Pesquisas em Andamento

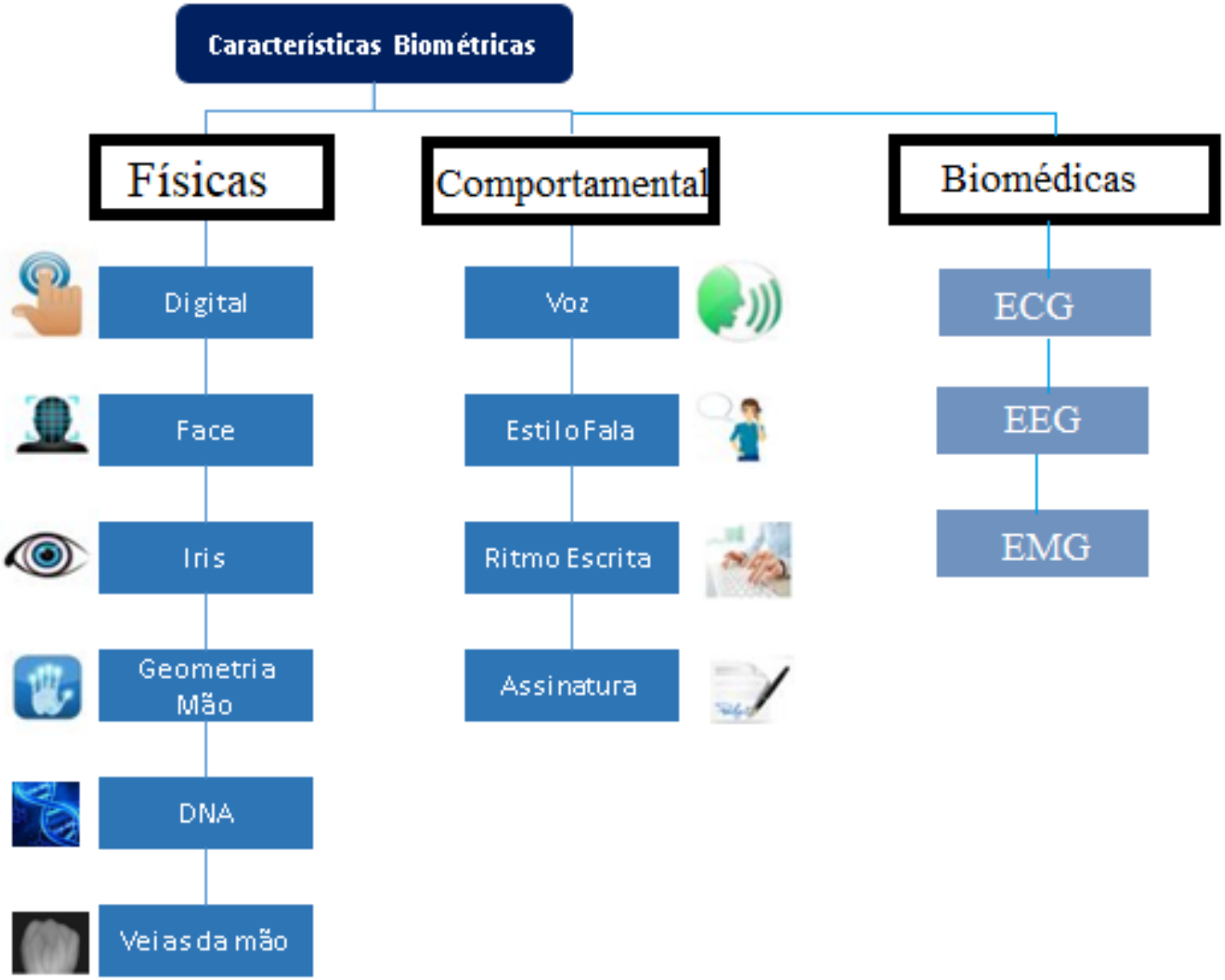
Clodoaldo A. M. Lima

19 de agosto de 2021

Definição

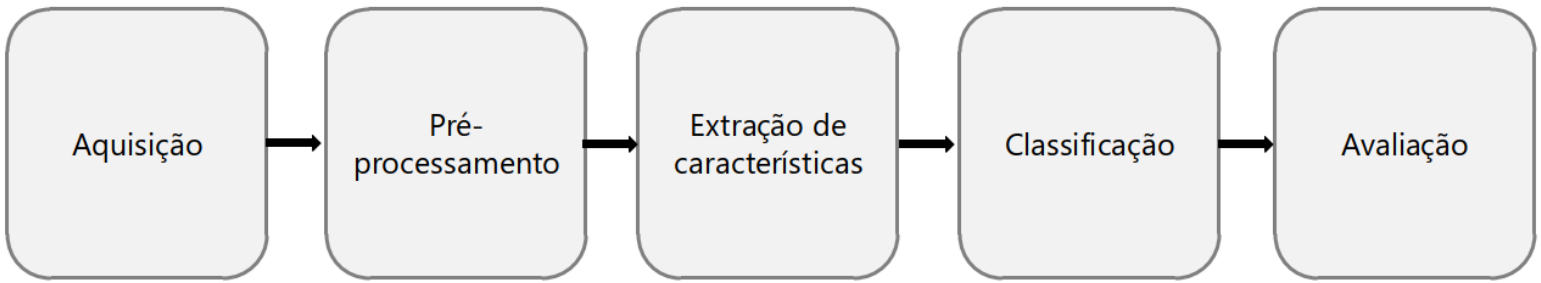
Biometria é uma palavra de origem grega que significa *'medida da vida'* e consiste em realizar medidas de traços humanos, que podem ser tanto físicos quanto comportamentais.

Modalidades Biométricas

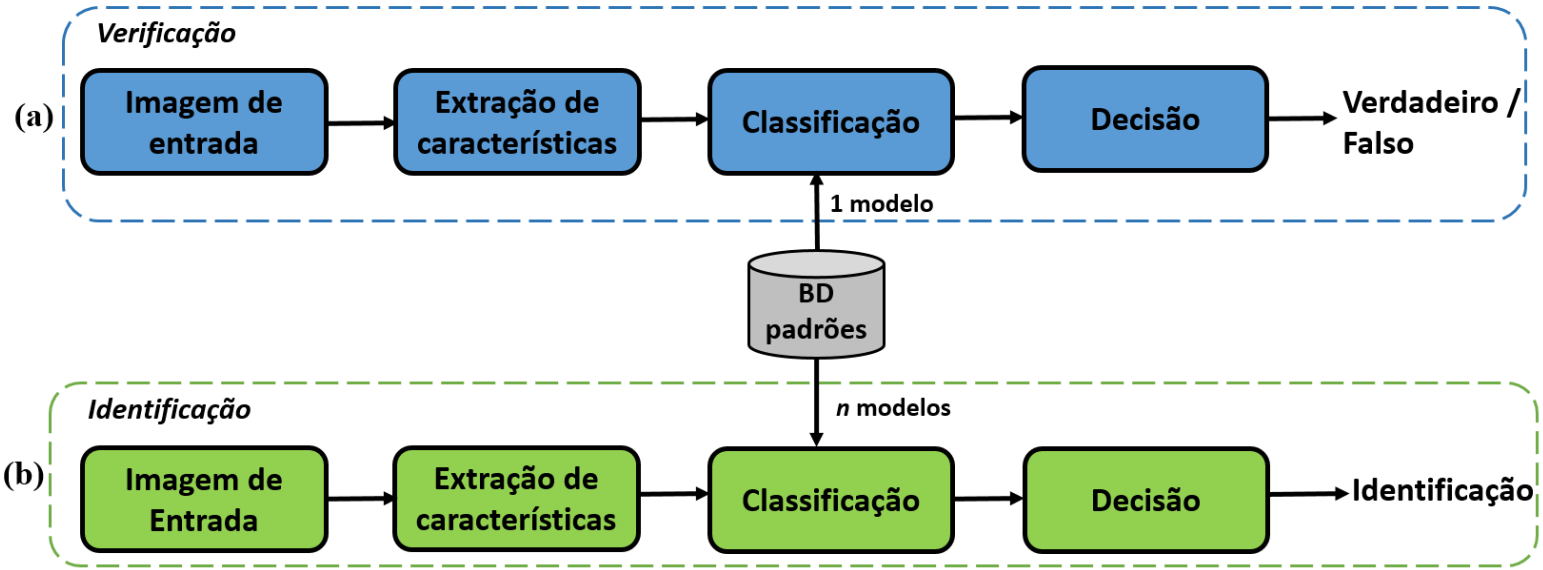


Reconhecimento Biométrico Tradicional

Diagrama de blocos

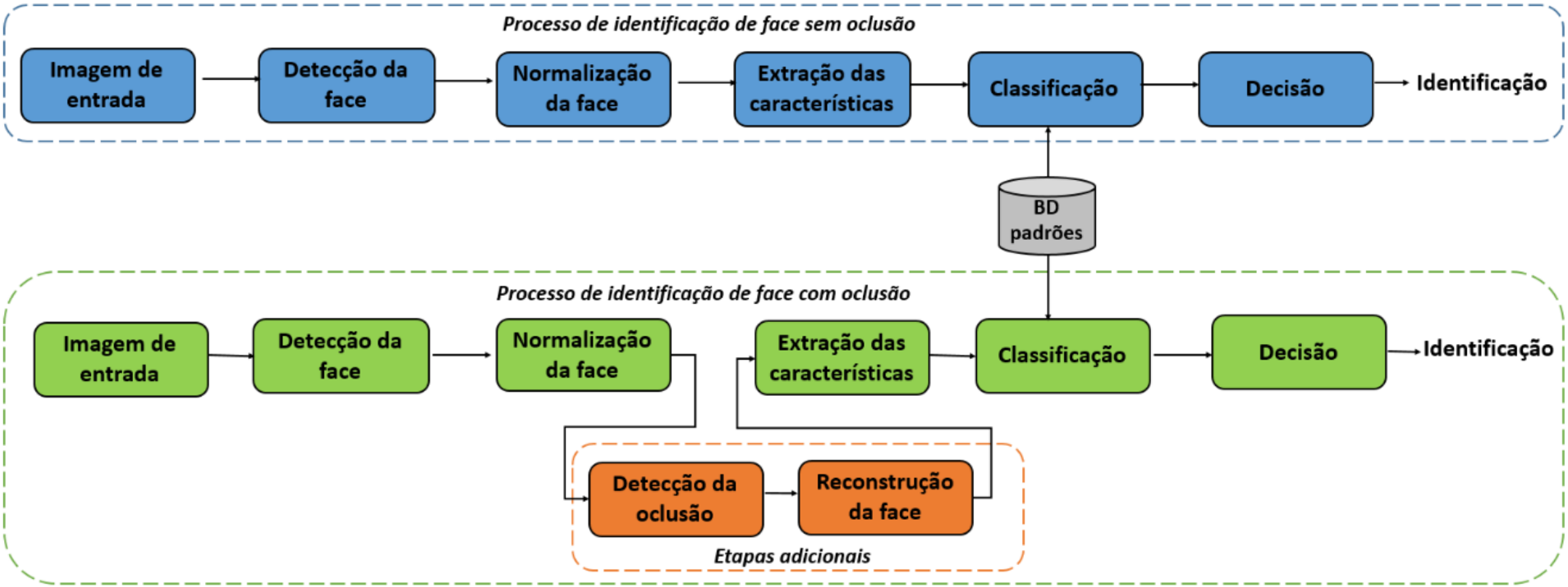


Verificação versus identificação



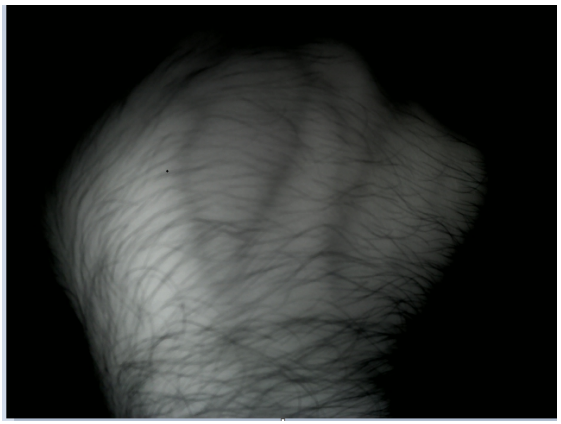
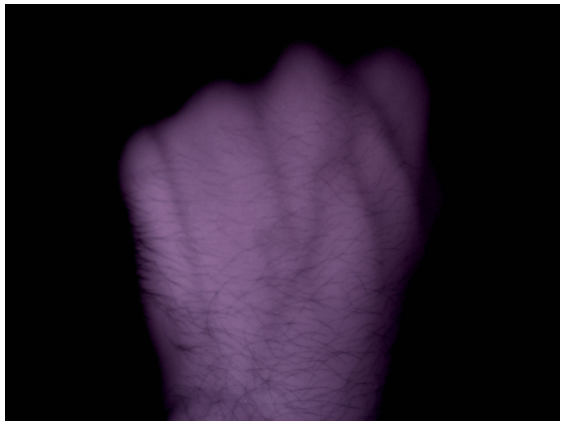
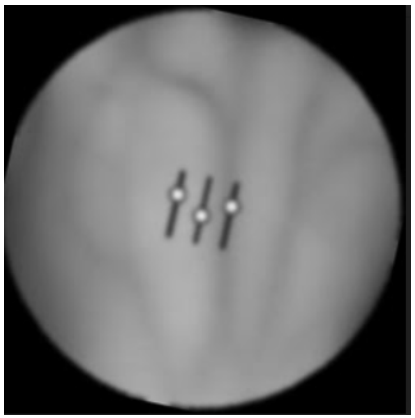
Tipos de variações

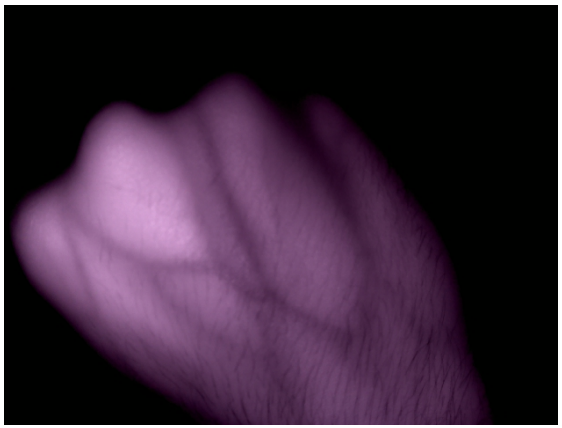
Reconhecimento facial com oclusão



Reconhecimento baseado no dorso da mão

Existem diferentes desafios que são encontrados em imagens de veias do dorso da mão:

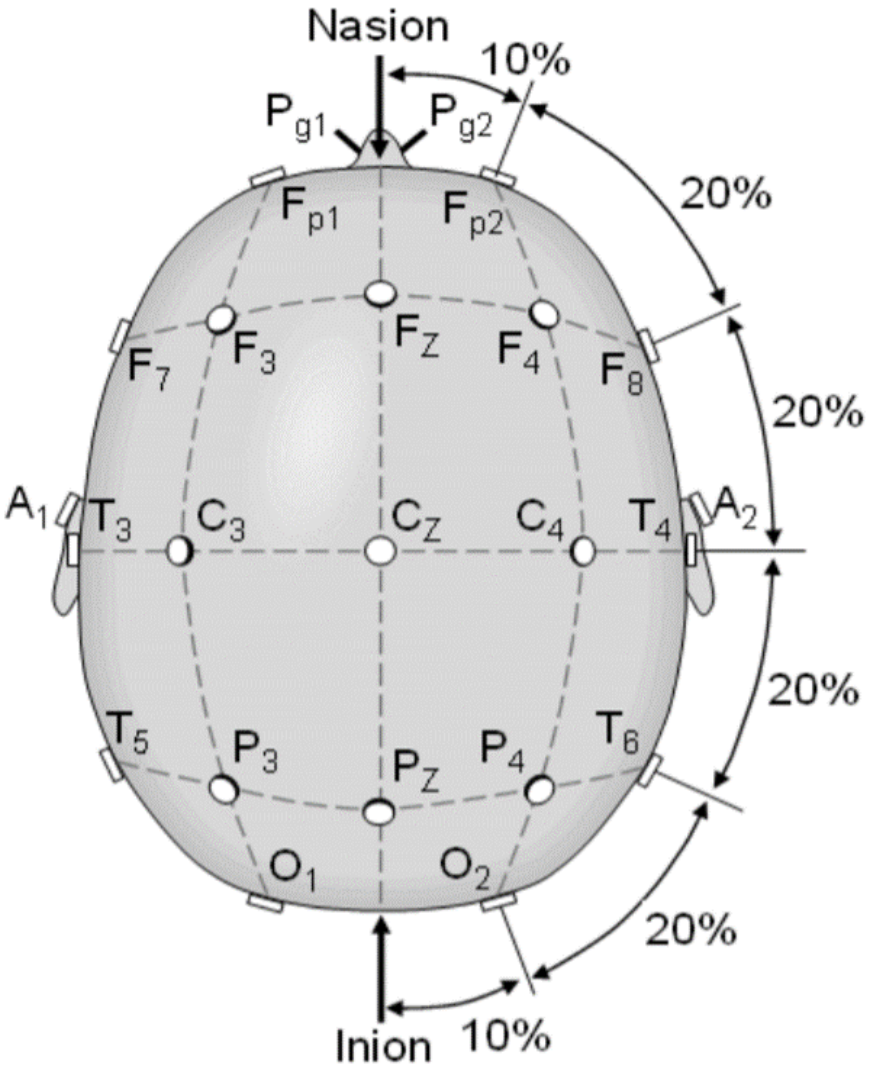


Forma de Captura

Composição do aparato:

- Eletrodos de Prata/Prata-Clorada com o diâmetro de 1 à 3 mm
- Amplificadores
- Dispositivo para aquisição de dados
- Sistema de processamento.

Sistema 10-20

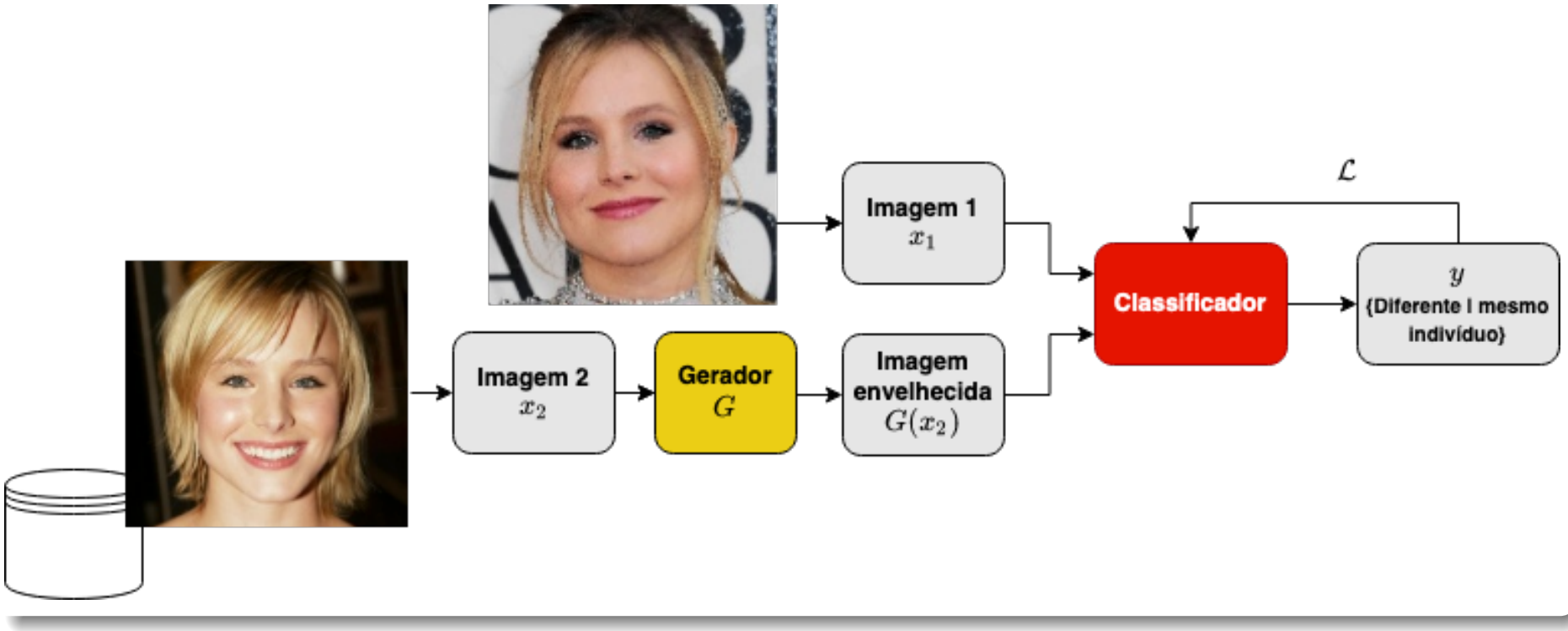


Envelhecimento

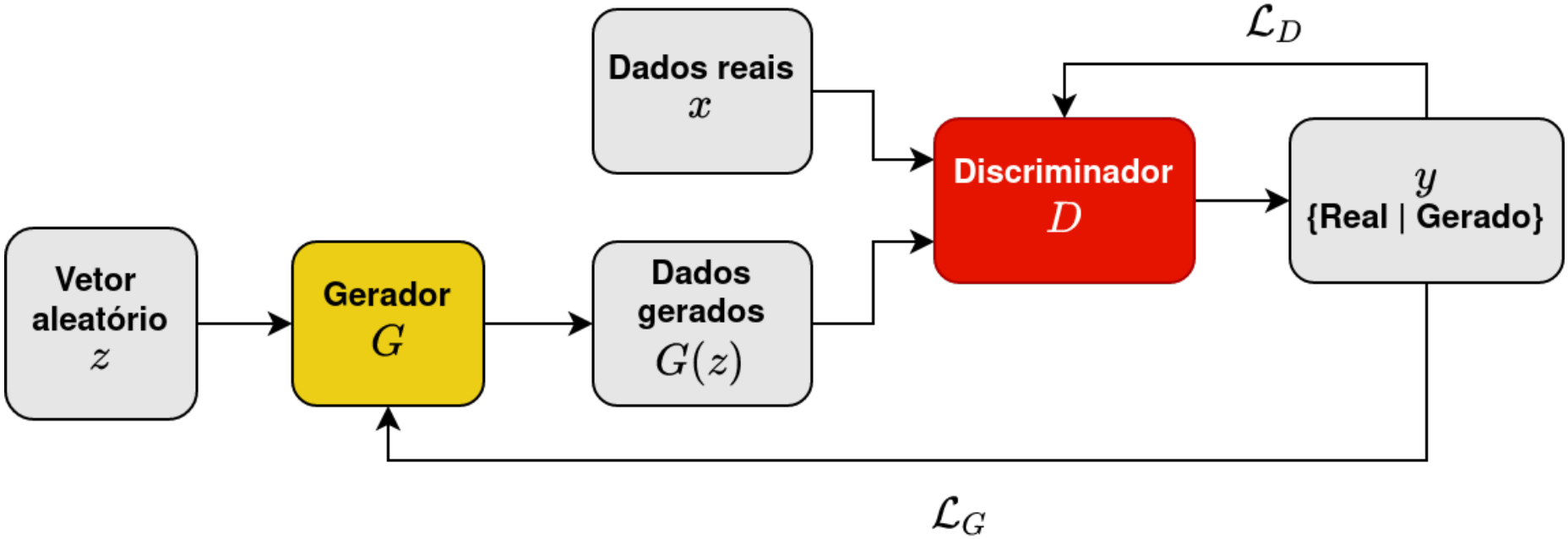
Envelhecimento facial

Os métodos computacionais de envelhecimento facial tem como objetivo gerar uma face envelhecida mantendo as caracterı´sticas individuais.

Motivação - Biometria



Redes adversárias generativas *GANs*



$$\min_G \max_D \mathbb{E}_{x \sim q_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z))]$$

GANs

Figura: Evolução dos resultados das redes.

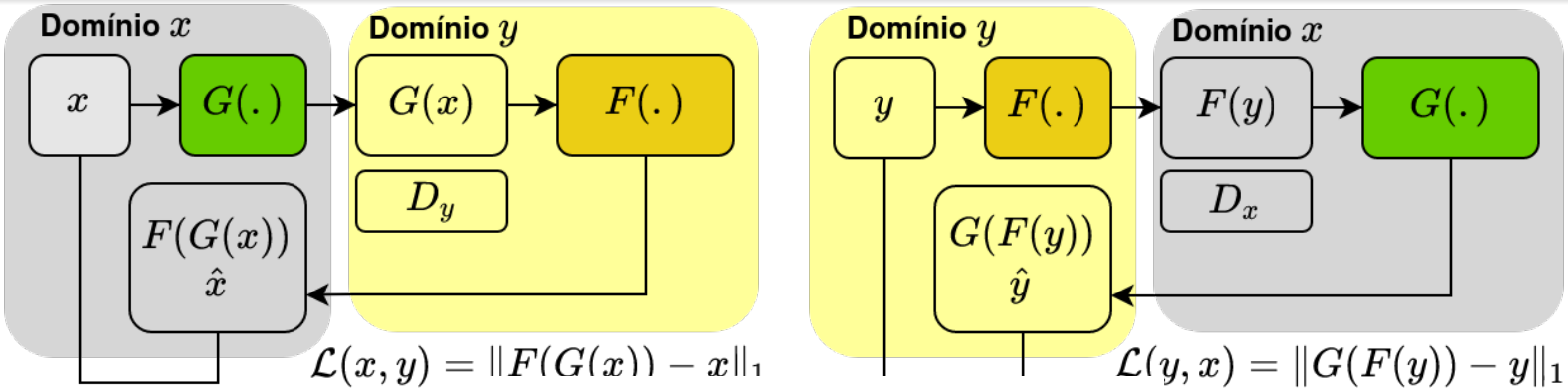
GANs for Good- A Virtual Expert Panel by DeepLearning.AI, 2020.

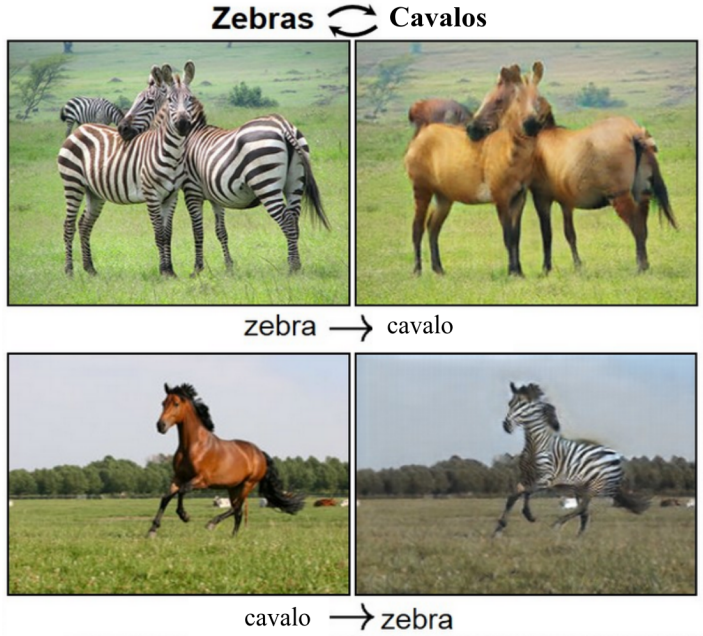
GANs

Redes adversárias generativas condicionais, *cGANs*

$$\min_G \max_D \mathbb{E}_{x|c\sim q_{data}(x|c)}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z|c)))]$$

CycleGAN


Domínio x
x
$G(.)$
$F(G(x))$
$\hat{x}$
Domínio y
$G(x)$
$F(.)$
D_y
$\mathcal{L}(x, y) = \|F(G(x)) - x\|_1$
Domínio y
y
$F(.)$
$G(F(y))$
$\hat{y}$
Domínio x
$F(y)$
$G(.)$
D_x
$\mathcal{L}(y, x) = \|G(F(y)) - y\|_1$


Zebras ⇄ Cavalos
zebra → cavalo
cavalo → zebra

Aula 01 - Conceitos Básicos

Clodoaldo A. M. Lima

19 de agosto de 2021

Programa de Pós-Graduação em Sistemas de Informação
Mestrado acadêmico - EACH - USP
http://ppgsi.each.usp.br

Sumário

Introdução

Objetivo de aprendizado de Máquina

O objetivo do aprendizado de máquina é construir modelos computacionais que podem adaptar-se e aprender a partir da experiência (MITCHELL,1997).

Aprendizado Indutivo - Segundo MITCHELL (1997)

"Um programa de computador aprende a partir de um elenco de experiências E, relacionadas a uma classe de tarefas T e dispondo de uma medida de desempenho M, se seu desempenho medido por M junto à tarefa T melhora com o elenco de experiências E."

Exemplo

No contexto de redes neurais artíciais, o seu processo de treinamento pode, então, ser caracterizado como aprendizado indutivo, sendo que o uso posterior da rede neural treinada para classificação, regressão ou agrupamento de dados é geralmente denominado de processo de inferência dedutiva.

Tipos de Raciocínio

Raciocínio Indutivo

- Conhece: p(a,b), p(a,d), p(d,e), p(d,g), p(e,f)
- Observa: a(a,e) e a(d,f)
- Aprende: $p(X,Y) \cap a(Y,Z) \rightarrow p(X,Z)$

Raciocínio Abdutivo

- Conhece: $p(a,b), p(a,d), p(d,e), p(d,g), p(e,f), p(X,Y) \cap p(Y,Z) \rightarrow a(X,Z)$
- Observa: a(a,c)
- Explica: p(b,c) ou p(d,c)

Raciocínio Dedutivo

- Conhece: $p(a,b), p(a,d), p(b,c), p(d,e), p(d,g), p(e,f), p(X,Y) \cap p(Y,Z) \rightarrow$ a(X,Z)
- Conclui: a(a,c), a(a,e), a(a,g) e a(d,f)

Aprendizado de Máquina

Em termos práticos, algoritmos de aprendizado de máquina têm como objetivo descobrir o relacionamento entre as variáveis de um sistema (entrada/saída) a partir de dados amostrados (CHERKASSKY & MULIER, 2007).

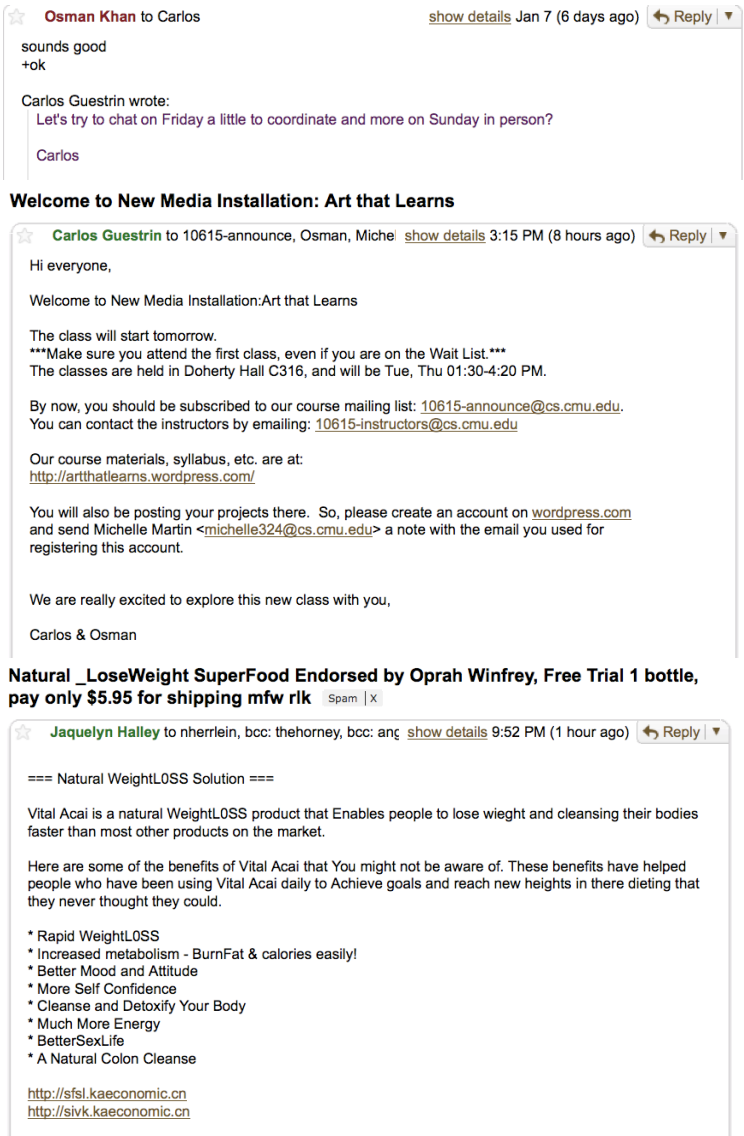
Sendo assim, eles não são necessários quando os relacionamentos entre todas as variáveis do problema (entrada/saída) são completamente compreendidos. Este definitivamente não é o caso de muitos dos problemas reais com os quais nos defrontamos em nosso dia-a-dia.

Há três frentes principais em aprendizado de máquina

- Aprendizado supervisionado, que será o centro de atenção deste curso;
- Aprendizado por reforço, que não será abordado formalmente, pois foge ao escopo do curso;
- Aprendizado não supervisionado, abordado especificamente em alguns tópicos.

Classificação

Dados



Osman Khan to Carlos show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
> Let's try to chat on Friday a little to coordinate and more on Sunday in person?
>
> Carlos

Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Miche show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Our course materials, syllabus, etc. are at:
http://artthatlearns.wordpress.com/

You will also be posting your projects there. So, please create an account on wordpress.com and send Michelle Martin <michelle324@cs.cmu.edu> a note with the email you used for registering this account.

We are really excited to explore this new class with you,

Carlos & Osman

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | X

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: anç show details 9:52 PM (1 hour ago) Reply

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

http://sfsl.kaeconomic.cn
http://sivk.kaeconomic.cn

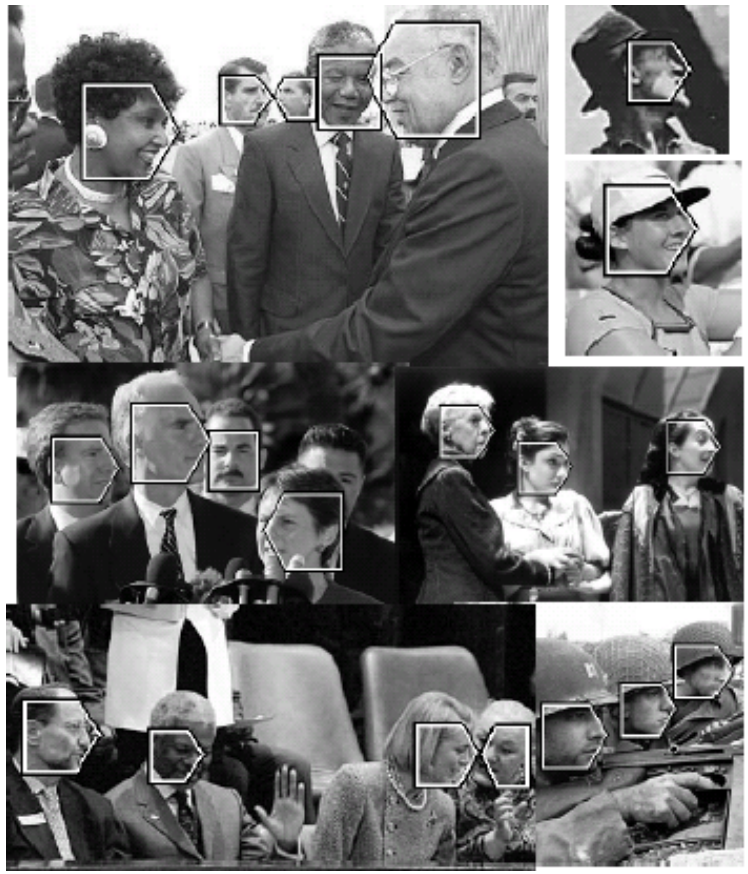
Predição

Spam vs Não Spam

Classificação


Figure 1. Examples of training images for each face orientation

Reconhecimento Facial



Classificação

Predição do tempo

Regressão

Predição do valor da ação

Regressão

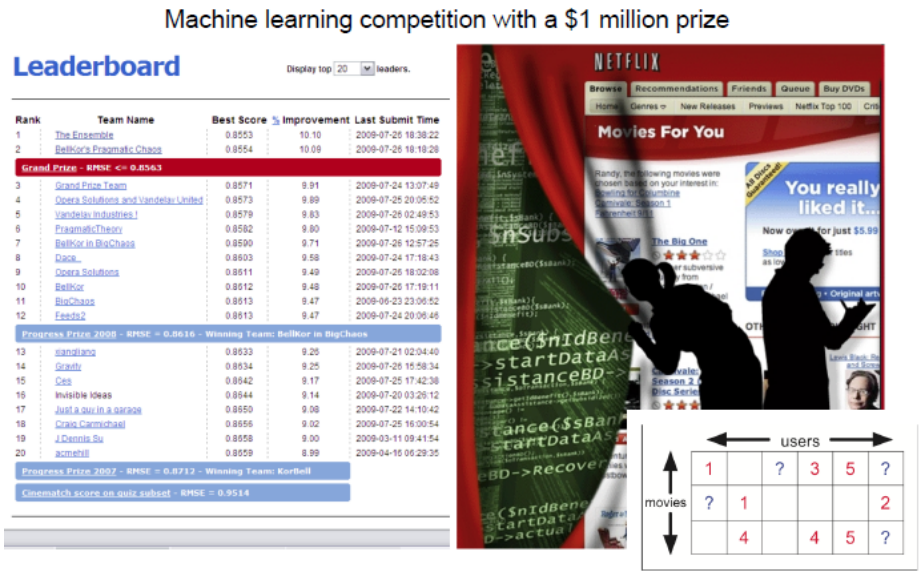
Predição do tempo

Filtragem Colaborativa

Sistema de Recomendação

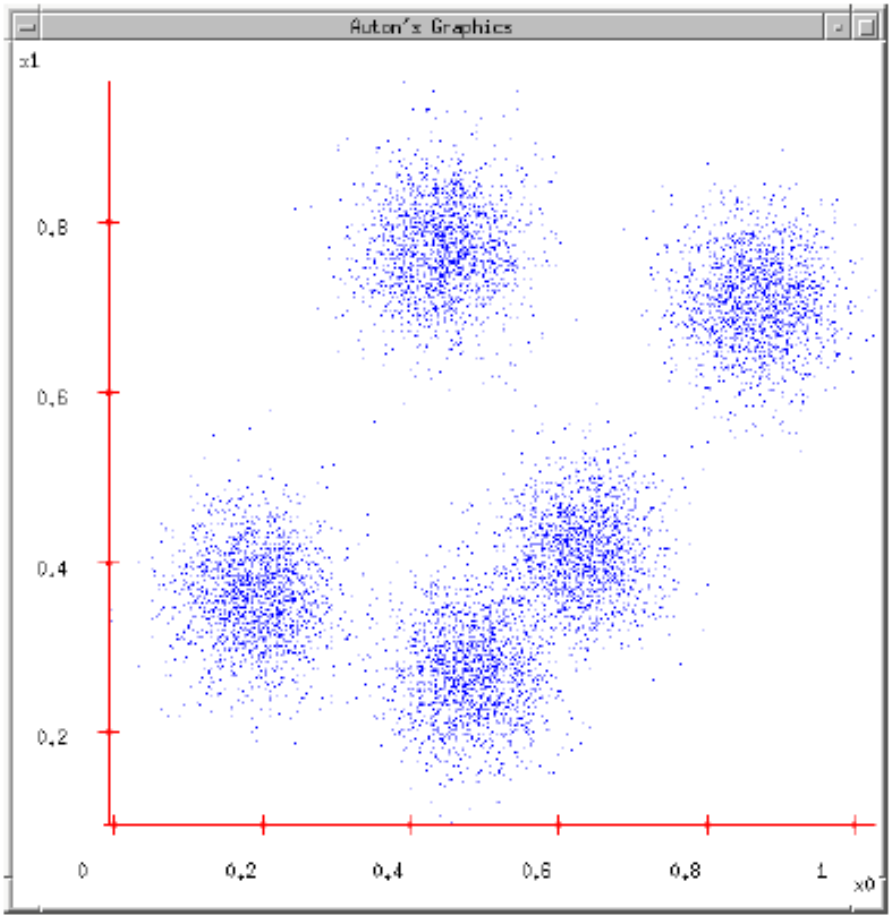
Filtragem Colaborativa

Sistema de Recomendação



Agrupamento - Clusterização

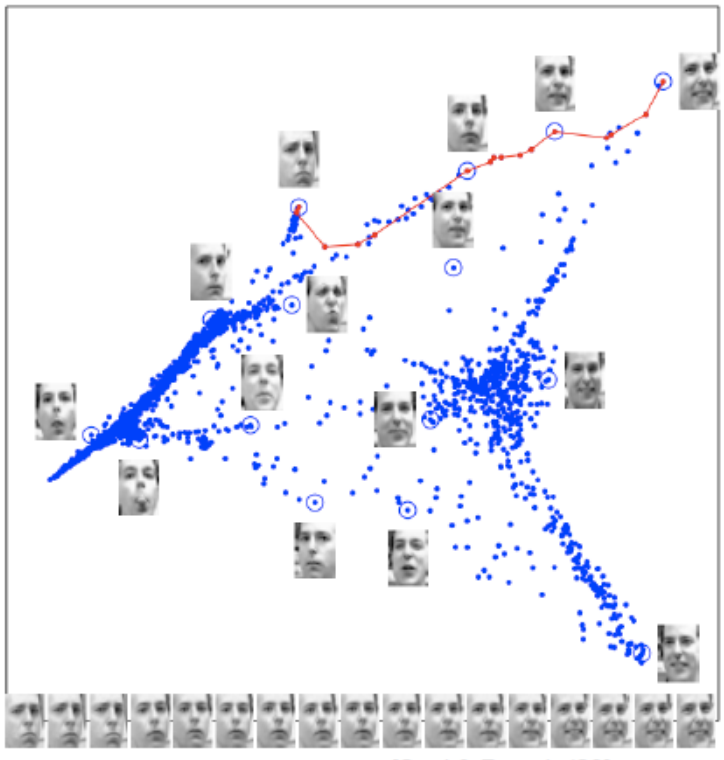
Agrupar coisas similares



Visualização

Imagens possuem pixels. Nós podemos fornecer uma coordenada tal que imagens similares estejam perto uma das outras?

Reconhecimento Facial



Crescimento de Aprendizado de Máquina

Aprendizado de Máquina e uma abordagem muito utilizada para

- Reconhecimento de fala, processamento de linguagem natural
- Visão computacional
- Análises médica
- Biologia computacional
- Redes de sensores

Esta tendência e acelerada por

- Big Data
- Melhoria de Algoritmos de aprendizado de máquina
- Computadores mais rápido

Aprendizado Supervisionado

Definição

- Dado um conjunto de treinamento $f(x_i; y_i)i = 1; \cdots ; N$
- Encontrar uma boa aproximação para $f : XY$

Exemplos: O que representam X e Y

- Detecção Spam
  - Mapear texto para (Spam, Não-Spam)
- Reconhecimento de Digito
  - Mapear pixels para 0,1,2,3,4,5,6,7,8,9
- Predição de Ações
  - Mapear preços históricos para $\Re$ (número real)

Problema de Aprendizado de Máquina

Conjunto de Dados

Exemplo	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Nosso objetivo é encontrar uma função
$f : XY$
$X = \{0, 1\}^4$
$Y = \{0, 1\}$

Questão 1

Como definir o espaço de hipótese, o conjunto possíveis de f

Questão 2

Como encontrar melhor f no espaço de hipotese

Espaço de Hipoteses mais gerais

Considere todas as possíveis funções booleanas sobre 4 características como entrada

2^{16} hipoteses possíveis
2^9 são consistente com nosso conjunto de dados.
Como escolher a melhor hipótese?

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Conjunto de Dados

Exemplo	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Espaço de Hipoteses mais gerais

Considere todas as funções booleanas conjuntivas

16 hipoteses possíveis
nenhuma é consistente com nosso conjunto de dados
Como escolher a melhor hipótese?

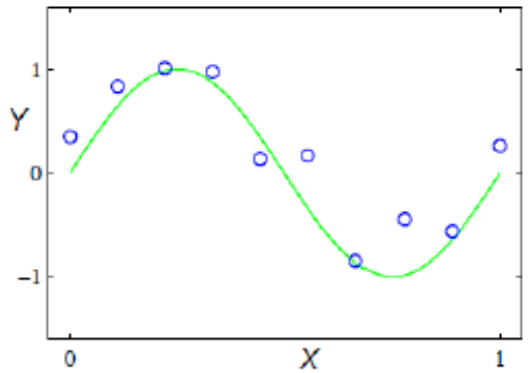
Rule	Counterexample
$\Rightarrow y$	1
$x_1 \Rightarrow y$	3
$x_2 \Rightarrow y$	2
$x_3 \Rightarrow y$	1
$x_4 \Rightarrow y$	7
$x_1 \wedge x_2 \Rightarrow y$	3
$x_1 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \Rightarrow y$	3
$x_2 \wedge x_4 \Rightarrow y$	3
$x_3 \wedge x_4 \Rightarrow y$	4
$x_1 \wedge x_2 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3

Conjunto de Dados

Exemplo	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Regressão

Conjunto de dados: 10 pontos (X, Y) gerados a partir da função seno com ruído



Regressão

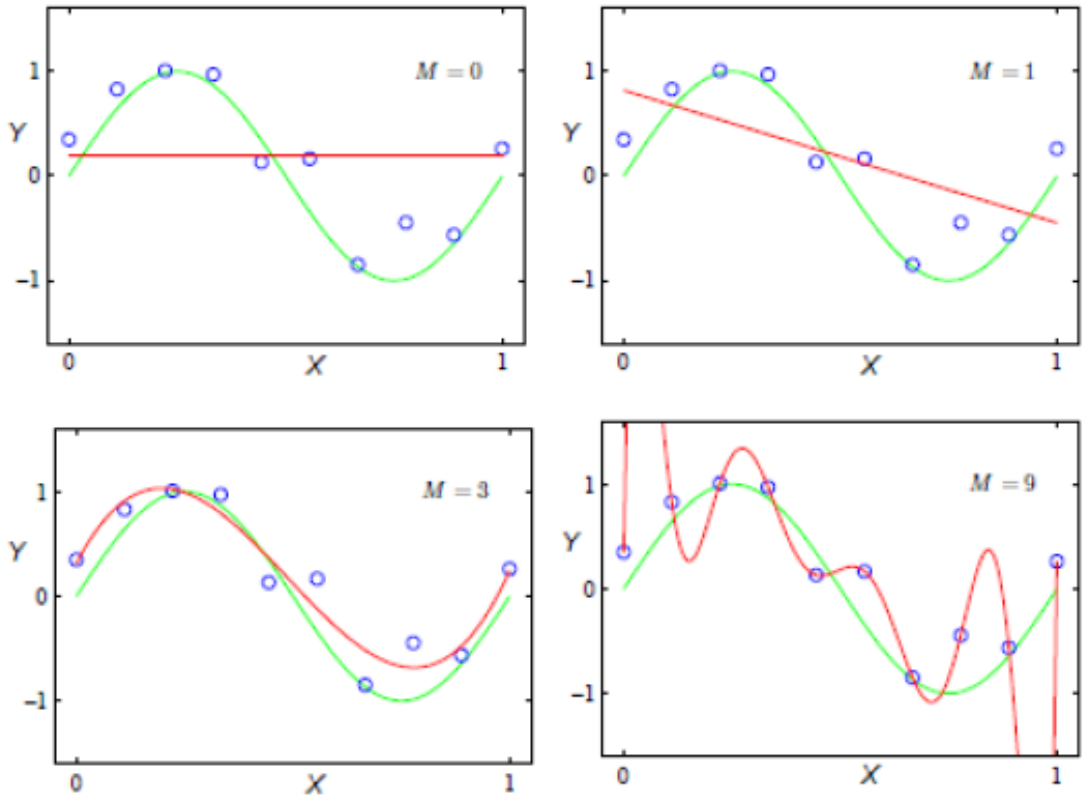
$$f : X \to Y$$

$$X = \mathcal{R}$$

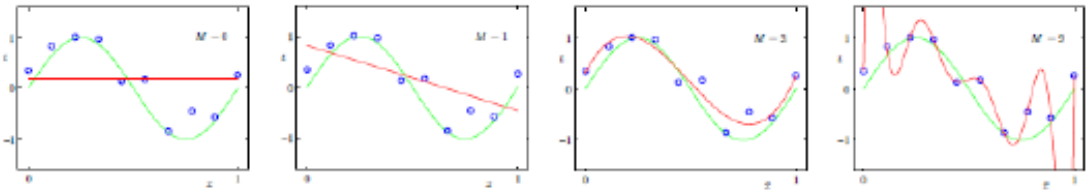
$$Y = \mathcal{R}$$

Regressão

Qual é o melhor grau do polinômio M?



Regressão

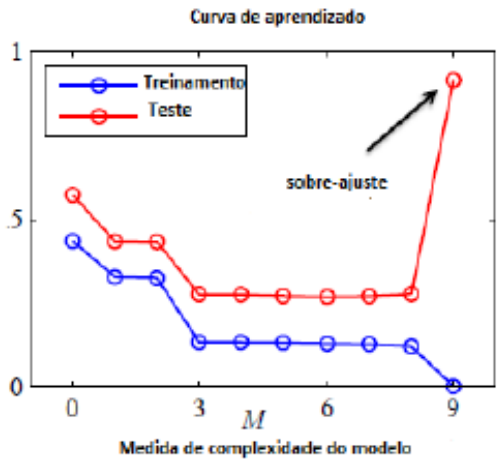


Nós medimos o erro usando uma função perda $L(y, \hat{y})$
Para regressão, uma escolha comum é perda quadrada:

$$L(y_i, f(x_i)) = (y_i - f(x_i))^2$$

A perda empírica da função f aplicada para dados de treinamento é então

$$\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$



Princípio de Occam's Razor

William of Occam: Monge viveu no século 14

Principio da parcimônia:

"One should not increase, beyond what is necessary, the number entities required to explain anything"

- Quando várias soluções estão disponíveis para uma dado problema, nós devemos selecionar aquela mais simples
- Mas o que nós queremos dizer por simples?
- Nós usaremos o conhecimento a priori do problema para solucionar e definir o que é uma solução simples
- Exemplo de um conhecimento a priori: suavidade

Questões chave em Aprendizado de Máquina

- Como vamos escolher um espaço de hipótese?
  - Frequentemente nós usamos conhecimento a piori para guiar esta escolha
- Como nós podemos avaliar a precisão de uma hipótese sobre dados não vistos?
  - Occam's razor: usa a hipótese mais simples consistente com dados. Isto ajudaria a evitar o sobre-ajuste
  - Teoria de Aprendizado vai nos ajudar a quantificar a capacidade de generalização como uma função da quantidade de dados de treinamento e o espaço de hipótese.
- Como iremos encontrar a melhor hipótese?
  - Esta é uma questão algorítmica, o tópico principal da ciência da computação
- Como modelar aplicações como problema de aprendizado de máquina? (Desafio de engenharia)

Conceitos

Hipótese Suficiente

Uma hipótese é chamada de suficiente se e somente se ela tem valores 1 para todos os exemplos de treinamento rotulados por um 1

Hipótese Necessária

Uma hipótese é chamada de necessária se e somente se ela tem valores 0 para todos os exemplos de treinamento rotulados por um 0

Hipótese consistente

Uma hipótese é consistente com o conjunto de treinamento se ambas suficiente e necessária

Conceitos

Capacidade de generalização

Definida como a capacidade do classificador de prever corretamente a classe de novos dados

Sobre-ajuste - Overfitting

No caso em que o modelo se especializa nos dados utilizados em seu treinamento, apresentando uma baixa taxa de acerto quando confrontado com novos dados, tem-se a ocorrência de um superajuste (overfitting).

Sub-ajuste - Underfitting

E também possível induzir hipóteses que apresentem uma baixa taxa de acerto mesmo no subconjunto de treinamento, configurando uma condiçãp de subajuste (underfitting).

Essas situações podem ocorrer, por exemplo, quando os exemplos de treinamento disponíveis são pouco representativos ou quando o modelo obtido é muito simples

Erro verdadeiro de uma hipótese

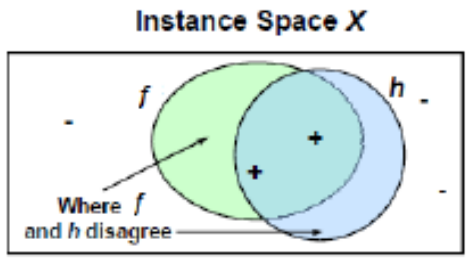
Duas noções de erro

- Erro de treinamento de uma hipótese h com respeito para um conceito f
  - Como frequentemente $h(x) \neq f(x)$ sobre as instâncias de treinamento
- Erro verdadeiro de uma hipótese h com relação á f
  - Como frequentemente $h(x) \neq f(x)$ sobre instâncias randômicas obtidas da distribuição D

Definição

O erro verdadeiro (denotado por $erro_D(h)$) de hipótese h com respeito para um conceito alvo f e distribuição D é a probabilidade que h irá classificar incorretamente uma instância obtida randomicamente de acordo com uma distribuição D.

$$erro_D(h) \equiv Prob_{x \in D}[f(x) \neq h(x)]$$



Erro verdadeiro de uma hipótese

Definição - Livro Mitchell

- Erro de treinamento de uma hipótese h com respeito para um conceito f
  - Como frequentemente $h(x) \neq f(x)$ sobre o instâncias de treinamento D
  - D - conjunto de exemplos de treinamento

$$erro_D(h) \equiv Prob_{x \in D}[f(x) \neq h(x)]$$

- Erro verdadeiro de uma hipótese h com relação à f
  - Como frequentemente $h(x) \neq f(x)$ sobre instâncias futuras obtidas randomicamente de D
  - D - Distribuição de probabilidade

$$erro_D(h) \equiv Prob_{x \in D}[f(x) \neq h(x)]$$

Teória de Aprendizado Computacional

Um determinado conjunto de treinamento de padrões pode ser suficiente para nos permitir selecionar uma função, consistente com os exemplos rotulados, dentre um conjunto restrito de hipóteses com alta probabilidade

A função selecionada será aproximadamente correta (probabilidade pequena de erro) sobre amostras subsequentes obtidas de acordo com a mesma distribuição a partir da qual as amostras rotuladas foram retiradas.

Essa percepção levou á Teoria de Aprendizado Provavelmente Aproximadamente Correta (PAC)

Notação e suposição para Teoria de Aprendizado PAC

Considere um conjunto de treinamento Ξ de vetores n-dimensionais, x_i, $i = 1; \cdots ; m$, rotulado (por 1 ou 0) de acordo com uma função alvo, f, é desconhecida pelo o preditor.

A probabilidade de qualquer vetor $x_i \in \Xi$, é dada por $P(x)$. A distribuição e probabilidade, P, pode ser arbitrária.

Na literatura de aprendizado PAC, a função alvo é usualmente chamada de conceito alvo é denotada por c, mas nós continuaremos a denotar por f .

Nosso problema é descobrir uma função, $h(x)$, baseado nos exemplos rotulados em Ξ. De acordo com teoria PAC tal função é chamada de hipótese.

Nós assumimos que a função alvo é algum elemento de um conjunto de funções, $\mathcal{C}$.

Nós assumimos que a hipótese h, é um elemento do conjunto de hipótese, $\mathcal{H}$, o qual inclui o conjunto , $\mathcal{C}$, de funções alvos. $\mathcal{H}$ é chamada de espaço de hipótese.

Em geral, h não seria idêntica a f , mas nós podemos nos esforçar para ter o valor de $h(x)$ igual ao valor de $f(x)$ para muitos $x's$. Isto é, queremos um h que seja aproximadamente correta. Para quantificar esta noção, nós definimos o erro de h, ϵ_h, como a probabilidade que um x obtido randomicamente de acordo com P seja classificado incorretamente.

$$erro_h = \sum_{x:h(x)\neq f(x)} P(x)$$

Nós dissemos que h é aproximadamente (exceto por ϵ) correta se $erro_h \leq \epsilon$, onde ϵ é o parâmetro de precisão.

Aprendizado PAC

Supondo que sejamos capazes de encontrar h que classifica todos os m exemplos de treinamento corretamente obtidos randomicamente, isto é, h é consistente com o conjunto de treinamento selecionado Ξ.

Se m for bastante grande, será h aproximadamente correto (e para qual valor de ϵ)?

Sobre alguns conjuntos de treinamento, usando m exemplos obtidos randomicamente, h será aproximadamente correto (para um dado valor de ϵ) e para outros pode não ser

Nós dissemos que h é provavelmente aproximadamente correto (PAC)(exceto para ϵ) se a probabilidade que seja aproximadamente correto é maior que $(1 - \delta)$, onde δ é o parâmetro de confiança.

Conclusão

Isto mostra que se m é maior que algum limiar cujo valor depende de ϵ e β, h é garantido ser provavelmente aproximadamente correto.

Aprendizado PAC

Em geral, podemos dizer que um algoritmo de aprendizado PAC aproxima uma função de $\mathcal{C}$ em termos de $\mathcal{H}$ se e somente se para toda função $f \in \mathcal{C}$, esta produz uma hipótese $h \in \mathcal{H}$, com probabilidade pelo menos $(1 - \delta)$, $erro_h \leq \epsilon$. Tal hipótese é chamada provavelmente (exceto para δ) aproximadamente (exceto para ϵ) correta

- Queremos algoritmos de aprendizado que sejam tratáveis, assim queremos um algoritmo PAC que aproxima funçoes em tempo polinomial. Isto pode ser feito para algumas classes de funções.
- Se há um número finito de hipótese no espaço de hipótese (como há para muitos dos conjuntos de hipótese que nós iremos considerar), nós poderemos sempre produzir uma hipótese consistente deste conjunto por testando todas contra os dados de treinamento.
- Mas se há um numero exponencial de hipótese isto poderá tomar um tempo exponencial.
- Nós procuramos métodos de treinamento que produz hipótese consistente em menos tempo.
- A complexidade temporal para várias hipóteses tem sido determinada

Aprendizado PAC

Uma classe, $\mathcal{C}$, é PAC aproximada polinomialmente em termos de $\mathcal{H}$ desde que exista um algoritmo de aprendizado PAC em tempo polinomial (polinomial no número de exemplos necessários, m, na dimensão, n, em $1/\epsilon$ e em $1/\delta$) aproxima funções em $\mathcal{C}$ em termos de $\mathcal{H}$.

- Trabalhos iniciais sobre PAC assumiram que $\mathcal{H} = \mathcal{C}$
- Mais tarde, foi mostrado que algumas funções não podem ser PAC aproximada polinomialmente sobre tal suposição.
- A nossa definição não especifica a distribuição, P, da qual os padrões são obtidos nem é dito nada a respeito das propriedades do algoritmos de aprendizado.
- PAC aproximada apropriadamente é uma classe $\mathcal{C}$ para qual existe um algoritmo PAC aproximada polinomialmente que aproxima funções de $\mathcal{C}$ em termos de $\mathcal{H}$.

Aprendizado PAC

Teorema Fundamental

Suponha que nosso algoritmo de aprendizado seleciona algum h randomicamente dentre aqueles que são consistente com os valores de f sobre m padrões de treinamento.
A probabilidade que o erro de h selecionado randomicamente é maior que algum ϵ, com h consistente com os valores de $f(x)$ para m instâncias (obtidas de acordo com P arbitrário), é menor que ou igual a $|\mathcal{H}|e^{-\epsilon m}$

Teorema - Blumer

Seja $\mathcal{H}$ um conjunto qualquer de hipótese, Ξ um conjunto de $m \geq 1$ exemplos de treinamento obtidos independentemente de acordo com alguma distribuição P, f seja qualquer função de classificação em $\mathcal{H}$, e $\epsilon > 0$. Então, a probabilidade que existe uma hipótese h consistente com f para os elementos de Ξ mas com erro maior que ϵ é no máximo $|\mathcal{H}|e^{-\epsilon m}$.

Aprendizado PAC

Prova

- Considere o conjunto de todas hipóteses, $\{h_1, h_2, \cdots, h_i, \cdots, h_S\}$, em $\mathcal{H}$, onde $S = |\mathcal{H}|$.
- O erro para h_i é $erro_{h_i}$ = a probabilidade que h_i irá classificar um padrão incorretamente (isto é, diferente do valor atribuído por f).
- A probabilidade que h_i ira classificar um padrão corretamente é $(1 - erro_{h_i})$.
- Um subconjunto, $\mathcal{H}_B$, de $\mathcal{H}$ irá ter um erro maior que ϵ. Nós chamamos a hipótese deste conjunto de *ruim*.
- A probabilidade que qualquer um destas hipóteses ruins, isto é h_b, possa classificar um padrão corretamente é $(1 - erro_{h_b})$
- Desde que $erro_{h_b} > \epsilon$, a probabilidade que h_b (ou qualquer outra hipótese ruim) possa classificar um padrão corretamente é menor que $(1 - \epsilon)$

Aprendizado PAC

Prova

A probabilidade que h_b possa classificar corretamente todos padrões obtidos independentemente é então menor que $(1-\epsilon)^m$

Isto é, prob[h_b classifica todos os m padrões corretamente | $h_b \in \mathcal{H}_B$] $\leq (1-\epsilon)^m$

prob[algum $h \in \mathcal{H}_B$ classifica todos os m padrões corretamente]
$= \sum_{h_b \in \mathcal{H}_B}$ prob [h_b classifica todos os m padrões corretamente $|h_b \in \mathcal{H}$]
$\leq K(1-\epsilon)^m$, onde $K = |\mathcal{H}_B|$.

Aprendizado PAC

Prova

Desde que $K \leq |\mathcal{H}|$ e $(1-\epsilon)^m \geq e^{-\epsilon m}$, nós temos
prob[há uma hipótese ruim que classifica todos os m padrões corretamente] = prob [há uma hipótese com erro $> \epsilon$ e que classifica todos os m padrões corretamente] $< |\mathcal{H}|e^{-\epsilon m}$.

Corolário

Dado $m \geq (1/\epsilon)(ln|\mathcal{H}| + ln(1/\delta)$ amostras independente, a probabilidade que há uma hipótese em $\mathcal{H}$ que é consistente com f sobre estas amostras e tem erro maior que ϵ é no máximo δ.

Aprendizado PAC

Prova

- Nós procuramos um limite sobre m que garante que prob[há uma hipótese com erro $> \epsilon$ e que classifica todos os m padrões corretamente] $\leq \delta$
- Então, usando o resultado do teorema, nós podemos mostrar que $|\mathcal{H}|e^{-\epsilon m} \leq \delta$.
- Aplicando o logaritmo natural em ambos os lados

$$ln|\mathcal{H}| - \epsilon m \leq ln\delta$$

ou

$$m \geq (1/\epsilon)(ln|\mathcal{H}| + ln(1/\delta))$$

Aprendizado PAC

Importância do Corolário

- Este resultado nos diz que podemos selecionar qualquer hipótese consistente com m exemplos e ser assegurado com a probabilidade $(1 - \delta)$ que seu erro será menor que ϵ.
- Também mostra que m incrementa polinomialmente com n, $|\mathcal{H}|$ não pode ser maior que $2^{O(n^k)}$. Nenhuma classe maior que esta pode ser garantida ser PAC aproximada apropriadamente.

Importante

O limite dado pelo corolário é um limite superior sobre o valor de m necessário para garantir aprendizado provavelmente aproximadamente correto. Valores de m maiores que o limiar são suficiente (mas pode não ser necessário). Nós iremos apresentar um limite inferior mais tarde.

Exemplo

Termos

Seja $\mathcal{H}$ o conjunto de termos (conjunções de literais). Então, $|\mathcal{H}| = 3^n$, e

$$m \geq (1/\epsilon)(ln(3^n) + ln(1/\delta))$$

$$m \geq (1/\epsilon)(1.1n + ln(1/\delta))$$

Exemplo

- Observe que o limitante sobre m incrementa polinomialmente com n, $1/\epsilon$, e $1/\delta$.
- Para $n = 50$, $\epsilon = 0.01$ e $\delta = 0.01$, $m \geq 5.961$ garante a capacidade de aprendizado PAC.
- A fim de mostrar que os termos são PAC aproximados apropriadamente, nós temos que mostrar que podemos encontrar em tempo polinomial em m e n uma hipótese consistente com um conjunto de m padrões rotulados pelo valor de um termo.
- Em Dietterich, 1990 é mencionado um procedimento para encontrar tal hipótese consistente que requer $O(nm)$ passos.

Limitações do limite de Haussler

- Pode não haver nenhum hipótese h consistente (onde $erro_{train}(h) = 0$)
- Tamanho do espaço de hipótese
  - E se $|\mathcal{H}|$ é grande?
  - E se for continuo?

Primeiro objetivo: podemos ter um limite para um preditor ter $erro_{train}(h)$ no conjunto de dados?

Questões: Qual é o erro esperado de uma hipótese?

- A probabilidade de uma hipótese classificar incorretamente: $\sum_{x,y} p(x,y)$
- Seja Z_i^h ser uma variável randômica que assume dois valores, 1 se h classifica corretamente amostra i, e 0 caso contrário.
- As variáveis Z são independente e identicamente distribuída (i.i.d) com $Pr(Z_i^h = 0) = \sum_{(x,y)} p(x,y)$
- Estimar a probabilidade de erro verdadeiro é como estimar o parâmetro de uma moeda
- Chernof bound: para m cara ou coroa i.i.d, $x_1, \cdots, x_m$, onde $x_i \in 0, 1$. Para $0 < \epsilon < 1$

$$P\left(\theta - \frac{1}{m}\sum_i x_i > \epsilon\right) \leq e^{-2m\epsilon^2}$$

Probabilidade de erro verdadeiro

Fração de pontos observados de pontos incorretamente classificados

$$p(X_i = 1) = \theta$$

Limite de Generalização para hipótese limitada $|\mathcal{H}|$

Teorema:

Espaço de hipótese $\mathcal{H}$ finito, conjunto de dados D com m amostras i.i.d, $0 < \epsilon < 1$, para qualquer hipótese h:

$$Pr(erro_{true}(h) - erro_D(h) > \epsilon) \leq |H|e^{-2m\epsilon^2}$$

$$Pr(erro_{true}(h) - erro_D(h) > \epsilon) \leq |H|e^{-2m\epsilon^2} \leq \delta$$

Assumindo que este resultado seja verdadeiro para uma probabilidade de no máximo δ

$$\ln(|\mathcal{H}|e^{-2m\epsilon^2}) \leq \ln \delta$$

$$\epsilon \geq \sqrt{\frac{\ln |\mathcal{H}| + \ln(1/\delta)}{2m}}$$

Limite PAC e Dilema Bias-Variância

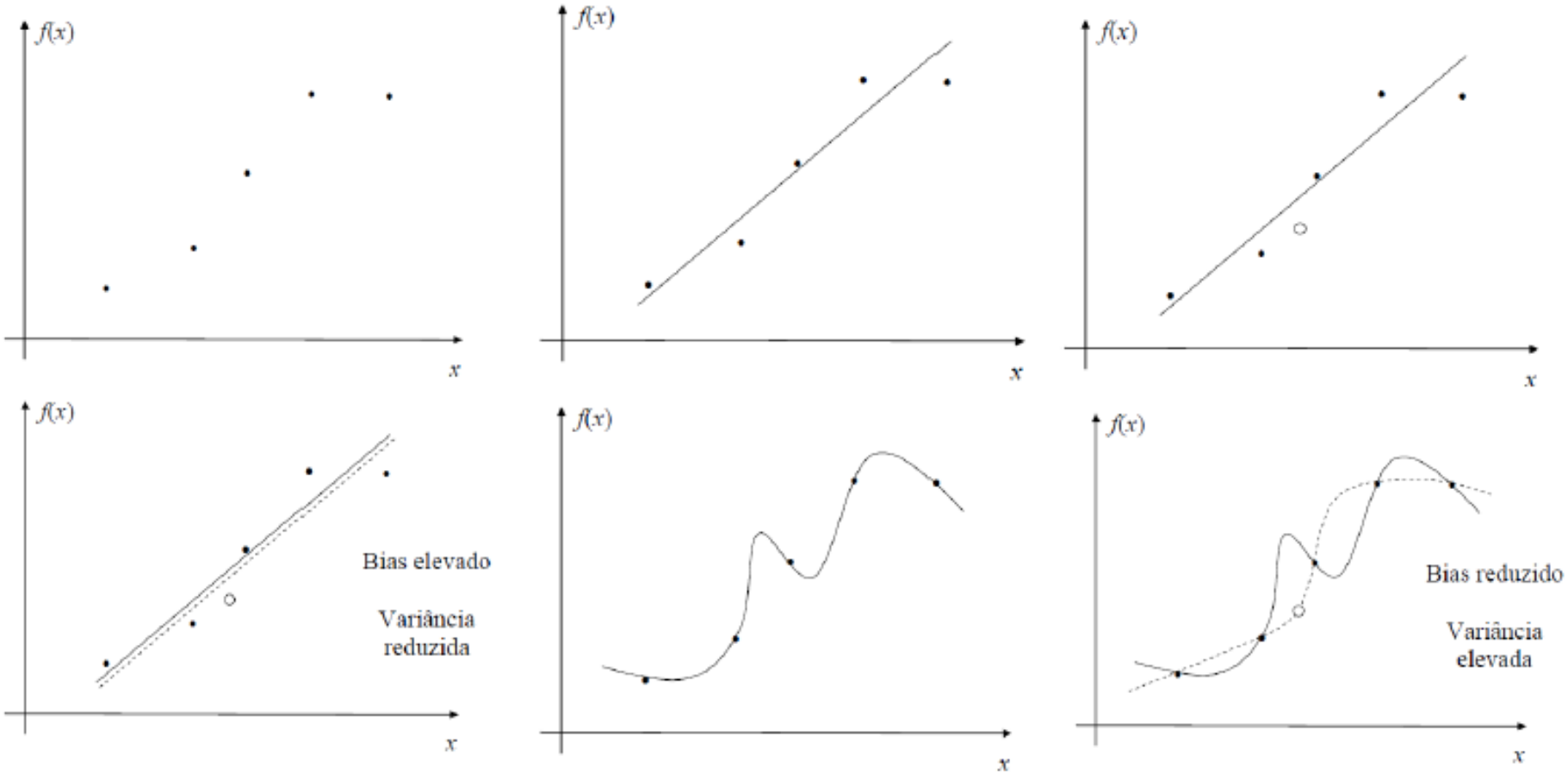
Para todo h, com probabilidade de pelo menos $1 - \delta$:

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

Conclusão

- Para $|\mathcal{H}|$ grande
  - Baixo bias (assumindo que encontramos um bom h)
  - Alta variância (por que é mais flexível)
- Para $|\mathcal{H}|$ pequeno
  - Alto bias (há um bom h)
  - Baixa variância (por que esta mais justo)

Dilema Bias-Variância



Risco Empíríco vs Risco Esperado

- Aprender uma função de classificação binária a partir dos dados
- Considere um conjunto de dados $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$ onde cada $y_i \in \{-1, 1\}$.
- Aprender uma função $y = f(x; \theta)$ que irá classificar corretamente os exemplos não observados
- Como é que vamos escolher o tipo de f e θ?
- Otimizando alguma medida de performance do modelo aprendido.
- O que é uma medida boa de performance?
- Uma medida boa de performance é o risco esperado

$$[R_f(\theta) = E[L(y, f(x; \theta)] = \int L(y, f(x; \theta) dP(x, y)$$

- Valor esperado da função de perda

Risco Empíríco vs Risco Esperado

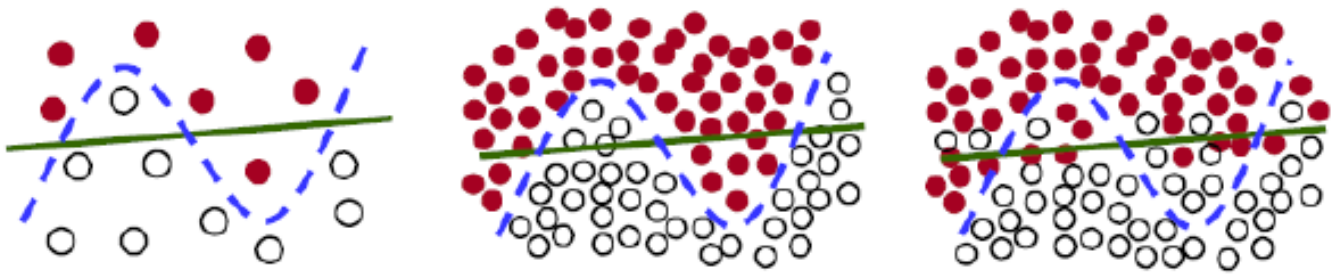
Minimização do Risco Empírico

- Um termo formal para um conceito simples: encontrar a função $f(x)$ que minimiza o risco médio sobre o conjunto de treinamento.
- Minimizar o risco empírico não é algo ruim a ser realizado, desde que dados de treinamento suficientes estejam disponiveis, a lei dos grandes número garante que o risco empírico irá convergir assintoticamente para o risco esperado ($n \to \infty$).
- Entretanto, para amostras pequenas, nós não podemos garantir que o ERM irá também minimizar o risco esperado. Esta é uma questão muito familiar de generalização.

Risco Empíríco vs Risco Esperado

Como podemos evitar o sobre-ajuste?

Controlando a complexidade do modelo. Intuitivamente, devemos preferir o modelo mais simples que explica os dados (Occam's razor).



Dimensão VC

A dimensão Vapnik-Chervonenkis

Esta é uma medida da complexidade / capacidade de uma classe de funções $\mathcal{F}$. Ela mede o maior número de exemplos que podem ser explicados pela família $\mathcal{F}$.
Compromisso entre Alta capacidade e Boa Generalização

Maior Capacidade

Se a família $\mathcal{F}$ tem capacidade suficiente para explicar todos os possíveis conjuntos de dados $\rightarrow$ há risco de sobre-ajuste

Menor Capacidade

Funções $f \in \mathcal{F}$ tendo pequena capacidade podem não ser capazes de explicar nosso conjunto de dados particular, entretanto, são menos propensa a sobre-ajuste.

Como a dimensão VC caracteriza este compromisso?

Dimensão Vapnik-Chervonenkis

$R_f(\theta) = E[\frac{1}{2}|y - f(x;\theta)|]$, $R_f^{emp}(\theta) = \frac{1}{n}\sum_{i=1}^{m}\frac{1}{2}|y_i - f(x_i,\theta)|$

- Dada uma classe de funções $\mathcal{F}$, seja VcDim ser sua dimensão VC
- VcDim é uma medida da capacidade de $\mathcal{F}$ (VcDim não depende da escolha do conjunto de treinamento)
- Vapnik mostrou que com probabilidade $1 - \delta$

$$R_f(\theta) \leq R_f^{emp}(\theta) + \sqrt{\frac{VcDim(\log(\frac{2m}{VcDim}) + 1) - \log(\frac{\delta}{4})}{m}}$$

Isto nos dá uma maneira de estimar o erro sobre dados futuros com base apenas no erro de treinamento e na dimensão VC de $\mathcal{F}$.
Dado $\mathcal{F}$ como nós podemos definir e calcular $VcDim$, sua dimensão VC?

Classificando

Uma função $f(x;\theta)$ pode classificar um conjunto de pontos $x_1, x_2, \cdots, x_m$ se e somente se

Para todo conjunto de treinamento possível da forma $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$ há algum valor de θ tal que $f(x_i, \theta) = y_i$ para $i = 1, \cdots, m$.

Há 2^m conjuntos de treinamento a considerar, cada um com uma combinação diferente de $+1's$ e $-1's$ para $y's$.

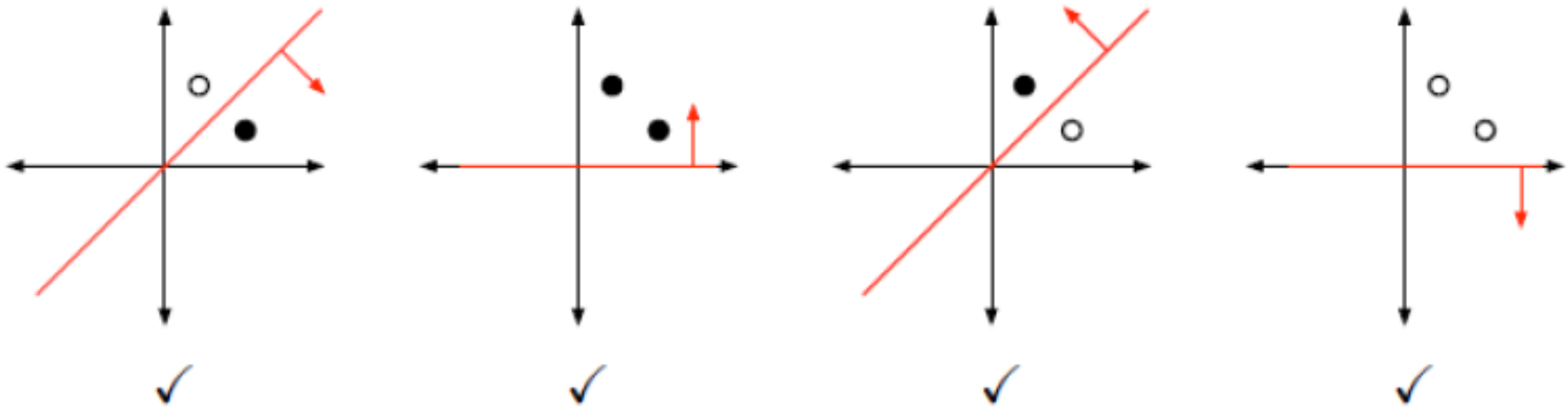
Classificando

Uma função $f(x;\theta)$ pode classificar um conjunto de pontos $x_1, x_2, \cdots, x_m$ se e somente se

Para todo conjunto de treinamento possível da forma $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$ há algum valor de θ tal que $f(x_i, \theta) = y_i$ para $i = 1, \cdots, m$.

Resposta

Nenhum problema. Há quatro conjuntos de dados a considerar.



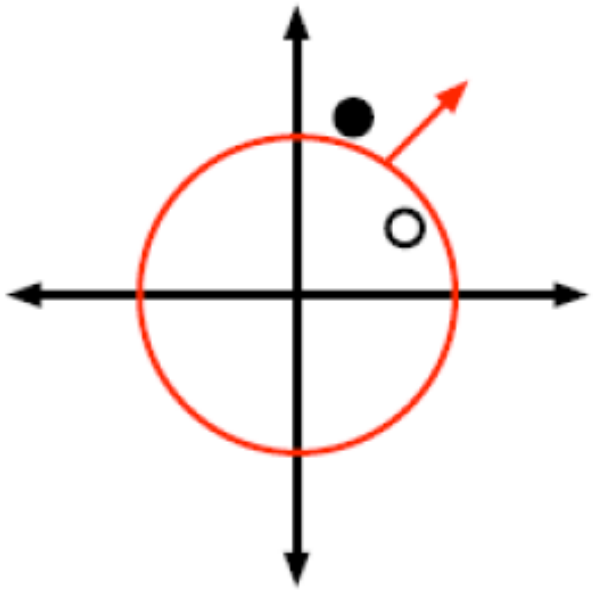
Classificando

Uma função $f(x;\theta)$ pode classificar um conjunto de pontos $x_1, x_2, \cdots, x_m$ se e somente se

Para todo conjunto de treinamento possível da forma $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$ há algum valor de θ tal que $f(x_i, \theta) = y_i$ para $i = 1, \cdots, m$.

Pode a seguinte função classificar os seguintes pontos?

$$f(x; b) = sign(x^T x - b)$$



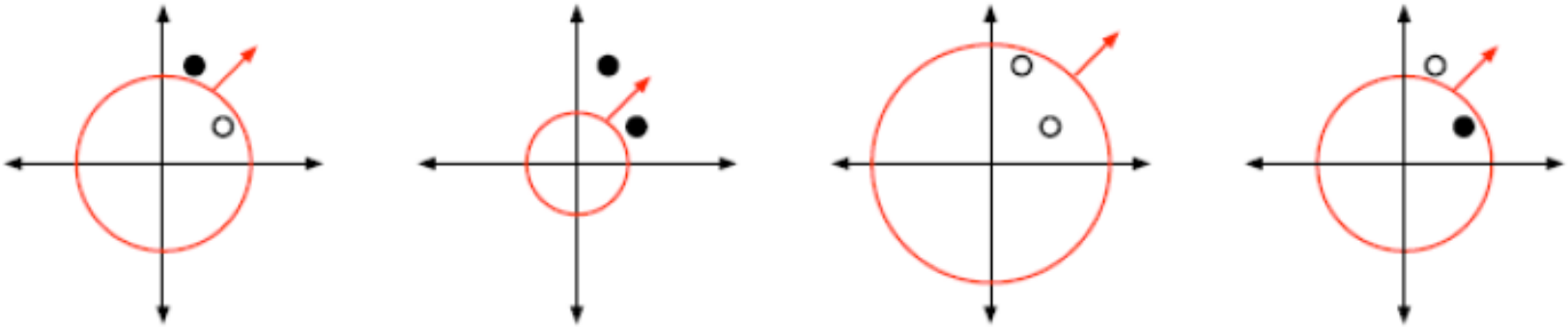
Classificando

Uma função $f(x;\theta)$ pode classificar um conjunto de pontos $x_1, x_2, \cdots, x_m$ se e somente se

Para todo conjunto de treinamento possível da forma $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$ há algum valor de θ tal que $f(x_i, \theta) = y_i$ para $i = 1, \cdots, m$.

Resposta

Não é possível.



Definição da Dimensão VC

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC $VCdim$ se há pelo menos um conjunto de $VCdim$ pontos que podem ser classificados por $f \in \mathcal{F}$ (em geral, não será verdade que todo conjunto de pontos pode ser classificado).

Questão

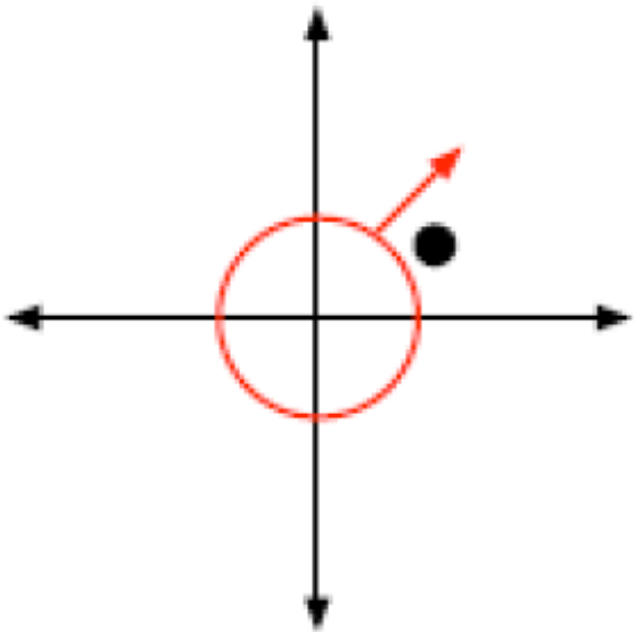
Qual é a dimensão VC de $f(x, b) = sign(x^T x - b)$?

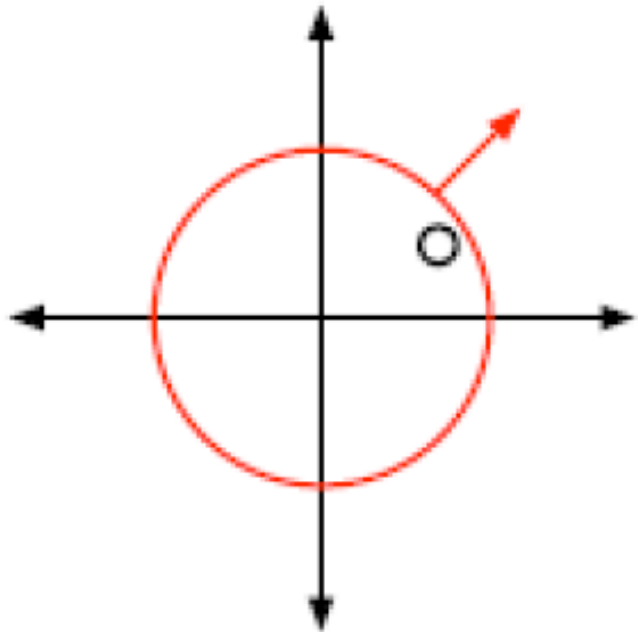
Definição da Dimensão VC

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC *VCdim* se há pelo menos um conjunto de *VCdim* pontos que podem ser classificados por $f \in \mathcal{F}$

Resposta

Nós não podemos mesmo classificar dois pontos. É claro que um ponto pode ser classificado





Definição da Dimensão VC

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC $VCdim$ se há pelo menos um conjunto de $VCdim$ pontos que podem ser classificados por $f \in \mathcal{F}$

Exemplo

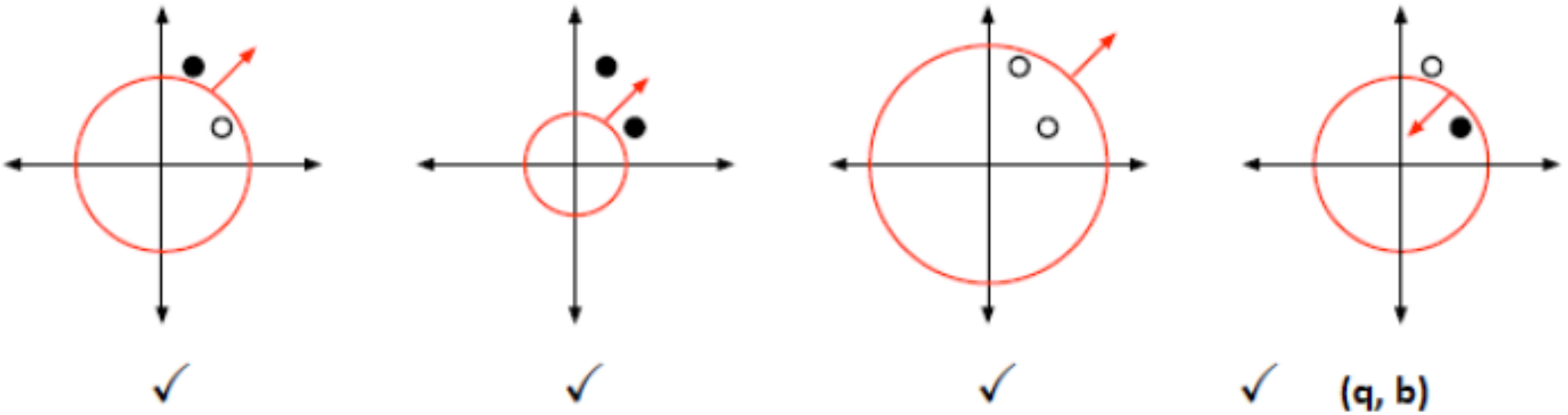
Para entradas bi-dimensional, qual é a dimensão de

$$f(x; q, b) = sign(qx^T x - b)$$

Definição da Dimensão VC

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC *VCdim* se há pelo menos um conjunto de *VCdim* pontos que podem ser classificados por $f \in \mathcal{F}$

Resposta: 2

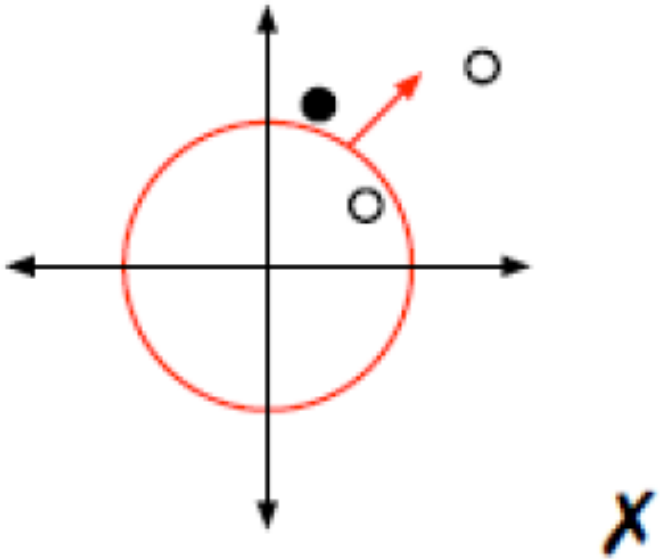


Definição da Dimensão VC

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC *VCdim* se há pelo menos um conjunto de *VCdim* pontos que podem ser classificados por $f \in \mathcal{F}$

Exemplo

Qual é a dimensão VC de $f(x; q, b) = sign(qx^T x - b)$
Resposta: 2 (claramente não pode ser 3)



Definição da Dimensão VC de uma reta

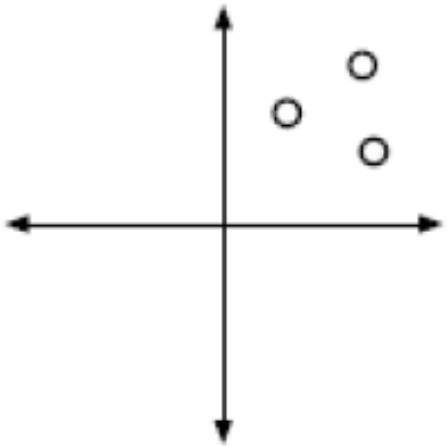
Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC $VCdim$ se há pelo menos um conjunto de $VCdim$ pontos que podem ser classificados por $f \in \mathcal{F}$

Resposta

Para entradas bidimensionais, qual é a dimensão VC de

$$f(x; w, b) = sign(w^T x + b)$$

Pode f classificar estes 3 pontos?

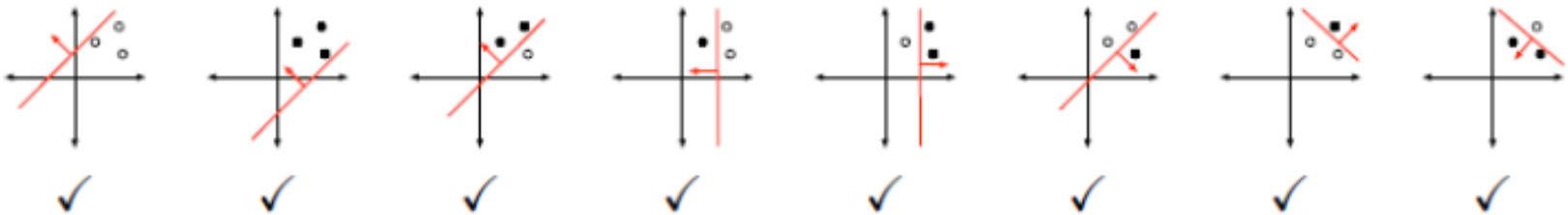


Definição da Dimensão VC de uma reta

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC *VCdim* se há pelo menos um conjunto de *VCdim* pontos que podem ser classificados por $f \in \mathcal{F}$

Exemplo

Qual é a dimensão VC de $f(x; w, b) = sign(w^T x + b)$
Resposta: Sim, pode classificar 3 pontos.



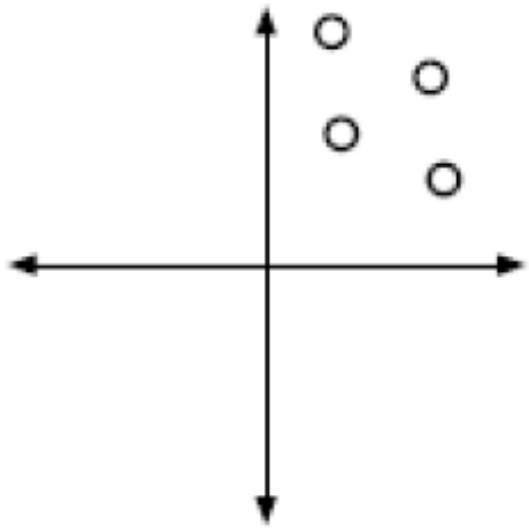
Definição da Dimensão VC de uma reta

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC *VCdim* se há pelo menos um conjunto de *VCdim* pontos que podem ser classificados por $f \in \mathcal{F}$

Resposta

Para entradas bidimensionais, qual é a dimensão VC de $f(x; w, b) = sign(w^T x + b)$?

Pode f classificar estes 4 pontos?



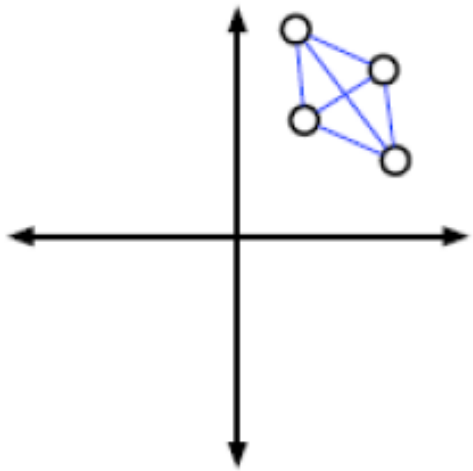
Definição da Dimensão VC de uma reta

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC *VCdim* se há pelo menos um conjunto de *VCdim* pontos que podem ser classificados por $f \in \mathcal{F}$

Resposta

Para entradas bidimensionais, qual é a dimensão VC de $f(x; w, b) = sign(w^T x + b)$?

Podemos sempre desenhar 6 linhas entre pares de 4 pontos



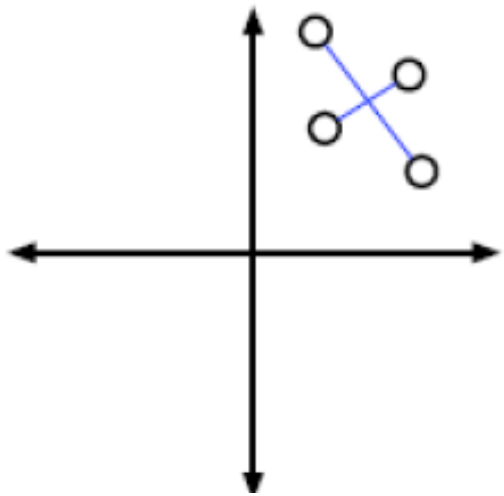
Definição da Dimensão VC de uma reta

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC *VCdim* se há pelo menos um conjunto de *VCdim* pontos que podem ser classificados por $f \in \mathcal{F}$

Resposta

Para entradas bidimensionais, qual é a dimensão VC de $f(x; w, b) = sign(w^T x + b)$?

Podemos sempre desenhar 6 linhas entre pares de 4 pontos
Duas destas linhas se cruzam



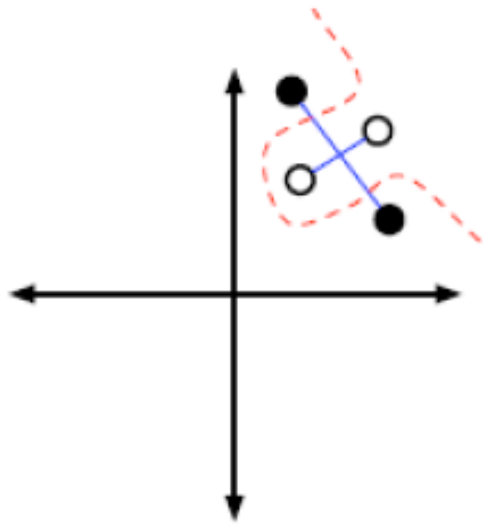
Definição da Dimensão VC de uma reta

Dada a classe de funções $\mathcal{F}$, ele tem dimensão VC *VCdim* se há pelo menos um conjunto de *VCdim* pontos que podem ser classificados por $f \in \mathcal{F}$

Resposta

Qual é a dimensão VC de $f(x; w, b) = sign(w^T x + b)$?

Podemos sempre desenhar 6 linhas entre pares de 4 pontos
Duas destas linhas se cruzam
Se analisarmos os pontos ligados pela reta que se cruzam, veremos que eles não poderão ser separados linearmente.



Uma linha pode classificar 3 pontos mas não 4 $\rightarrow$ a dimensão VC de uma linha de separação é 3.

O que mede a dimensão VC

È o número de parâmetros?

Relacionado, mas não é a mesma coisa

Pode-se esperar que intuitivamente modelos com um número grande parâmetros livres teriam maior dimensão VC, enquanto os modelos com poucos parâmetros teria dimensões VC baixa

Exemplo

Considere uma função de um parâmetro α

$$f_{alpha}(x) = sign(sin(\alpha x)), x, \alpha \in \mathcal{R}$$

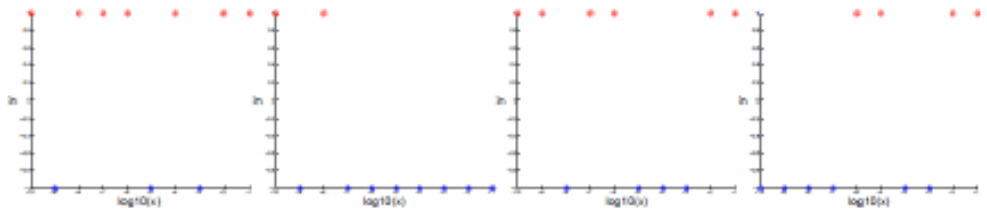
Escolha um número arbitrário VcDim e faça $x_i = 10^{-i}, i = 1, \cdots, VcDim$
Escolha os correspondentes rotulos y_i arbitrariamente como $y_i \in \{-1, +1\}$
Seja α

$$\alpha = \pi \left(1 + \sum_{i=1}^{VcDim} \frac{(1 - y_i)10^i}{2}\right)$$

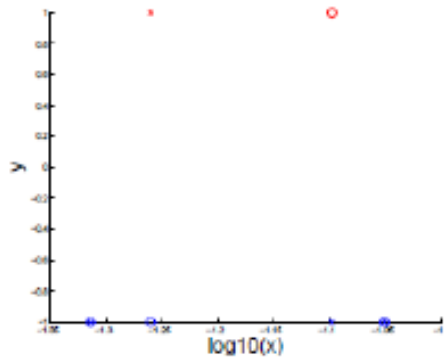
Apesar de ter apenas um parâmetro, a função f_α classifica um número arbitrário de pontos escolhidos de acordo com o procedimento acima.

Exemplo

Quaisquer que sejam a rotulação, as predições são corretas.
Círculos representam a predição e a cruz as rotulações



Mas, nós podemos também encontrar 4 pontos que não podem ser classificados pela função



Assim o que nós podemos tirar disso? A dimensão VC é uma medida mais sofisticada da complexidade do modelo que a dimensionalidade ou número de parâmetros livres.

Minimização Risco Estrututal

Outro termo formal para um conceito intuitivo: o modelo ótimo é encontrado por estabelecendo um equilíbrio entre risco empírico e a dimensão VC.
Relembre

$$R(\theta) \leq R^{emp}(\theta) + \sqrt{\frac{VcDim(\log(\frac{2m}{VcDim}) + 1) - \log(\delta/4)}{m}}$$

O princípio SRM procede como a seguir

- Construa uma estrutura aninhada para a família de classes de funções $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_k$ com dimensão VC não decrescente $(VcDim_1 \leq VcDim_2 \leq \cdots VcDim_k$
- Para cada classe de $\mathcal{F}_)$, compute a solução f_i que minimiza o risco empírico
- Escolha a classe de função $\mathcal{F}_)$, e correspondente solução f_i, que minimiza o limitante do risco.

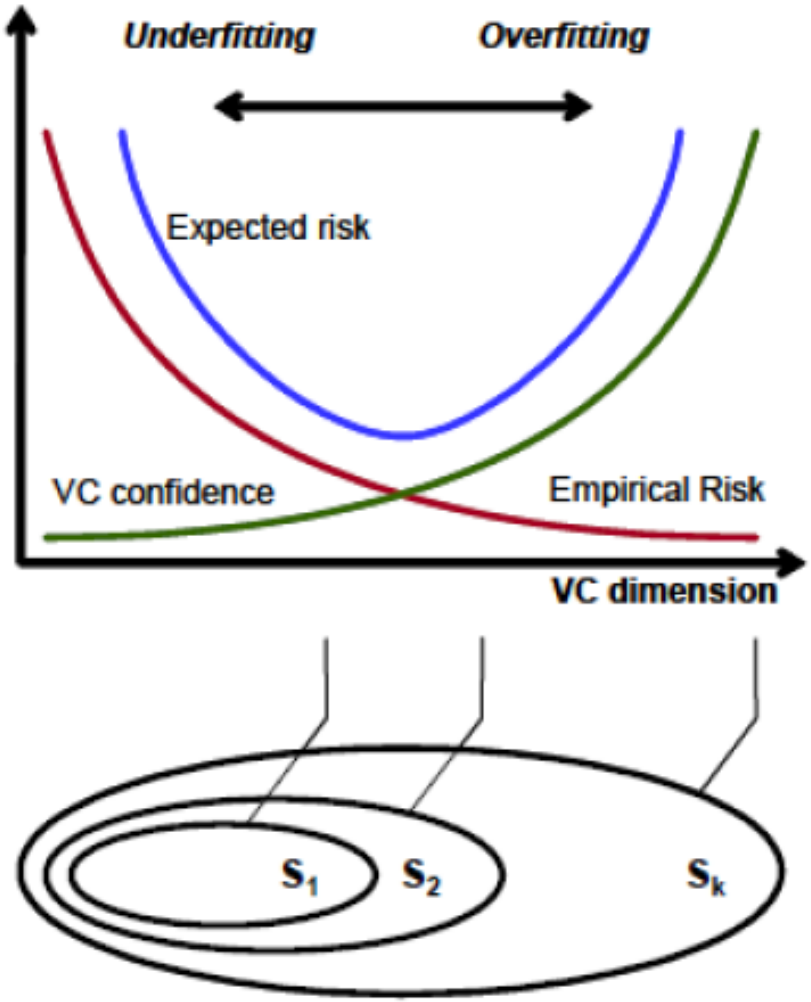
Em outras palavras

- Treine um conjunto de máquinas, um para cada subconjunto
- Para um dado subconjunto, minimize o risco empírico
- Escolha a maquina cuja soma do risco empírico e a confiança VC é mínimo

i	$\mathcal{F}_i$	$R^{\text{emp}}(\boldsymbol{\theta})$	VC Confidence	Probable Upper bound	Choice
1	$\mathcal{F}_1$				
2	$\mathcal{F}_2$				
3	$\mathcal{F}_3$				
4	$\mathcal{F}_4$				
5	$\mathcal{F}_5$				
6	$\mathcal{F}_6$				

Observe que o termo da dimensão VC é usualmente muito, muito conservador (pelo menos centena de vezes maior que o efeito de sobre-ajuste)

Minimização Risco Estrutural



Usando dimensão VC

Vários pesquisadores tem trabalhado arduamente para encontrar a dimensão VC para

- Perceptrons
- Redes Neurais
- Máquinas de Vetores Suporte
- e Muito mais

Tudo com o objetivo de

Entender quais máquinas de aprendizado são mais ou menos poderosa sobre algumas circunstâncias
Usar a Minimização do Risco Estrutural para escolher a melhor máquina de aprendizado

Alternativas para Dimensão VC baseado seleção de modelo

Poderíamos potencialmente usar k-fold validação cruzada:

i	$\mathcal{F}_i$	$R^{\text{emp}}(\boldsymbol{\theta})$	VC Confidence	Probable Upper bound	Choice
1	$\mathcal{F}_1$				
2	$\mathcal{F}_2$				
3	$\mathcal{F}_3$				
4	$\mathcal{F}_4$				
5	$\mathcal{F}_5$				
6	$\mathcal{F}_6$				

Note erro de CV pode ter mais variância

Dimensão VC na prática

- Infelizmente, calcular o limitante superior sobre o risco estrutural não é prático em várias situações
- A dimensão VC não pode ser precisamente estimada para modelos não lineares como rede neurais
- Implementação da Minimização do Risco Estrutural pode conduzir a problema de otimização não linear
- A dimensão VC pode ser infinita (por exemplo, Vizinho Mais Próximo com k=1), requerendo uma quantidade infinita de dados
- O limitante superior pode algumas vezes ser trivial (maior que 1)
- Felizmente, Teoria de Aprendizado Estatística pode ser rigorosamente aplicada a modelos lineares.

Modelos Paramétricos vc Modelos não Paramétricos

- análise multivariável clássica: fornece ferramentas poderosas na obtenção de associações lineares entre as variáveis.
- se todas as informações relevantes puderem ser extraídas com base nestas ferramentas clássicas, nenhum passo adicional se faz necessário.
- nos casos em que associações não lineares arbitrárias estão presentes, a determinação do tipo de não linearidade é fundamental para a obtenção do melhor modelo de aproximação a partir dos dados de entrada-saída
- quando a forma da não linearidade é conhecida previamente e passível de descrição matemática, modelos paramétricos são normalmente empregados, simplificando o problema de aproximação, já que os parâmetros podem ser determinados com base em técnicas de regressão não linear.

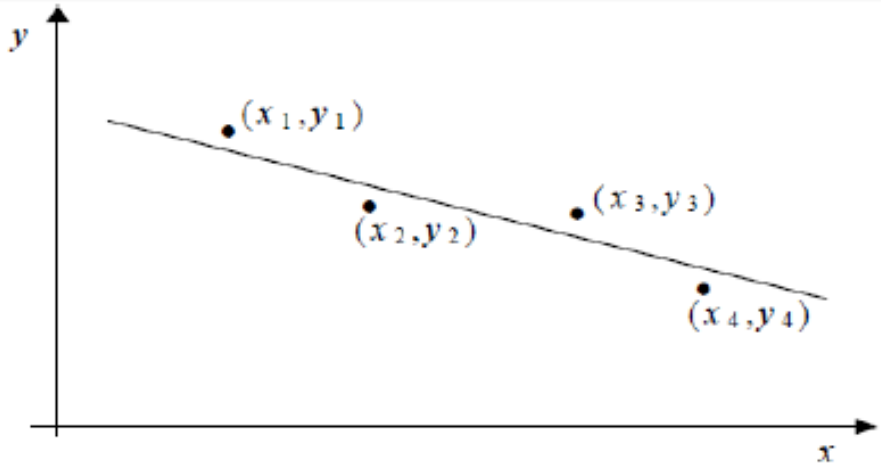
Modelos Paramétricos vc Modelos não Paramétricos

- regressão paramétrica: a forma do relacionamento funcional entre as variáveis dependentes e independentes é conhecida, mas podem existir parâmetros cujos valores são desconhecidos, embora passíveis de serem estimados a partir do conjunto de treinamento.
- em problemas paramétricos, os parâmetros livres, bem como as variáveis dependentes e independentes, geralmente têm uma interpretação física.
- Exemplo: ajuste de uma reta a uma distribuição de pontos

$f(x) = y = ax + b$

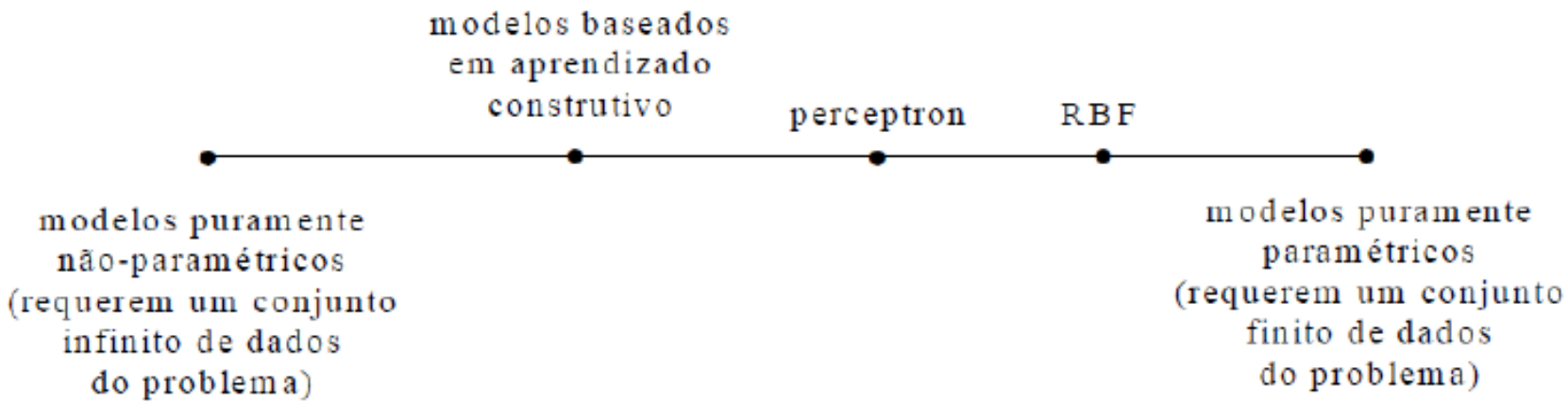
a, b desconhecidos

y: sujeito a ruído



Modelos Paramétricos vc Modelos não Paramétricos

- regressão não paramétrica: sua característica distintiva é a ausência (completa ou quase completa) de conhecimento a priori a respeito da forma da função que está sendo estimada. Sendo assim, mesmo que a função continue a ser estimada a partir do ajuste de parâmetros livres, o conjunto de "formas" que a função pode assumir (classe de funções que o modelo do estimador pode prever) é muito amplo.
- como consequência, vai existir um número elevado de parâmetros (por exemplo, quando comparado ao número de dados de entrada-saída para treinamento), os quais não mais admitem uma interpretação física isolada.



Modelos Paramétricos vc Modelos não Paramétricos

- todos os modelos de regressão que não são puramente paramétricos são denominados não paramétricos ou semi-paramétricos. Esta denominação não deve causar confusão, principalmente levando-se em conta que modelos de regressão puramente não paramétricos são intratáveis.
- com base no exposto acima, fica evidente que redes neurais artificiais para treinamento supervisionado pertencem à classe de modelos de regressão nãoparamétricos. Sendo assim, os pesos não apresentam um significado físico particular em relação ao problema de aplicação.
- além disso, estimar os parâmetros de um modelo não paramétrico (por exemplo, pesos de uma rede neural artificial) não é o objetivo primário do aprendizado supervisionado. O objetivo primário é estimar a "forma" da função em uma região compacta do espaço de aproximação (ou ao menos a saída para certos valores desejados de entrada).

Modelos Paramétricos vc Modelos não Paramétricos

- quando a forma da não linearidade é desconhecida, a utilização de modelos paramétricos pode representar uma perda acentuada de flexibilidade de representação, principalmente nos casos em que as restrições paramétricas impostas ao modelo de aproximação não correspondem à forma da não linearidade que deve ser aproximada.
- mesmo no caso de problemas de aproximação passíveis de tratamento paramétrico, é recomendada uma abordagem inicial utilizando modelos não paramétricos para auxiliar na determinação do tipo de parametrização que pode ser utilizada.
- STONE (1977) faz um estudo mais aprofundado desta e outras motivações para o emprego de modelos não paramétricos.
- dificuldade: o emprego de modelos não paramétricos provoca uma significativa acentuação de uma característica já presente em alguns problemas de aproximação que utilizam abordagens paramétricas: a maldição da dimensionalidade.

Modelos Paramétricos vc Modelos não Paramétricos

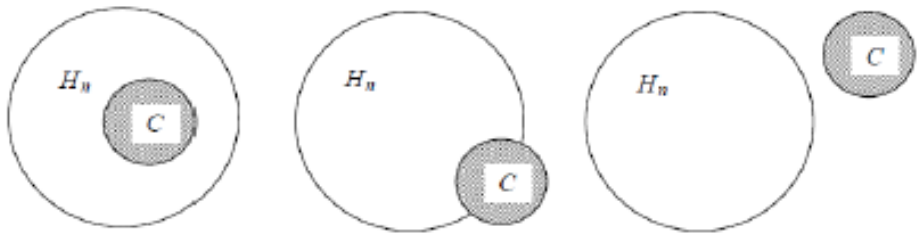
Maldição da dimensionalidade

- esta expressão foi empregada originalmente por BELLMAN (1961) e se refere basicamente à existência de uma relação direta entre a dimensionalidade dos dados e a quantidade de dados necessária para possibilitar o sucesso da tarefa de aproximação.
- a consequência prática da maldição da dimensionalidade é a necessidade de um aumento exponencial no número de dados para a manutenção do poder de aproximação com um aumento da dimensão do espaço de aproximação.

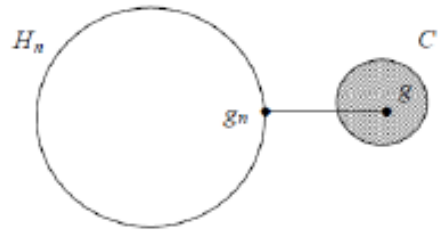
- apesar de estar invariavelmente associada a uma redução da capacidade de aproximação, a imposição de um conjunto de restrições (inclusive, restrições paramétricas) aos modelos não paramétricos pode reduzir significativamente o efeito da maldição da dimensionalidade.
- conclusão: os melhores modelos de aproximação são aqueles capazes de conciliar o nível de dependência da dimensionalidade com a flexibilidade do modelo de aproximação.

Representação

Modelagem paramétrica x modelagem não paramétrica



$$g_n = \arg \min_{\hat{g}_n \in H_n} \|g - \hat{g}_n\|, \text{ com } g \in C$$

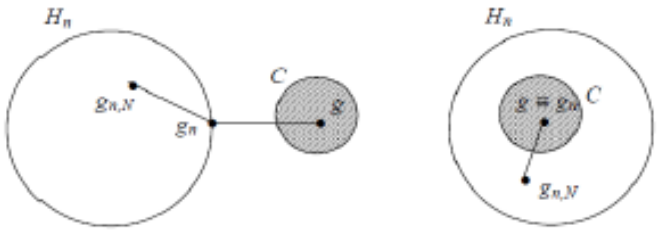


sobe que condições e que taxa $g_n \to g$ quando $n \to \infty$

Representação

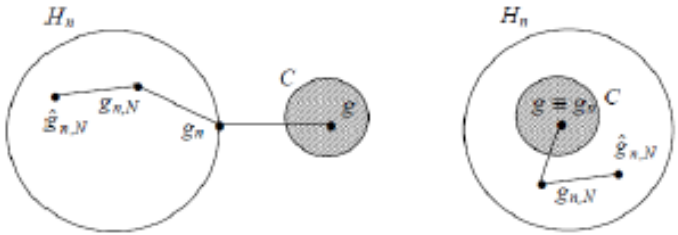
Estimação

$$g_{n,N} = \arg \min_{\hat{g}_n \in H_n, T_N} \left\| s_l - \hat{g}_{n,N}(x_l) \right|$$



sob que condições e a que taxa $g_{n,N} \rightarrow g_n$ quando $N \rightarrow \infty$

Computação



Composição do erro de estimação

$$\hat{e}_{n,N} = g - \hat{g}_{n,N}$$

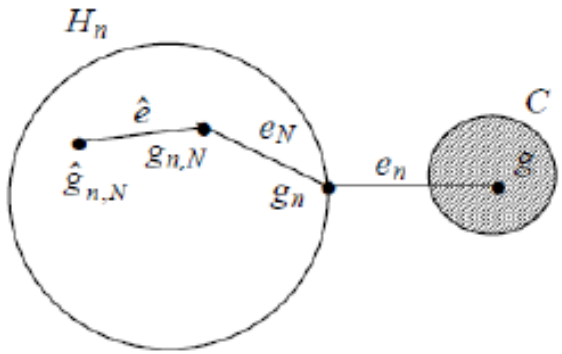
$$\hat{e}_{n,N} = \underbrace{g - g_n}_{e_n} + \underbrace{g_n - g_{n,N}}_{e_N} + \underbrace{g_{n,N} - \hat{g}_{n,N}}_{\hat{e}}$$

e_n: erro de aproximação

e_N: erro de estimação

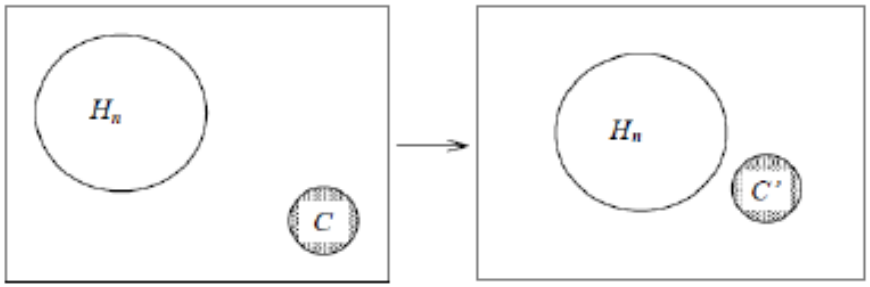
$\hat{e}$: erro de computação

Usando a desigualdade triangular: $\left\|\hat{e}_{n,N}\right\| \leq \left\|e_n\right\| + \left\|e_N\right\| + \left\|\hat{e}\right\|$



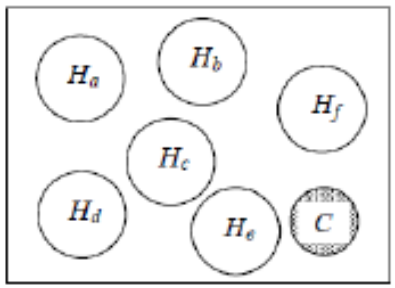
Técnicas para redulçao do erro de aproximação

1) Simplificar o problema de aproximação (modelo e/ou função)



2) Escolher uma melhor classe de modelos de aproximação

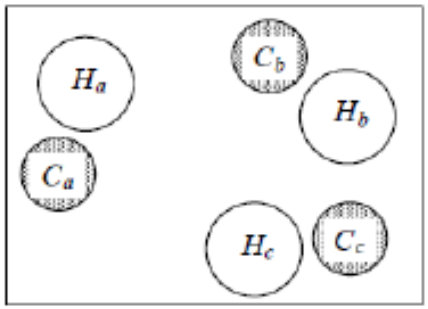
- *não existe uma classe de modelos que seja a mais adequada para todos os problemas*



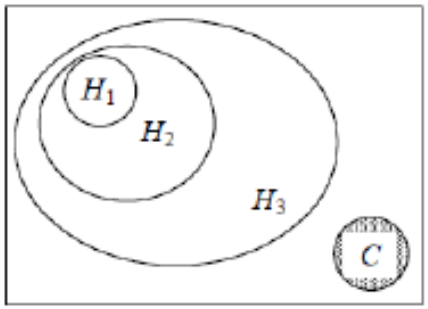
Técnicas para redulçao do erro de aproximação

3) Quebrar o problema em problemas menores (mais simples)

3(a) Mixture of experts



3(b) Métodos construtivos



4) Redução de dimensionalidade