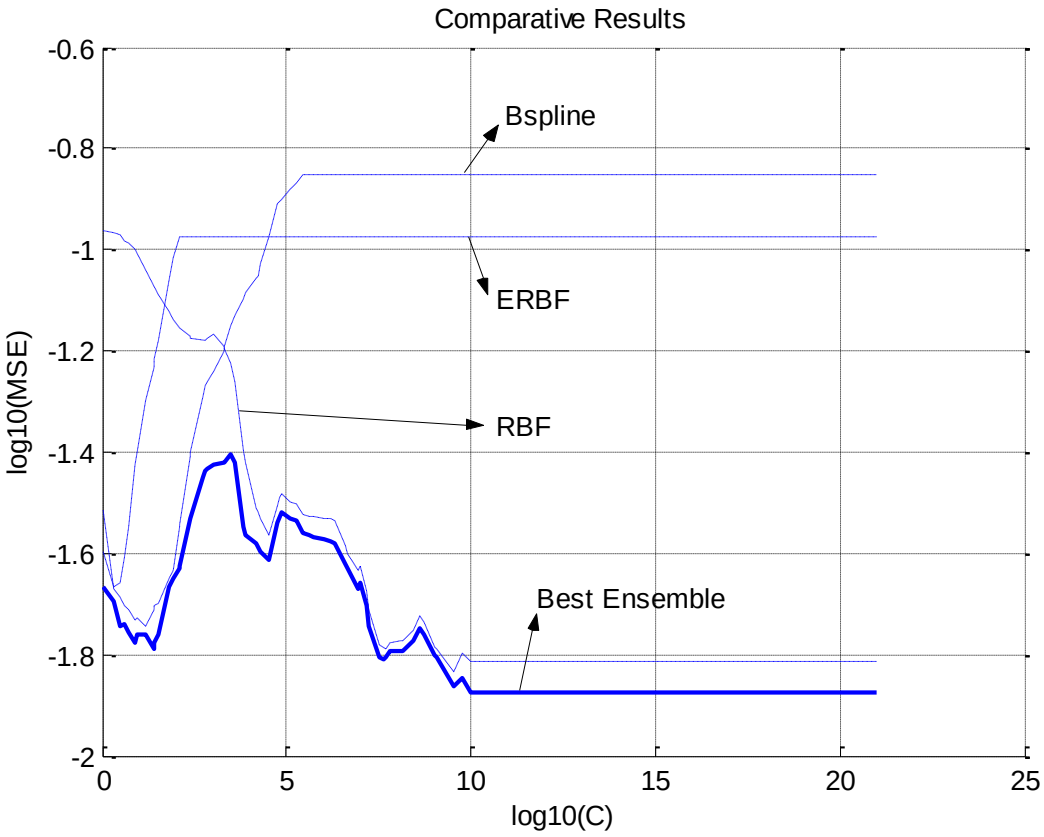
Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - Métodos de Busca Exploratória:
    - Já no trabalho de Khare & Yao (2002), tal conceito foi estendido para problemas de classificação, com a utilização da entropia de Kullback-Leibler como métrica de diversidade a ser otimizada durante a busca.
    - Um aspecto importante que deve ser ressaltado aqui é que, em ensembles, deseja-se que os componentes apresentem diversidade de erros, o que pode ser bem diferente de diversidade de indivíduos em uma população:
      - Ex.: Em redes MLP, duas redes com conjunto de pesos distintos (ou seja, diversas) podem levar a um mesmo padrão de saídas, o que não é desejável em um ensemble.

Ensembles: Etapas de Construção

- **Manip. da Forma de Expl. do Espaço de Hipóteses**
  - Métodos de Busca Exploratória:
    - Diante disso, Coelho & Von Zuben (2006) propuseram a aplicação de um sistema imunológico artificial para treinamento de redes MLPs que tratasse especificamente desta situação;
    - Como as redes geradas ao final do treinamento seriam candidatas a formarem um ensemble, o mecanismo de manutenção de diversidade do algoritmo foi modificado de forma a não gerar redes de tal modo a produzir conjuntos de pesos diversos durante o treinamento, mas redes com padrões de saída distintos (não importando assim o grau de diversidade de seus vetores de pesos).

Dados com ruído (IJCNN – 2002)



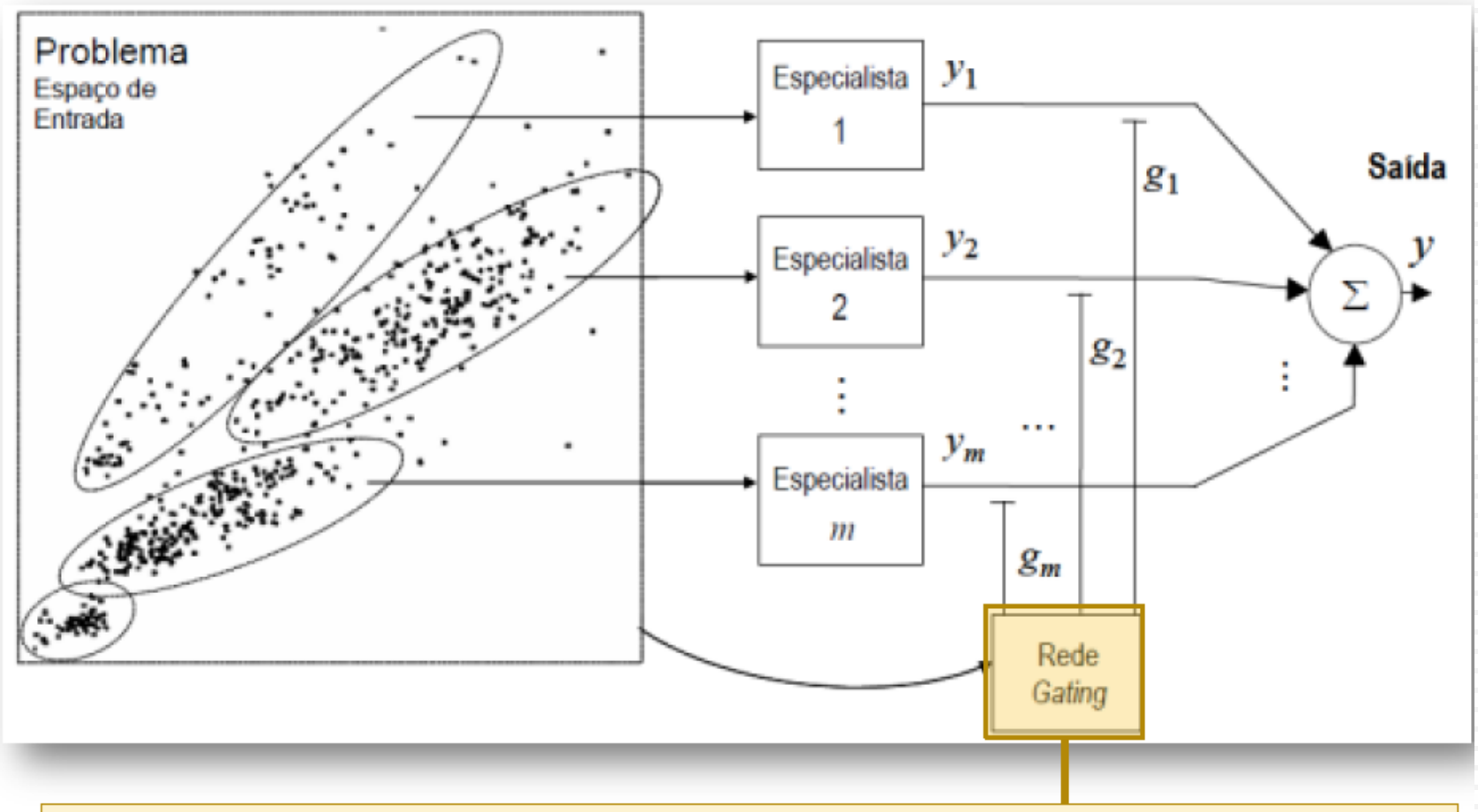
- Under the presence of noisy data, the performance of the HE-SVM approach when approximating the sinc function presented in Fig. 1, is superior to the S-SVM approaches for all values of C

Mistura de Especialista

Mistura de Especialistas

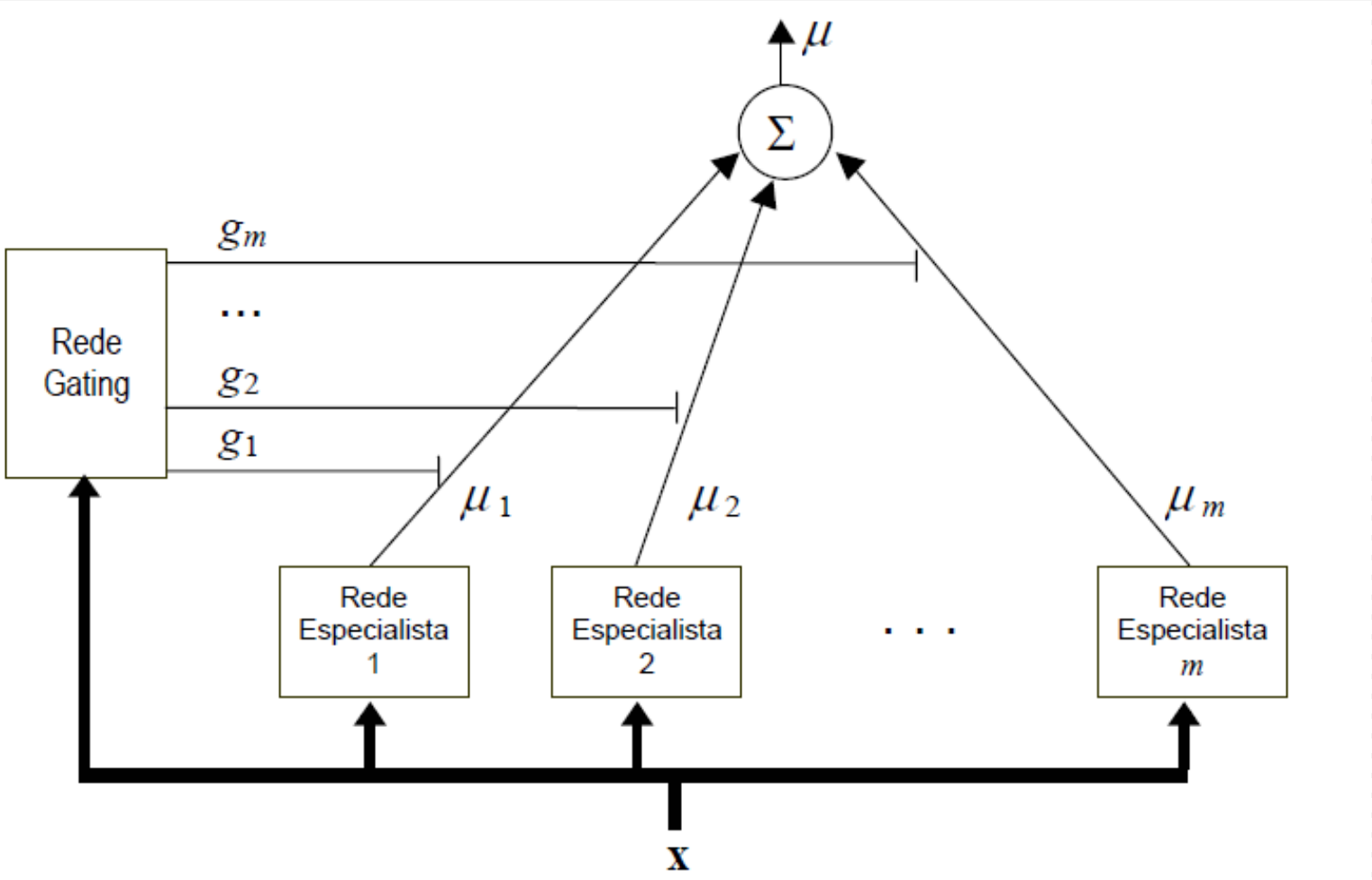
- Mistura de Especialistas (ME) é uma proposta de comitê de máquinas na qual o espaço de entrada é automaticamente dividido em regiões durante o treinamento;
- Se baseia no princípio de que estimadores são capazes de se especializar no tratamento de regiões particulares do problema;
- Com isso, para cada região existirá um único ou um subconjunto de especialistas mais indicados para atuar;
- O caráter dinâmico de MEs deve-se ao fato de que as regiões de atuação a serem alocadas para os especialistas não são definidas a priori.

Mistura de Especialista

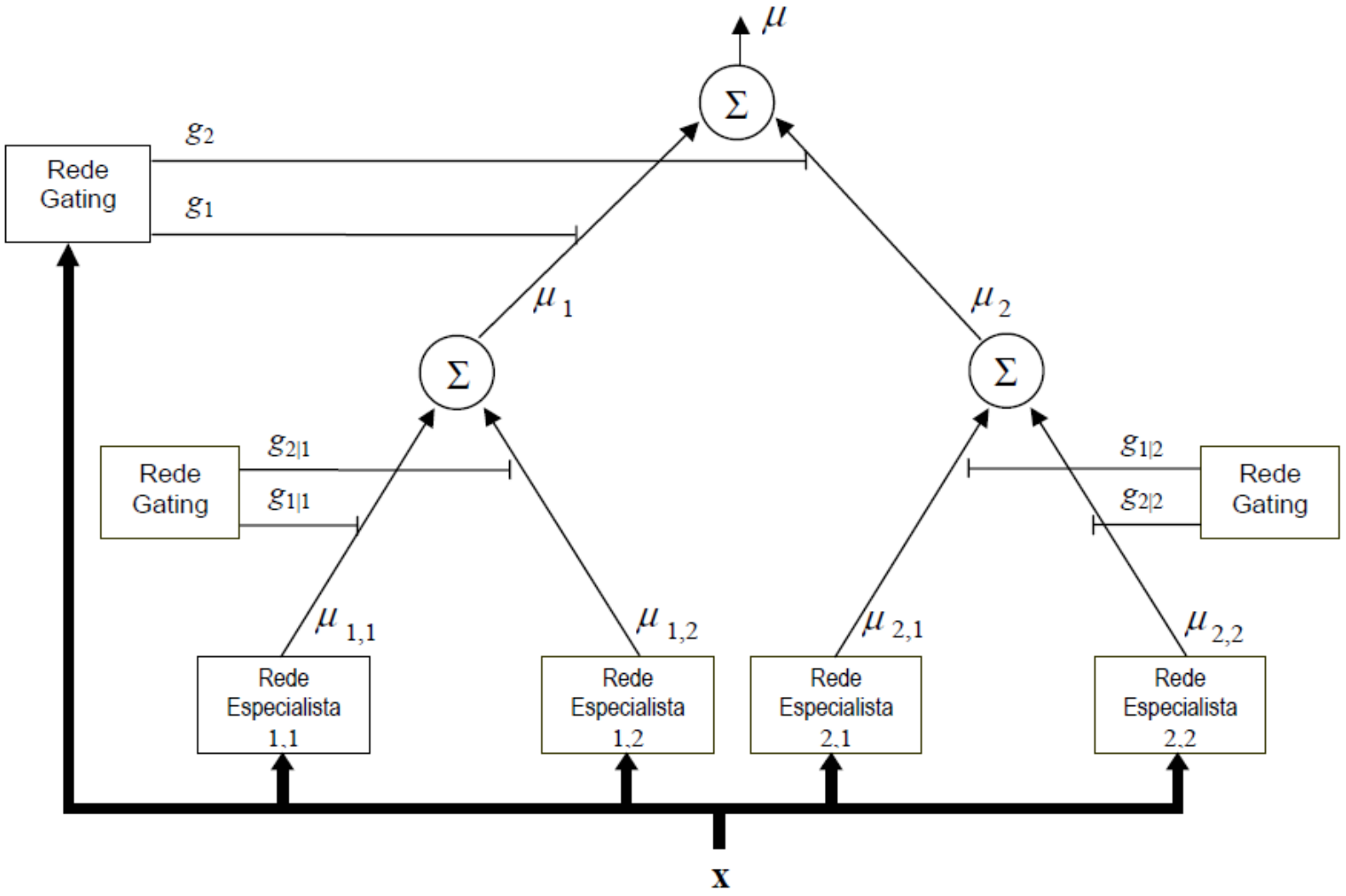


Responsável pelo aprendizado da ponderação apropriada dos especialistas para cada entrada

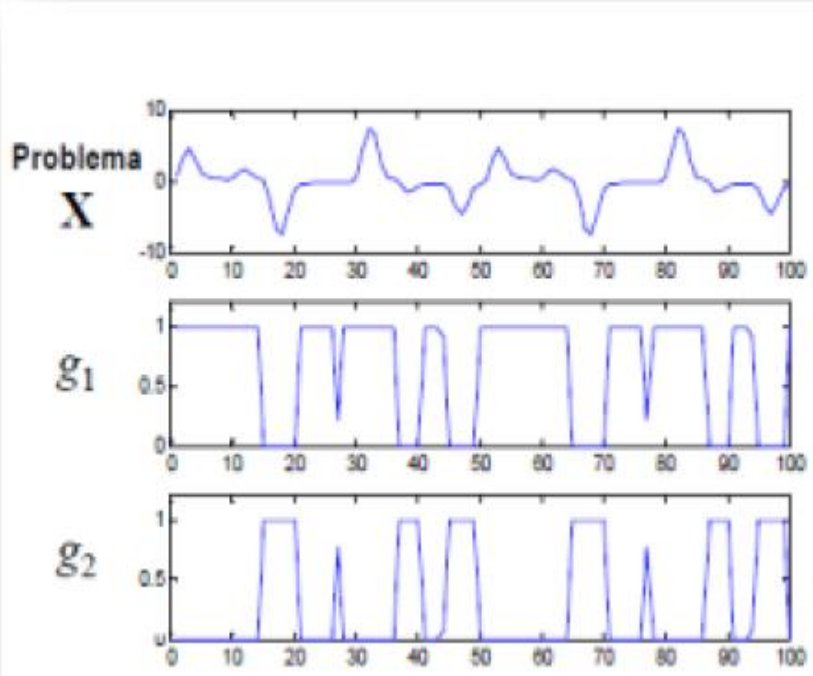
Mistura de Especialista

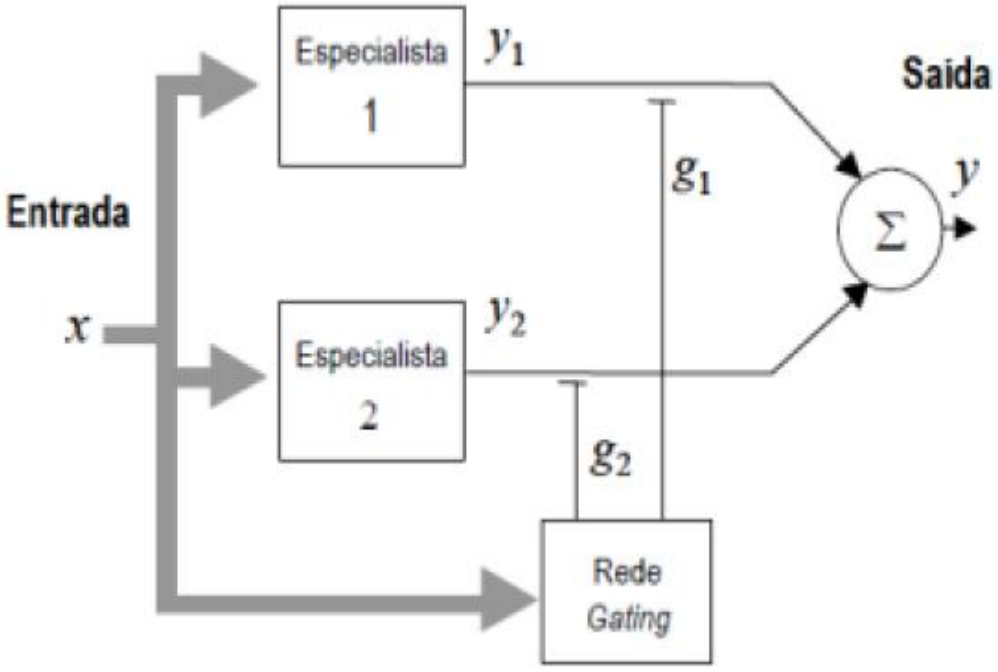


Mistura Hierárquica de Especialista

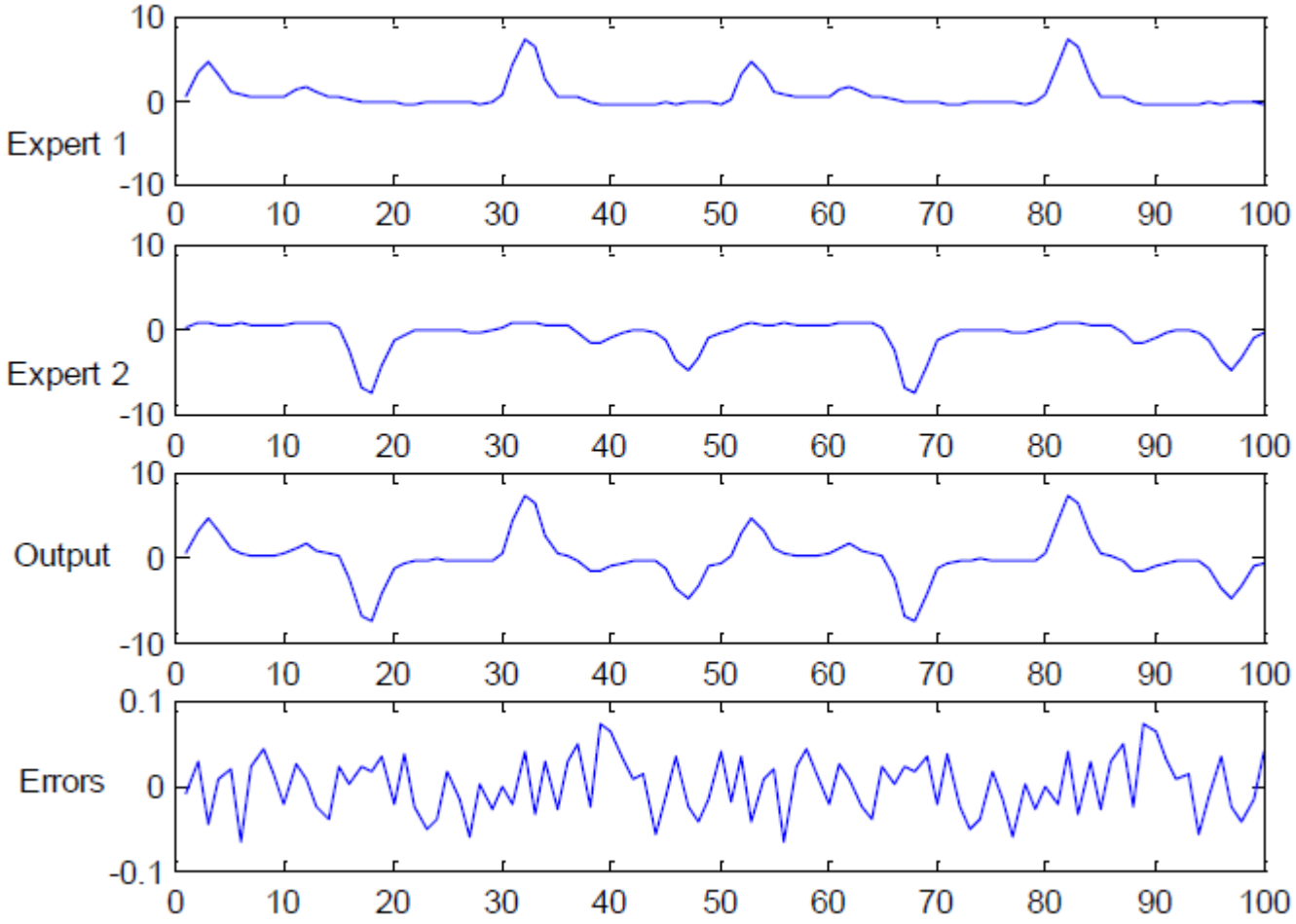


Exemplo: Mistura de Especialista

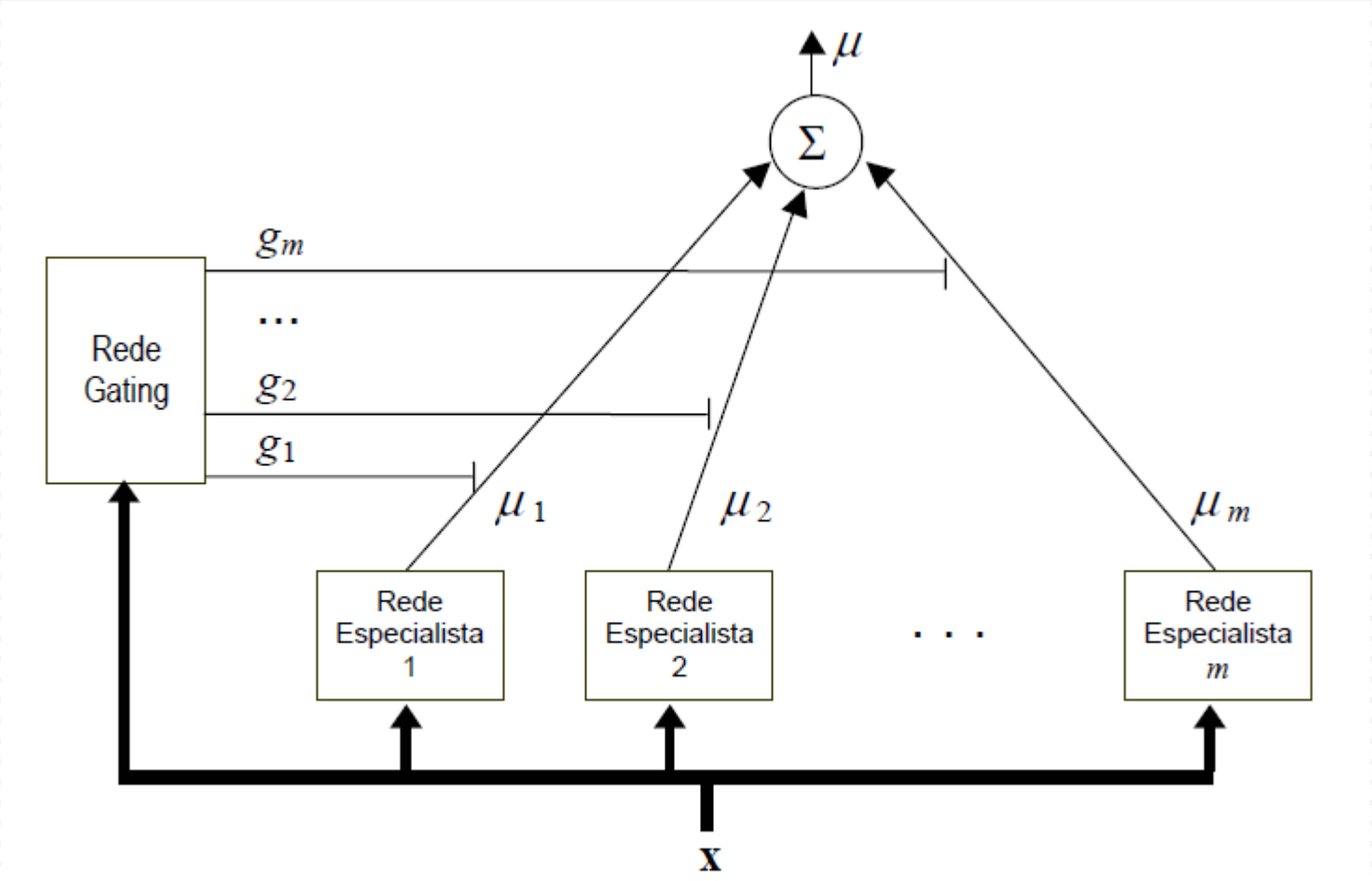




Exemplo: Mistura de Especialista



Mistura de Especialista: Formulação básica



Mistura de Especialista: Formulação básica

- Considere a arquitetura modular anterior composta de *m* módulos referidos como redes especialista, cada um implementando uma função parametrizada

$$\mu_i = f_i(\theta_i, x)$$

- onde θ_i é o vetor de parâmetros do especialista *i*
- *Considere que o especialista gera a saída* ***y*** *com probabilidade* $P(y \mid x, \theta_i)$ *, onde* μ_i *é a média da função densidade de probabilidade P*

Mistura de Especialista: Formulação básica

- Considerando que diferentes redes especialistas são apropriadas para diferentes regiões do espaço de entrada, a arquitetura requer um mecanismo capaz de identificar, para cada entrada **x**, o especialista ou combinação de especialistas mais capazes de produzir a saída correta, em termos probabilísticos.
- Isto é realizado por meio de uma rede auxiliar, conhecida como rede *gating*.

Mistura de Especialista: Formulação básica

- A interpretação probabilística da rede *gating é de um sistema que calcula, para cada* especialista, a probabilidade dele gerar a saída desejada, com base apenas no conhecimento da entrada **x.**

- Estas probabilidades são expressas pelos coeficientes g_i *(i=1,...,m), de modo* que estes devem ser não-negativos e devem produzir sempre o valor unitário quando somados, para cada **x.**

- Estes coeficientes não são constantes fixos, mas variam em função da entrada x.
  - Caso permaneçam estáticos para todas as entradas, a mistura de especialistas se torna um ensemble.

Mistura de Especialista: Formulação básica

- Há muitas formas de garantir que os coeficientes g_i *(i=1,...,m) atendam as restrições* anteriores.
- Uma abordagem é utilizar a função softmax (JACOBS et al., 1991).
- A função *softmax define* um conjunto de variáveis intermediárias $\xi_i (i=1,...,m)$ *como funções da entrada x e de um* vetor de parâmetros v_i *(i=1,...,m) na forma:*

$$\xi_i = \xi_i(x, v_i)$$

- Com isso, os coeficientes g_i *(i=1,...,m) podem ser definidos em termos de* ξ_i *(i=1,...,m) como segue*

$$g_i = \frac{\exp(\xi_i)}{\sum_{k=1}^{m} \exp(\xi_k)}$$

Mistura de Especialista: Formulação básica

- Considere que o conjunto de treinamento $\chi = \{(x_i, y_i)\}_{i=1}^{N}$ é gerado da seguinte forma:
  - dada uma entrada **x**, um especialista *i é escolhido com* probabilidade $P(i \mid x, v^0)$ *(onde o sobrescrito "0" será usado para distinguir os valores reais* dos parâmetros do modelo de probabilidade adotado daqueles estimados pela rede *gating* ou pela rede especialista).
  - Dada a escolha do especialista e dada a entrada, a saída desejada y é suposta ser gerada de acordo com a probabilidade $P(y \mid x, \theta_i^0)$.
  - *Cada um dos pares de* entrada-saída é suposto ser gerado independentemente

Mistura de Especialista: Formulação básica

- Assim, a probabilidade total de geração de y a partir de x é dada pela soma sobre i, na forma:

$$P(y \mid x, \Theta^o) = \sum_{i=1}^{m} P(i \mid x, v^o) P(y \mid x, \theta_i^o)$$

- onde Θ^o denota o vetor contendo todos os parâmetros, na forma

$$\Theta^o = [\theta_1^0, \theta_2^0, \ldots, \theta_m^0, v_1^0]^T$$

- A função densidade na equação acima é conhecida como *mistura de densidade* ou *função de verossimilhança*.
- *Tomando o logaritmo de m densidades acima e dado o conjunto de treinamento chega-se à seguinte medida de verossimilhança*

$$l(\chi, \Theta) = \sum_{j=1}^{N} \log \sum_{i=1}^{m} P(i \mid x_j, v) P(y_j \mid x_j, \theta_i)$$

Mistura de Especialista: Formulação básica

- Uma abordagem para maximizar o logaritmo da verossimilhança é usar o gradiente ascendente
- Calculando o gradiente de $l(.;.)$ com respeito a μ_i e ξ_i, resulta

$$\frac{\partial l(.;.)}{\partial \mu_i} = \sum_{i=1}^{N} h_i^{(t)} \frac{\partial}{\partial \mu_i} \log P(y^{(t)} \mid x^{(t)}, \theta_i)$$

$$\frac{\partial l(.;.)}{\partial \xi_i} = \sum_{t=1}^{N} \left(h_i^{(t)} - g_i^{(t)} \right)$$

$$h_i^{(t)} = P(i \mid x^{(t)}, y^{(t)}) = \frac{P(i, x^{(t)}) P(y^{(t)} \mid x^{(t)}, \theta_i)}{\sum_{i=1}^{m} P(j, x^{(t)}) P(y^{(t)} \mid x^{(t)}, \theta_j)}$$

Mistura de Especialista: Formulação básica

- $h_i^{(t)}$ é definida como a probabilidade a posteriori do i-ésimo especialista, condicionada à entrada $x^{(t)}$ e à saída $y^{(t)}$
- $g_i^{(t)}$ pode ser interpretada como a probabilidade a priori $P(i, x^{(t)})$, ou seja, a probabilidade da gating escolher o i-ésimo especialista, da somente a entrada $x^{(t)}$
- A equação de atualização da gating pode ser interpretada como uma forma de aproximar a probabilidade a posteriori utilizando a probabilodade a priori.

Mistura de Especialista: Formulação básica

- Um caso especial interessante é uma arquitetura na qual as redes especialistas e a rede *gating* são *modelos lineares* e a *densidade de probabilidade associada com* os especialistas é uma gaussiana com matriz de covariância igual à identidade

$$P(y^{(t)} \mid x^{(t)}, \theta_i) = \exp\left\{-\frac{1}{2}\left(y^{(t)} - \mu_i^{(t)}\right)^T \left(y^{(t)} - \mu_i^{(t)}\right)\right\}$$

$$h_i^{(t)} = \frac{g_i^{(t)} \exp\left\{-\frac{1}{2}\left(\mathbf{y}^{(t)} - \mu_i^{(t)}\right)^T \left(\mathbf{y}^{(t)} - \mu_i^{(t)}\right)\right\}}{\sum_{j=1}^{m} g_j^{(t)} \exp\left\{-\frac{1}{2}\left(\mathbf{y}^{(t)} - \mu_j^{(t)}\right)^T \left(\mathbf{y}^{(t)} - \mu_j^{(t)}\right)\right\}}.$$

$$\theta_i^{(k+1)} = \theta_i^{(k)} + \rho h_i^{(t)} (\mathbf{y}^{(t)} - \mu_i^{(t)}) \mathbf{x}^{(t)^T}$$

$$\mathbf{v}_i^{(k+1)} = \mathbf{v}_i^{(k)} + \rho (h_i^{(t)} - g_i^{(t)}) \mathbf{x}^{(t)^T}$$

- onde ρ é a taxa de aprendizado

Mistura de Especialista: Algoritmo EM

- Aprendizado EM
  - O algoritmo de aprendizado por maximização da esperança, proposto por DEMPSTER *et al.*(1977), é uma técnica geral para estimação do máximo da verossimilhança aplicado principalmente a aprendizado não-supervisionado, isto é, clusterização e estimação da mistura de densidades.

- JORDAN & JACOBS (1994) derivaram um algoritmo de maximização da esperança (EM, *expectation maximization na literatura em inglês*) *para estimação dos parâmetros das* arquiteturas MEs e HMEs.

- Este algoritmo é uma alternativa ao método do gradiente apresentado anteriormente e particularmente útil para modelos nos quais as redes especialistas e a rede *gating têm uma forma paramétrica simples*.

Mistura de Especialista: Algoritmo EM

- Cada iteração do algoritmo EM consiste de duas fases:
  - (1) uma propagação na árvore para calcular a probabilidade a posteriori (o passo E);
  - (2) uma solução de um conjunto de problemas de maximização da verossimilhança, ponderados nós não terminais e terminais da árvore (o passo M).

- O passo E (Esperança) define uma nova função de verossimilhança a cada iteração, a qual é maximizada durante o passo M (Maximização).

Mistura de Especialista: Algoritmo EM

- A aplicação do algoritmo EM para a arquitetura MEs envolve a definição de variáveis ausentes que facilitam a otimização do logaritmo da verossimilhança.

- Seja z_i, i =1,..., N, um conjunto de variáveis indicadoras binárias para a rede gating

- Para qualquer vetor de entrada x, exatamente um dos $z_{i's}$ vale 1, com todos os outros valendo 0.

- Observe, portanto, que os $z_{i's}$ *não é conhecido* e deve ser tratado como variável randômica

Mistura de Especialista: Algoritmo EM

- Anteriormente, temos

$$l(\chi,\Theta)=\sum_{j=1}^{N}\log\sum_{i=1}^{m}P(i\mid x_j,v)P(y_j\mid x_j,\theta_i)$$

- Podemos reescrever

$$l_c(\chi,\Theta)=\sum_{t=1}^{N}\log\prod_{i=1}^{m}P(i\mid x^{(t)},v)P(y^{(t)}\mid x^{(t)},\theta_i)^{z_i^{(t)}}$$

$$l_c(\chi,\Theta)=\sum_{t=1}^{N}\sum_{i=1}^{m}z_i^{(t)}\{\log P(i\mid x^{(t)},v)+\log P(y^{(t)}\mid x^{(t)},\theta_i)\}$$

$$l_c(\chi,\Theta)=\sum_{t=1}^{N}\sum_{i=1}^{m}z_i^{(t)}\log g_i^{(t)}+\sum_{t=1}^{N}\sum_{i=1}^{m}z_i^{(t)}\log P(y^{(t)}\mid x^{(t)},\theta_i)$$

Mistura de Especialista: Algoritmo EM

- É possível provar que as densidades de probabilidade a posteriori h_i podem ser usadas como valores esperados para as variáveis desconhecidas $z_{i's}$

- Usando este fato, define-se a função Q *para* o passo *E do algoritmo EM na forma:*

$$Q(\theta, \theta^k) = \sum_{t=1}^{N} \sum_{i=1}^{m} h_i^{(t)} \{\log g_i^{(t)} + \log P(y^{(t)} \mid x^{(t)}, \theta_i)\}$$

Mistura de Especialista: Algoritmo EM

- O passo M requer a maximização da função Q com respeito aos modelos dos parâmetros

- Nota-se agora o benefício da aplicação do algoritmo EM, já que a maximização divide o problema original em um conjunto de problemas de maximização separáveis, os quais podem ser solucionados independentemente no passo M, *como segue:*

$$v_i^{k+1} = \arg\max_{v_i} \sum_{t=1}^{N} \sum_{i=1}^{m} h_i^{(t)} \log g_i^{(t)}$$

$$\theta_i^{k+1} = \arg\max_{\theta_i} \sum_{t=1}^{N} \sum_{i=1}^{m} h_i^{(t)} \log P(y^{(t)} \mid x^{(t)}, \theta_i)$$

Mistura de especialistas gaussianos

- XU *et al. (1995) propuseram usar uma forma paramétrica baseada em modelos* gaussianos para a rede *gating.*

- FRITSH (1996) mostrou que a mesma forma paramétrica pode ser adotada também para as redes especialistas.

- Tal arquitetura é muito atrativa, pois há mecanismos eficazes para fornecer uma inicialização próxima a uma solução ótima, reduzindo assim o tempo de convergência do algoritmo de aprendizado (FRITSH, 1996).

Mistura de especialistas gaussianos

- Parametrização

$$g_i(\mathbf{x},\mathbf{v}) = \frac{\alpha_i P(\mathbf{x} \mid \mathbf{v}_i)}{\sum_{k=1}^{m} \alpha_k P(\mathbf{x} \mid \mathbf{v}_k)}, \text{ com } \sum_{k=1}^{m} \alpha_k = 1 \text{ e } \alpha_k \geq 0 \text{ e}$$

$$P(\mathbf{x} \mid \mathbf{v}_i) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\Sigma_i\right|^{\frac{1}{2}}} \exp\left\{\frac{1}{2}(\mathbf{x}-\gamma_i)^T \Sigma_i^{-1} (\mathbf{x}-\gamma_i)\right\}.$$

Mistura de especialistas gaussianos

- O algoritmo EM conduz ao seguinte método de estimação iterativo:

- Passo E – Para cada vetor de treinamento, calcule a probabilidade a posteriori h_i de acordo com:

$$h_i^{(k)}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)}) = \frac{\alpha_i^{(k)} P(\mathbf{x}^{(t)} \mid \mathbf{v}_i^{(t)}) P_i(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)}, \Theta_{(i)}^{(k)})}{\sum_{j=1}^{m} \alpha_j^{(k)} P(\mathbf{x}^{(t)} \mid \mathbf{v}_j^{(t)}) P_j(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)}, \Theta_{(j)}^{(k)})}$$

Mistura de especialistas gaussianos

- Passo *M* – *Use os $h_{i's}$ para calcular as novas estimativas para os parâmetros αi, v_i e Σi da* rede *gating. A nova estimativa pode ser computada diretamente, pois o problema de* maximização da verossimilhança é agora analiticamente solucionável, produzindo:

$$\alpha_i^{(k+1)} = \frac{\sum_{t=1}^{N} h_i^{(t)}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})}{\sum_{t=1}^{N}\sum_{k=1}^{m} h_k^{(t)}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})},$$

$$\gamma_i^{(k+1)} = \frac{\sum_{t=1}^{N} h_i^{(t)}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})\mathbf{x}^{(t)}}{\sum_{t=1}^{N} h_i^{(t)}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})},$$

$$\Sigma_i^{(k+1)} = \frac{\sum_{t=1}^{N} h_i^{(t)}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})\left[\mathbf{x}^{(t)} - \gamma_i^{(k+1)}\right]\left[\mathbf{x}^{(t)} - \gamma_i^{(k+1)}\right]^T}{\sum_{t=1}^{N} h_i^{(t)}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})}.$$

Mistura de especialistas gaussianos

- O problema de maximização da verossimilhança para as redes especialistas permanece sem solução analítica (no caso de classificação) e seus parâmetros devem ser estimados iterativamente pelo gradiente descendente ou por quadrados mínimos recursivos.

Variância adaptativa na mistura de especialistas

- Para MEs, é usual introduzir uma variância local para cada especialista (WEIGEND, 1995) dada por:

$$P(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)}, \theta_{ij}) = \frac{1}{(2\pi\sigma_j^2)^{\frac{d}{2}}} \exp\left\{ - \frac{\left\| \mathbf{y}^{(t)} - \mu_{ij}^{(t)}(\mathbf{x}) \right\|}{2\sigma_j^2} \right\}$$

$$\sigma_j^2 = \frac{1}{d} \frac{\sum_{t=1}^{N} h_j^{(t)} \left\| \mathbf{y}^{(t)} - \mu_j^{(t)} \right\|^2}{\sum_{t=1}^{N} h_j^{(t)}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})}$$

$$\frac{\partial Q^e(\theta^{k+1}, \theta)}{\partial \mu_{ijk}} = h_{ij}^{(t)} \frac{1}{\sigma_j^2} (\mathbf{y}_k^{(t)} - \mu_{ijk})$$

$$\frac{\partial Q(\theta^{k+1}, \theta)}{\partial \xi_i} = \sum_{t=1}^{N} \left(h_i^{(t)} - g_i^{(t)} \right)$$

Variância adaptativa na mistura de especialistas

- Passo E

$$P(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)}, \theta_{ij}) = \frac{1}{(2\pi\sigma_j^2)^{\frac{d}{2}}} \exp\left\{ - \frac{\left\| \mathbf{y}^{(t)} - \mu_{ij}^{(t)}(\mathbf{x}) \right\|}{2\sigma_j^2} \right\}$$

$$h_i^{(t)} == \frac{g_i^{(t)} P(y^{(t)} \mid x^{(t)}, \theta_i)}{\sum_{j=1}^{m} g_j^{(t)} P(y^{(t)} \mid x^{(t)}, \theta_j)}$$

- Passo M

$$\frac{\partial Q(\theta^{k+1}, \theta)}{\partial \mu_i} = \sum_{t=1}^{N} h_i^{(t)} \frac{1}{\sigma_i^2} (y^{(t)} - \mu_i^{(t)})$$

$$\sigma_j^2 = \frac{1}{d} \frac{\sum_{t=1}^{N} h_j^{(t)} \left\| \mathbf{y}^{(t)} - \mu_j^{(t)} \right\|^2}{\sum_{t=1}^{N} h_j^{(t)}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})}$$

$$\frac{\partial Q(\theta^{k+1}, \theta)}{\partial \xi_i} = \sum_{t=1}^{N} \left(h_i^{(t)} - g_i^{(t)} \right)$$

Resumo (Variações de MEs)

- Mistura tradicional (ME):
  - Introduzida por Jacobs et al. (1991);
  - Tanto os especialistas quanto a rede gating são modelos lineares, com exceção de que a rede gating possui função de ativação *softmax (divide o espaço de entrada com hiperplanos);*
- Mistura de Especialistas Gated (GE):
  - Introduzida por Weigend et al. (1995);
  - Emprega especialistas não-lineares;
  - Rede gating: MLP (divisão mais flexível do espaço de entradas);

Resumo (Variações de MEs)

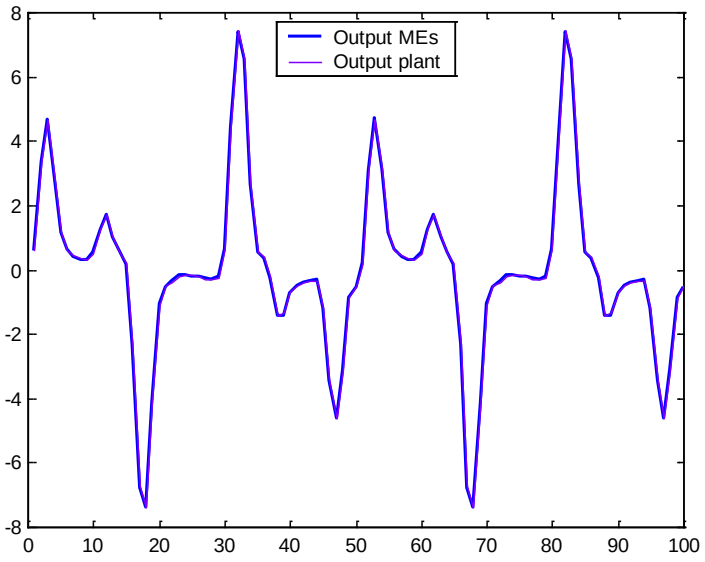
- Mistura de Especialistas Localizados (LME):
  - Introduzida por Xu et al. (1995);
  - Usa kernel gaussiano normalizado para a rede gating (divide o espaço de entradas com hiper-elipsóides suaves centrados nas regiões de atuação de cada especialista);
  - Especialistas são modelos lineares ou não-lineares;

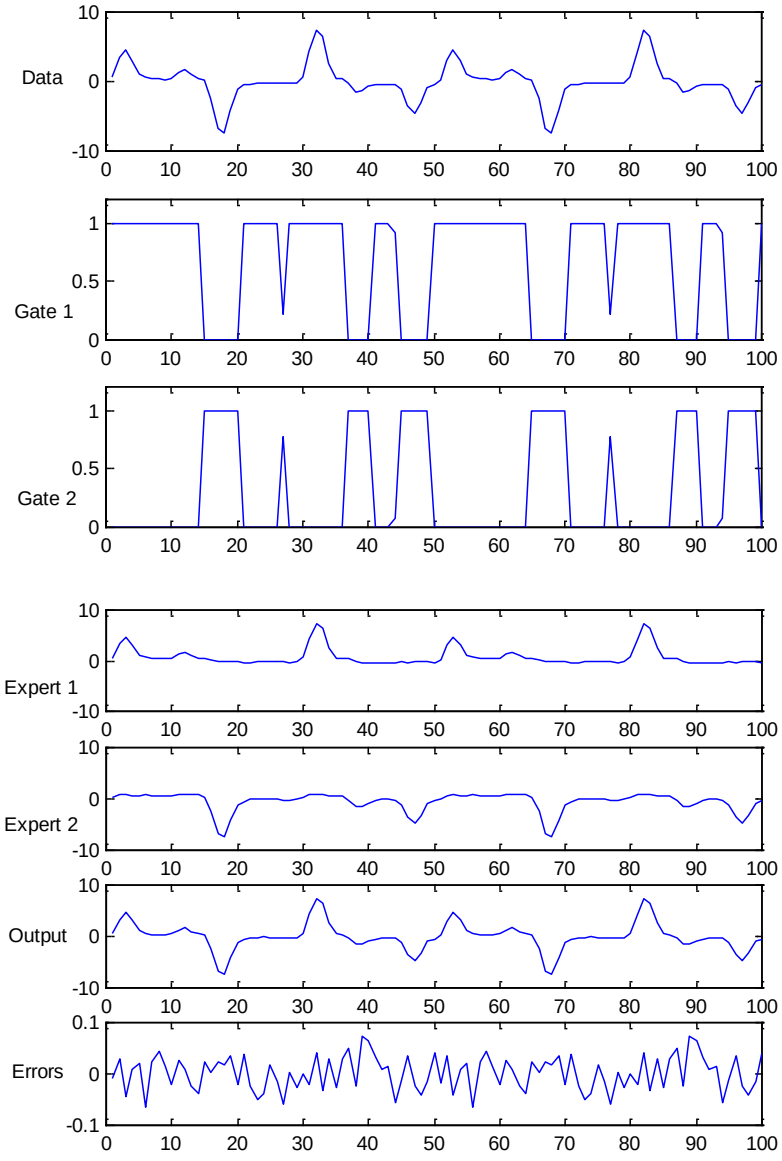
Experimento # 1

$$\hat{y}(k+1) = N[y(k), u(k)]$$

$$y(k+1) = \frac{y(k)}{1+y^2(k)} + u^3(k)$$

$$u(k) = sin(2\pi k/25) + sin(2\pi k/30)$$



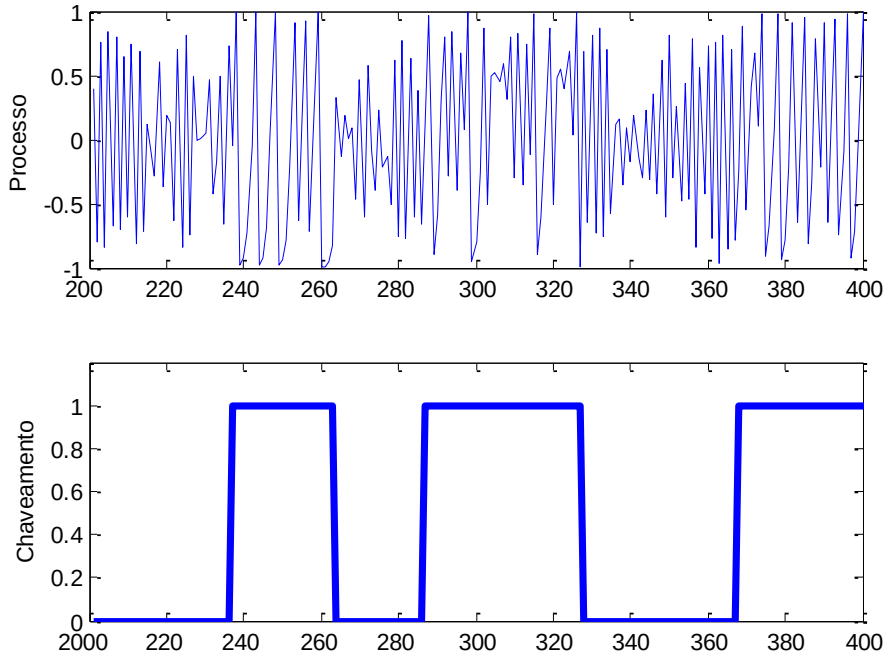


Plant versus GE behavior (first test set)

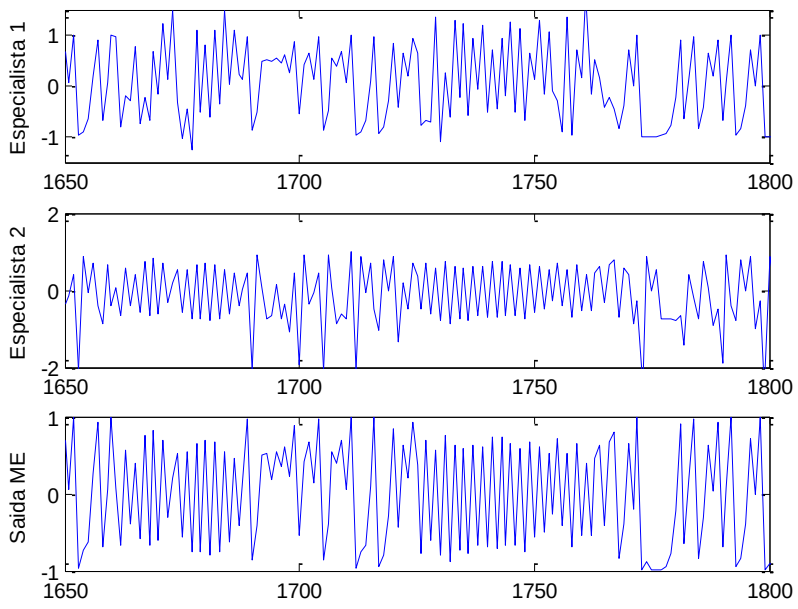
Experimento # 2 - Chaveamento entre processos

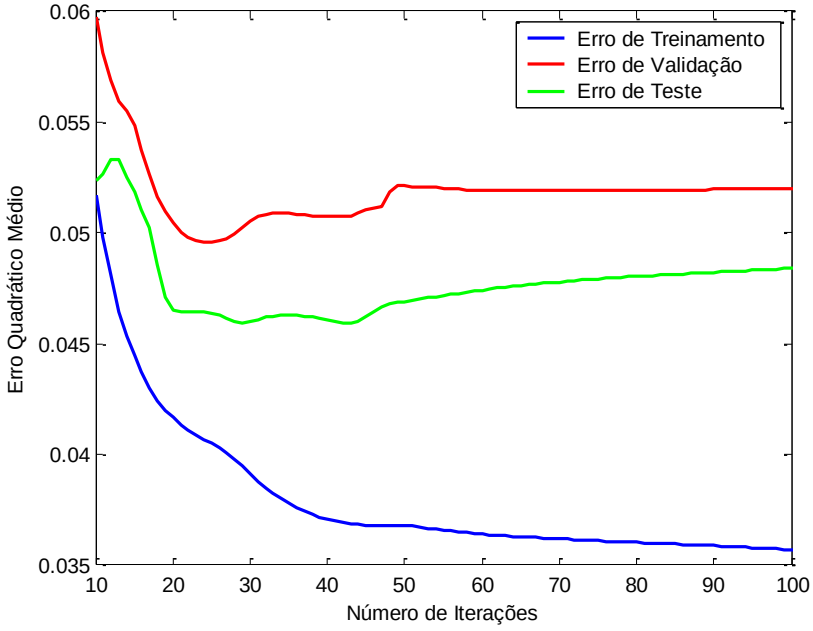
$$x^{(t+1)} = 2*(1-(x^{(t)})^2)-1$$, se chaveamento = 1

$$x^{(t+1)} = \tanh(-1.2*x^{(t)} + \xi^{(t+1)}); \xi \sim N(0,0.1)$$, se chaveamento =0

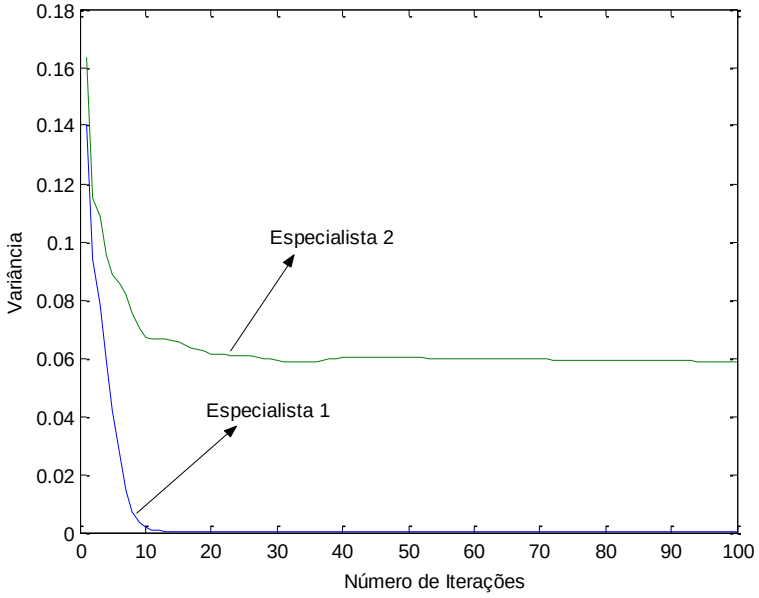


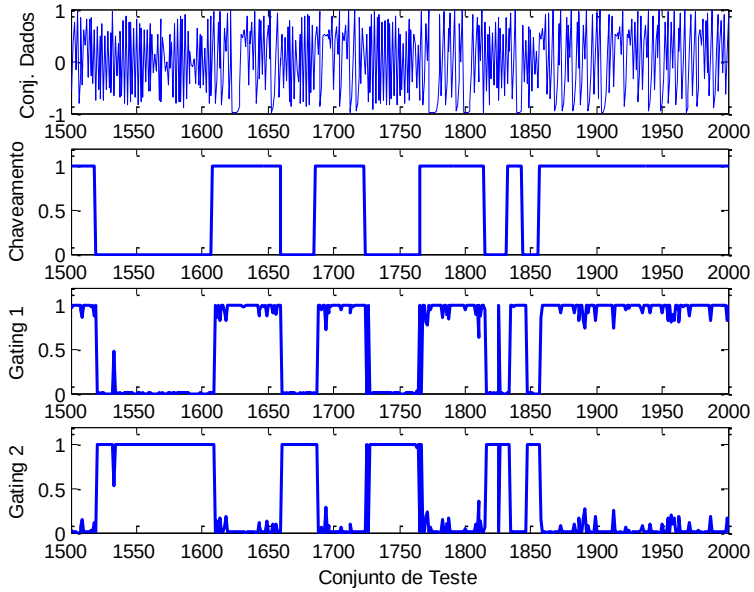
Experimento # 2 - Chaveamento entre processos



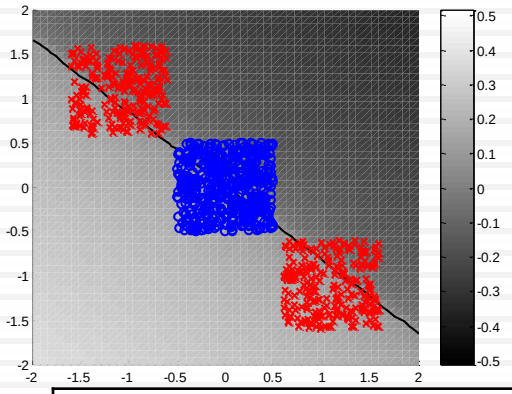


Experimento # 2 - Chaveamento entre processos

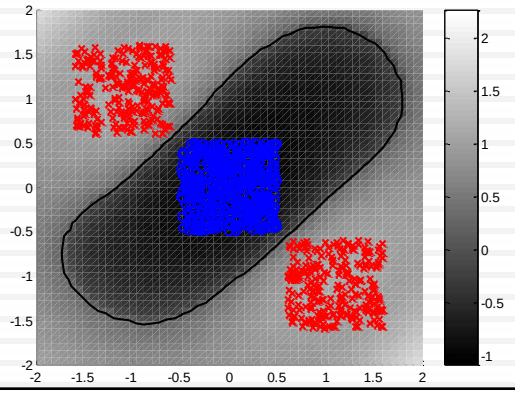




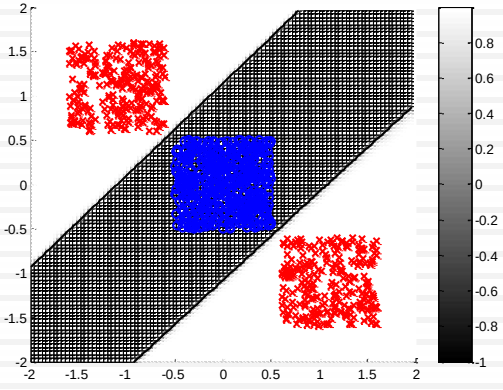
Experimento # 3 – Classificação binária



LS-SVM com Kernel Linear

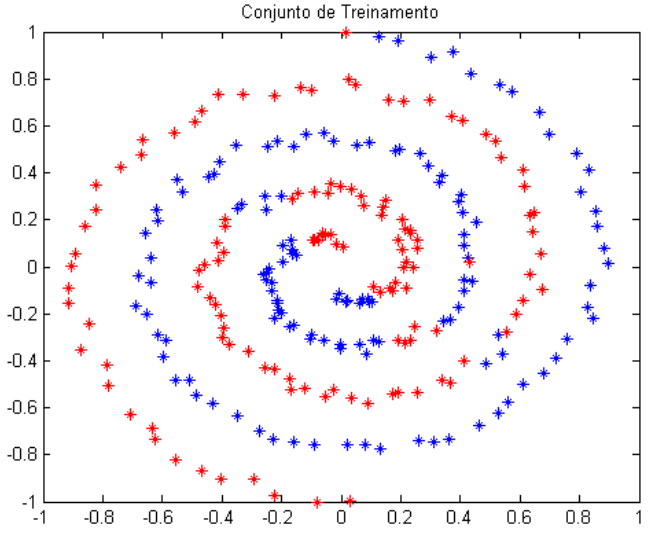


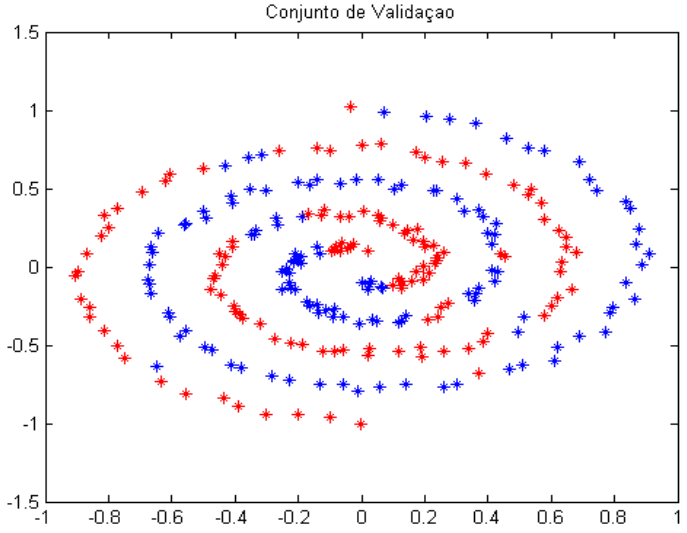
LS-SVM com kernel RBF

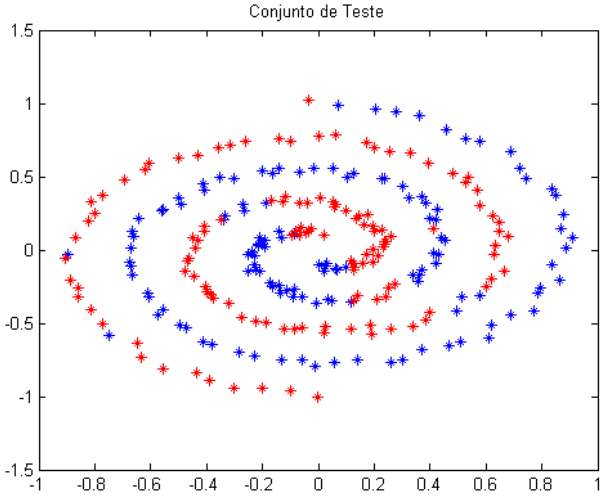


Mistura de LS-SVM com kernel Linear

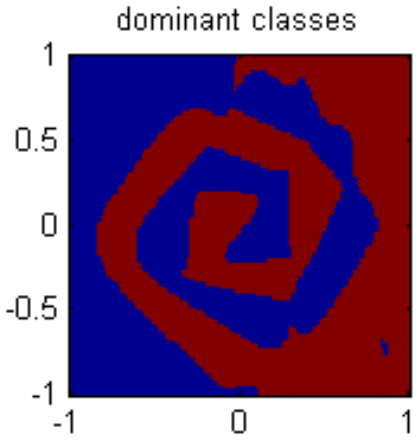
Experimento # 4 - Spiral (Rede Neural)

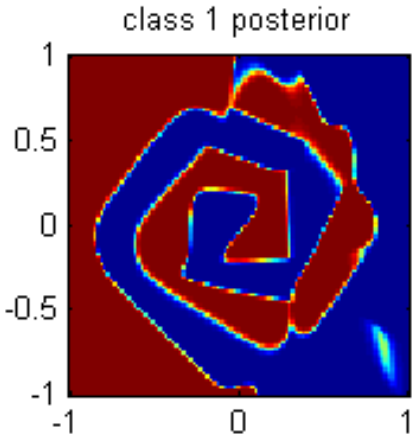


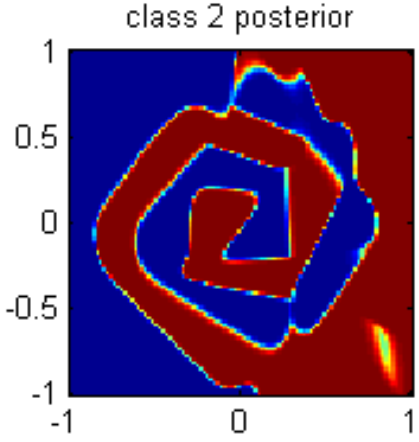




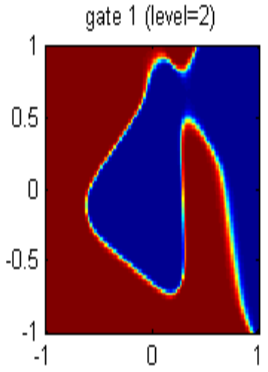
Experimento # 4 - Spiral (MEs)

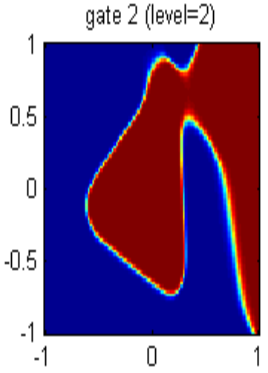


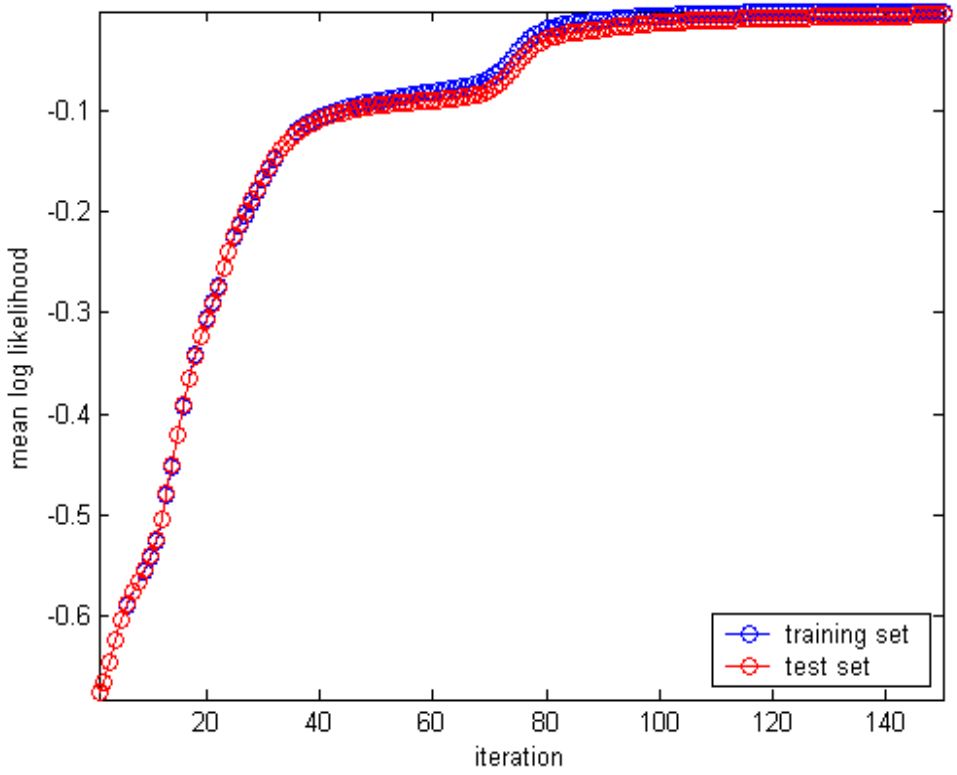




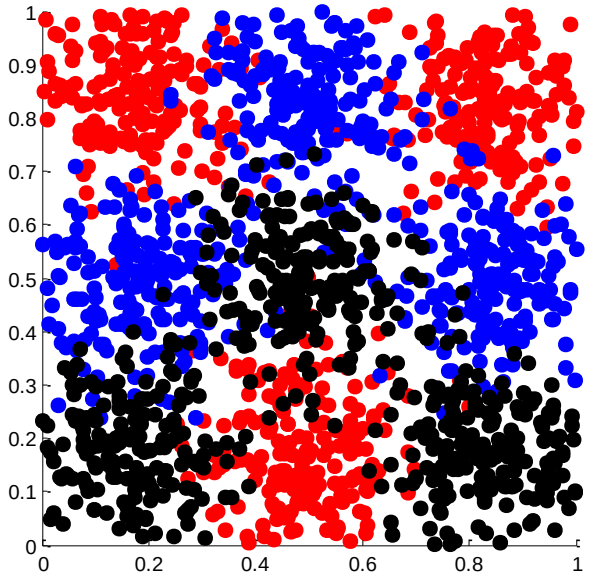
Experimento # 4 - Spiral (Rede Neural)

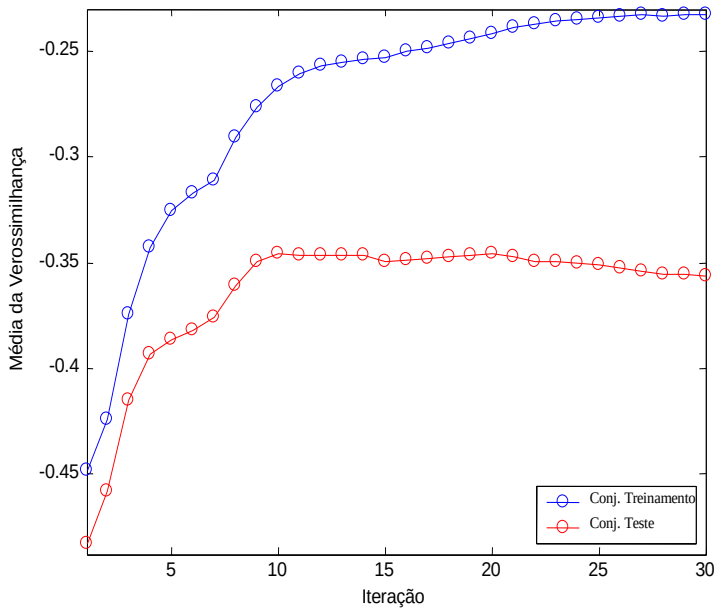




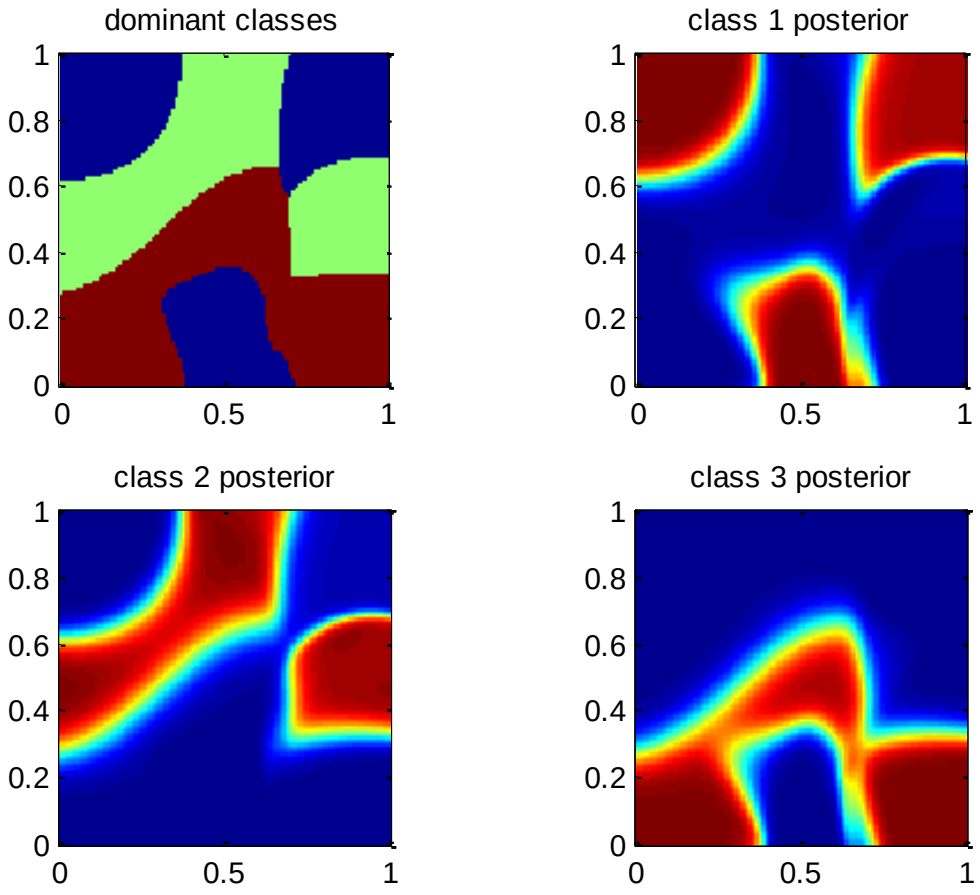


Experimento n°5 (Classificação com várias classes)





Experimento nº5 (Classificação com várias classes)



Experimento nº5 (Classificação com várias classes)