Tema 4

Treinamento de HMMs e Estimação de Desempenho

Professora:
Ariane Machado Lima

Vídeo 1

O que estamos estudando mesmo?

Exemplo Viterbi e posterior decoding

Aulas anteriores

- Aula 1: Conceitos básicos de RP (sintático e não sintático)
- Aula 2: Gramáticas e Hierarquia de Chomsky, Gramáticas regulares, AFDs e AFNs
- Aula 3: Modelos Probabilísticos Regulares
  - gramáticas regulares estocásticas, autômatos probabilísticos, modelos ocultos de Markov (HMM)
  - Como calcular a probabilidade de uma cadeia dado um modelo (no HMM algoritmo forward ou backward)
  - Como calcular a probabilidade de uma cadeia em um dado caminho pela HMM (algoritmo Viterbi)
  - Como usar essas probabilidades para classificação (score log-odd, modelo nulo, classificação Bayesiana baseada na posteriori)

Aulas anteriores

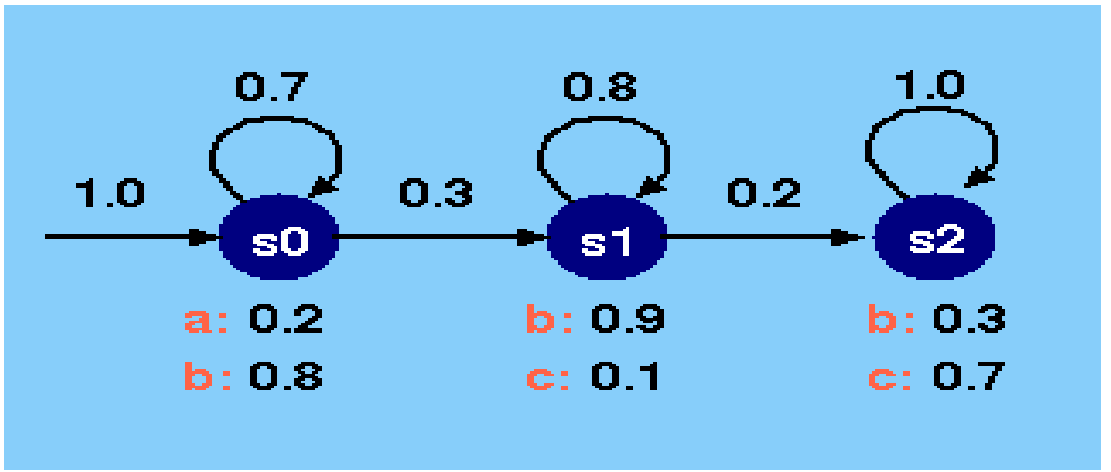
- Aula 1: Conceitos básicos de RP (sintático e não sintático)
- Aula 2: Gramáticas e Hierarquia de Chomsky, Gramáticas regulares, AFDs e AFNs
- Aula 3: Modelos Probabilísticos Regulares
  - gramáticas regulares estocásticas, autômatos probabilísticos, modelos ocultos de Markov (HMM)
  - Como calcular a probabilidade de uma cadeia dado um modelo (no HMM algoritmo forward ou backward)
  - Como calcular a probabilidade de uma cadeia em um dado caminho pela HMM (algoritmo Viterbi)
  - Como usar essas probabilidades para classificação (score log-odd, modelo nulo, classificação Bayesiana baseada na posteriori)

Aula de hoje

- Exemplo de utilidade do Viterbi
- Como treinar esses modelos (HMMs)
- Estimação de desempenho / calibração de parâmetros

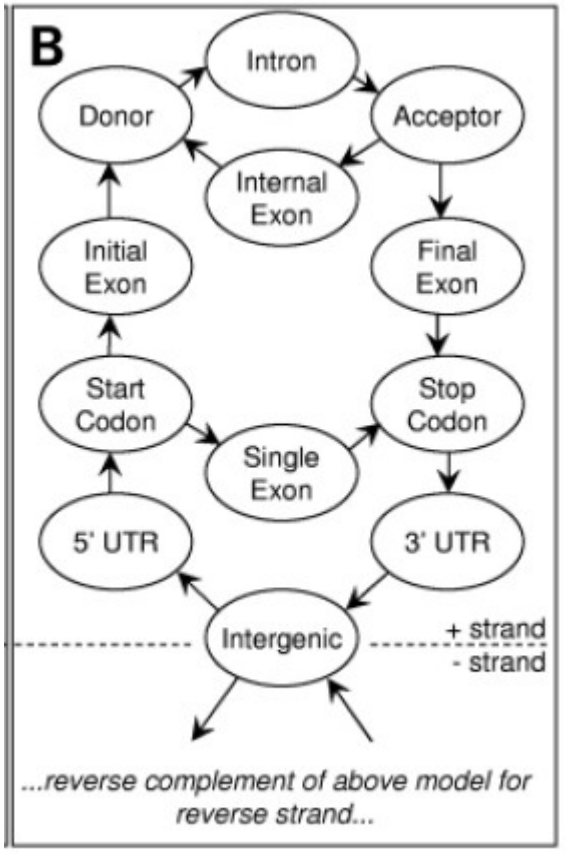
Aula de hoje

- Exemplo de utilidade do Viterbi
- Como treinar esses modelos (HMMs)
- Estimação de desempenho / calibração de parâmetros



Exemplo de uso do Viterbi

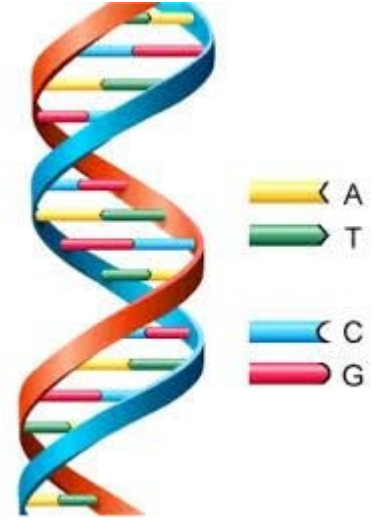
- Preditor de gene Exonomy



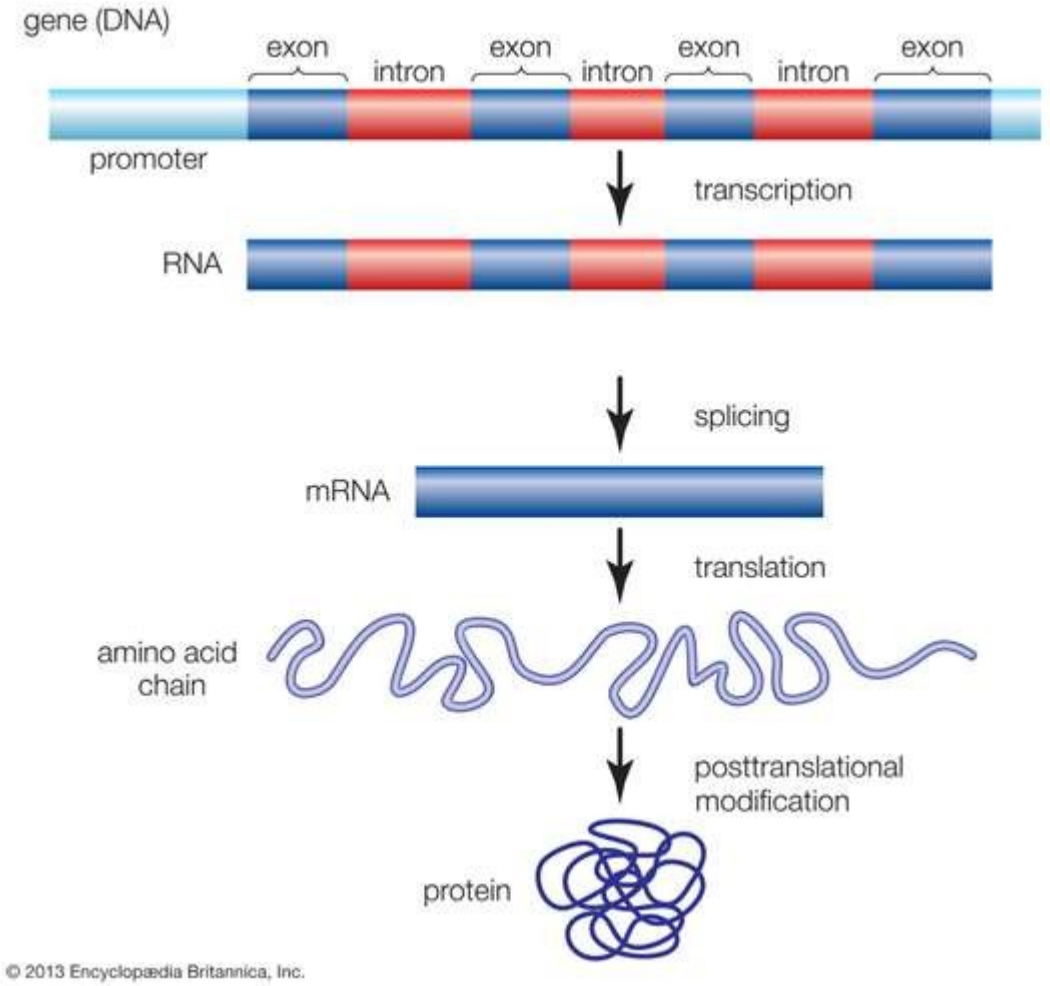
Nucleic Acids Research, Vol. 31, No. 13

Exemplos

- Predição de genes
- codificantes de proteínas



https://www.todamateria.com.br/dna/

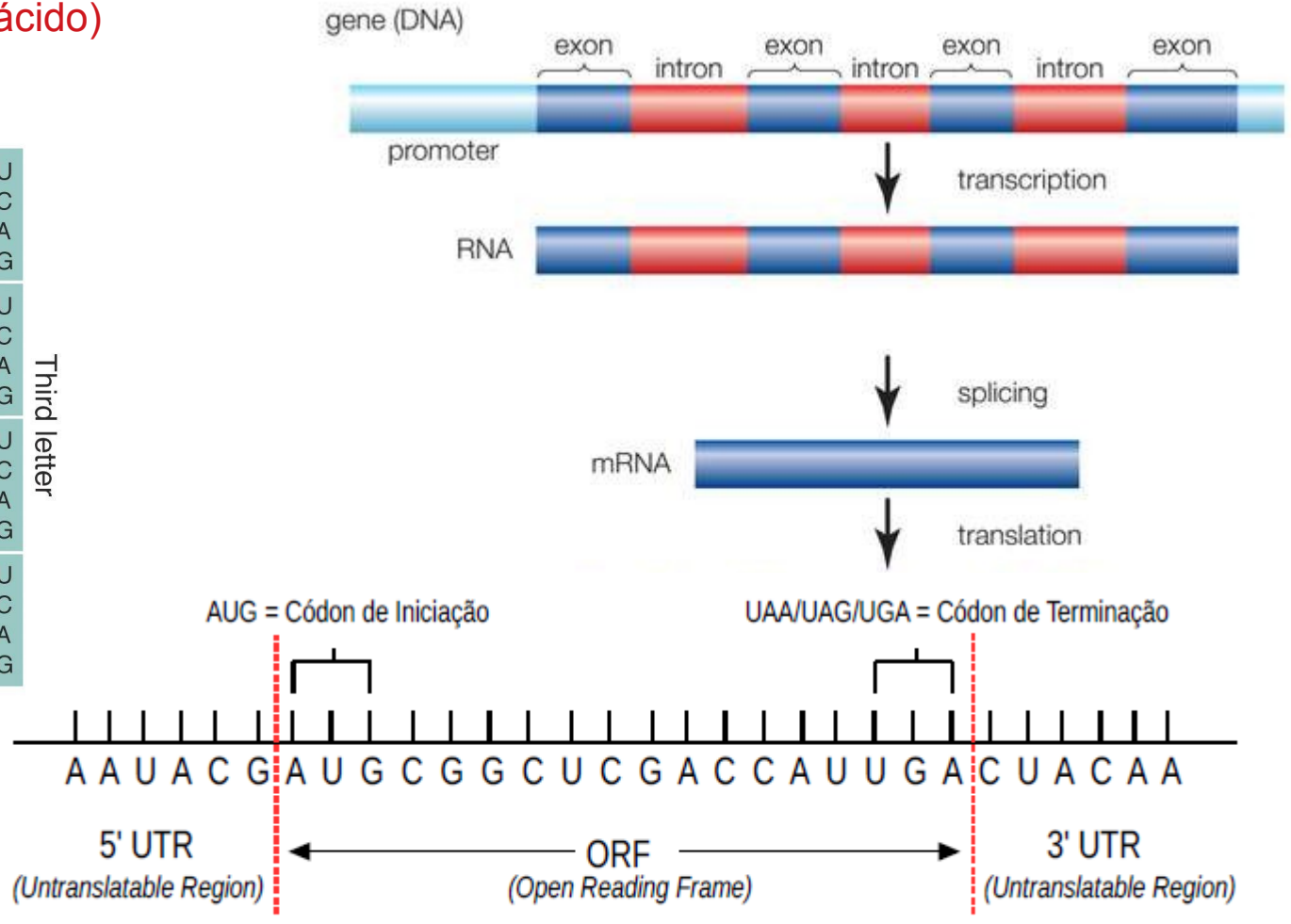


https://www.britannica.com/science/gene

Código genético (3 bases = 1 aminoácido)

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

https://sciencemusicvideos.com/wp-content/uploads/2015/01/genetic-code-tabular.jpg

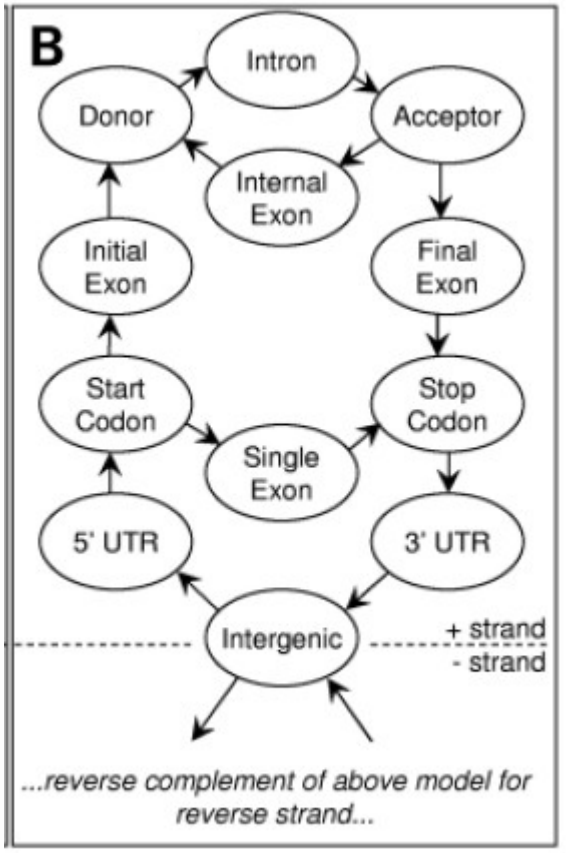


3 BASES = 1 CÓDON = 1 AMINOÁCIDO DA PROTEÍNA

Diss.USP Clebiano da Costa Sá

Exemplo de uso do Viterbi

- Preditor de gene Exonomy

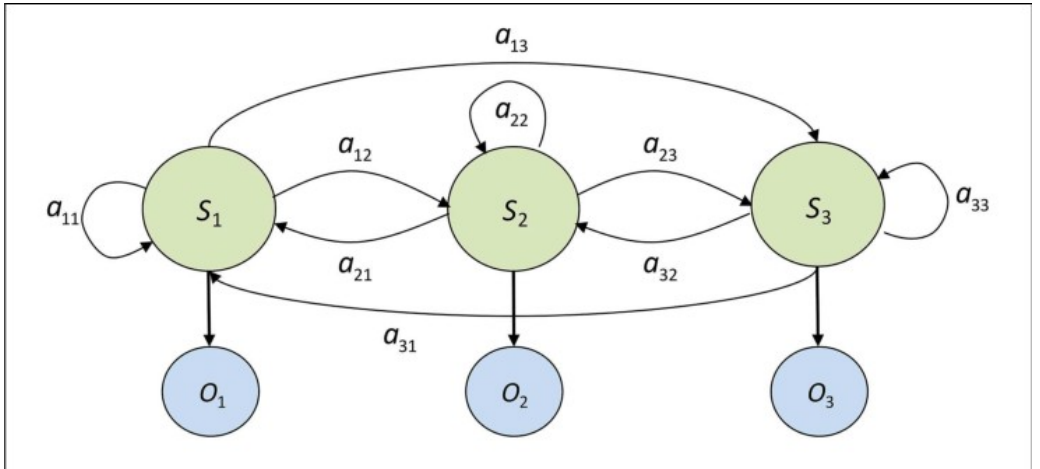


Nucleic Acids Research, Vol. 31, No. 13

Posterior decoding

- Diferente de identificar o melhor caminho, o objetivo pode ser identificar “qual o ESTADO mais provável no instante i dada a cadeia x”
  - $P(r_i = q_k \mid x, \theta)$

Ex: qual o escritor de um trecho de texto escrito por várias pessoas?



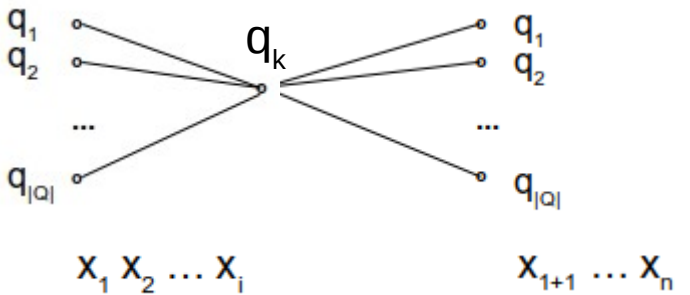
Front. Psychol. 6:919.
doi: 10.3389/fpsyg.2015.00919

Posterior decoding

- Diferente de identificar o melhor caminho, o objetivo pode ser identificar "qual o ESTADO mais provável no instante i dada a cadeia x"
  - $P(r_i = q_k \mid x, \theta)$ $\qquad P(A|B) = P(A,B)/P(B)$
  - $P(r_i = q_k \mid x, \theta) = P(x, r_i = q_k \mid \theta) / P(x \mid \theta)$
- $P(x, r_i = q_k \mid \theta) = P(x_1 \ldots x_i, r_i = q_k \mid \theta)\, P(x_{i+1} \ldots x_n | x_1 \ldots x_i, r_i = q_k, \theta)$

  $P(x, r_i = q_k \mid \theta) = P(x_1 \ldots x_i, r_i = q_k \mid \theta)\, P(x_{i+1} \ldots x_n | r_i = q_k, \theta)$

  $P(x, r_i = q_k \mid \theta) = f_k(i)\, b_k(i)$
- $P(r_i = q_k \mid x, \theta) = f_k(i)\, b_k(i) / P(x \mid \theta)$



Fim do vídeo 1

O que estamos estudando mesmo?

Exemplo Viterbi e posterior decoding

Vídeo 2

Treinamento de HMMs

Problemas relacionados a HMM

1) Dados um HMM e uma cadeia, calcular a probabilidade dessa cadeia - Algoritmo forward ou backword
2) Dados um HMM e uma cadeia, calcular o caminho mais provável dessa cadeia - Algoritmo viterbi
3) Dados um HMM e um conjunto de cadeias (amostra de treinamento), estimar os parâmetros (probabilidades de emissão e transição) - Algoritmo Baum-Welch
4) Projetar a topologia de uma HMM

Problemas relacionados a HMM

1) Dados um HMM e uma cadeia, calcular a probabilidade dessa cadeia - Algoritmo forward ou backword
2) Dados um HMM e uma cadeia, calcular o caminho mais provável dessa cadeia - Algoritmo viterbi
3) **Dados um HMM e um conjunto de cadeias (amostra de treinamento), estimar os parâmetros (probabilidades de emissão e transição) - Algoritmo Baum-Welch**
4) **Projetar a topologia de uma HMM**

Problemas relacionados a HMM

1) Dados um HMM e uma cadeia, calcular a probabilidade dessa cadeia - Algoritmo forward ou backword

2) Dados um HMM e uma cadeia, calcular o caminho mais provável dessa cadeia - Algoritmo viterbi

3) **Dados um HMM e um conjunto de cadeias (amostra de treinamento), estimar os parâmetros (probabilidades de emissão e transição) - Algoritmo Baum-Welch**

4) **Projetar a topologia de uma HMM**

Estimação dos parâmetros probabilísticos

- Aprendizado das probabilidades com base na amostra de treinamento $S = \{x^1, x^2, \ldots x^N\}$
- Para um dado conjunto de valores de parâmetros θ, a probabilidade conjunta $P(S \mid \theta)$:

$$P(x^1, x^2, \ldots x^N \mid \theta) = \prod_{j=1}^{N} P(x^j \mid \theta)$$

Estimação dos parâmetros probabilísticos
Caso 1: caminhos conhecidos

- Aprendizado das probabilidades com base na amostra de treinamento $S = \{x^1, x^2, \ldots x^N\}$
- Para um dado conjunto de valores de parâmetros θ, a probabilidade conjunta $P(S \mid \theta)$:

$$P(x^1, x^2, \ldots x^N \mid \theta) = \prod_{j=1}^{N} P(x^j \mid \theta)$$

Vamos utilizar as contagens de estados iniciais, transições e emissões por esses caminhos !!!

Lembrando os nomes das variáveis

Q: conjunto de estados ocultos

Σ: conjunto de símbolos de emissão (não confundir com o símbolo de somatório)

Probabilidades **iniciais** $\pi: Q \rightarrow [0,1]$

$\Sigma_k = \pi(q_k) = 1$

Probabilidades de **transições**

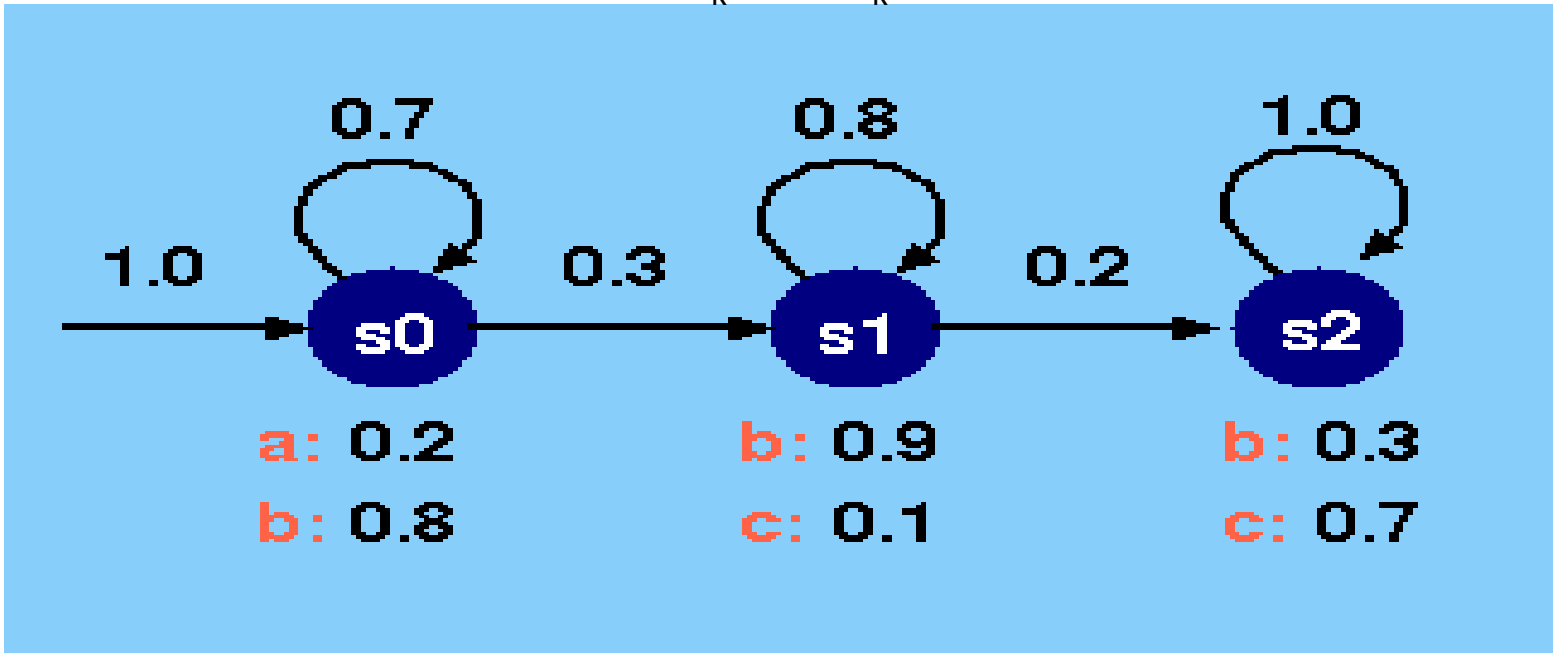
$t_{kl} = P(q_k \mid q_l)$

$\Sigma_l t_{kl} = 1$

Probabilidades de **emissões**

$e_k(a) = P(a \mid q_k)$

$\Sigma_{a \epsilon \Sigma} e_k(a) = 1$



Estimação por máxima verossimilhança

Probabilidades **iniciais** $\pi: Q \rightarrow [0,1]$

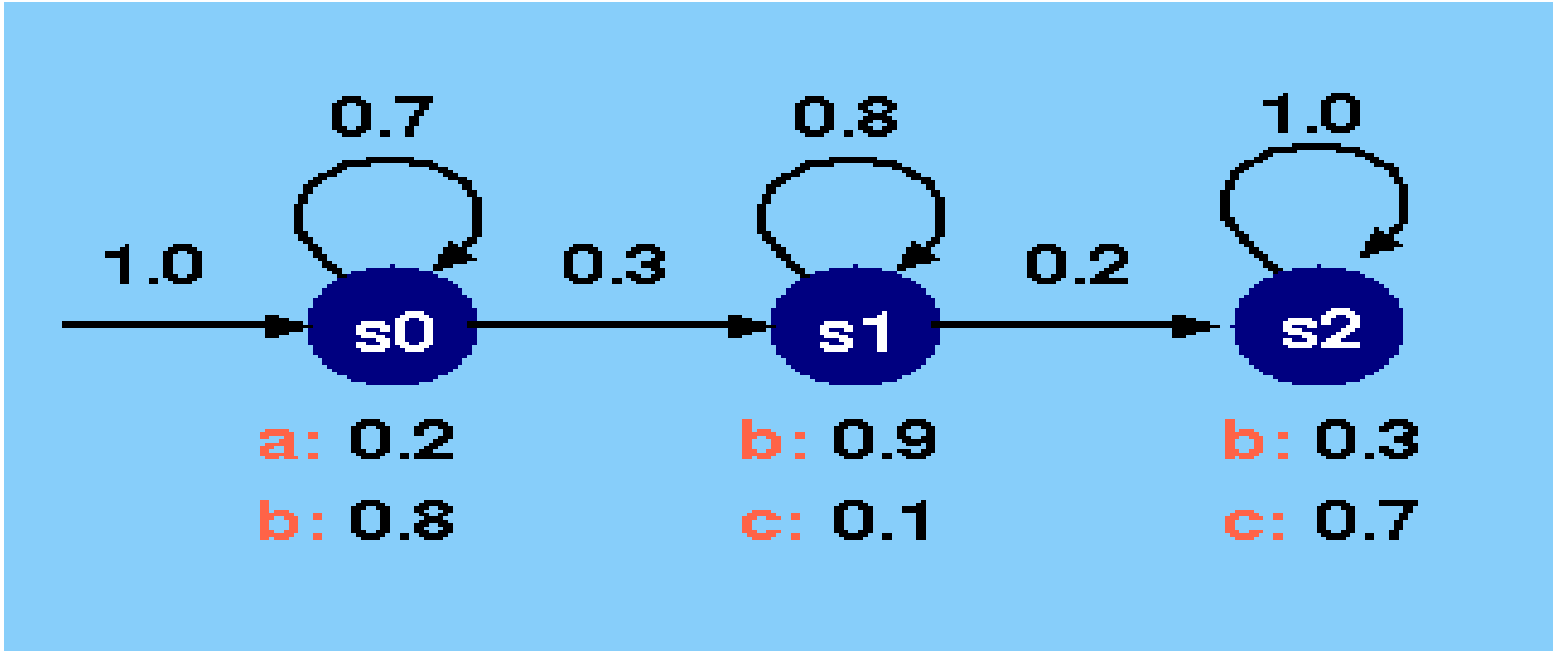
$\Sigma_k = \pi(q_k) = 1$

Simplificando a notação:

$\pi(q_k) = \pi_k$

$\hat{\pi}_k = \#\Pi_k / \Sigma_j \#\Pi_j$

$\#\Pi_k$: nr de inícios no estado q_k

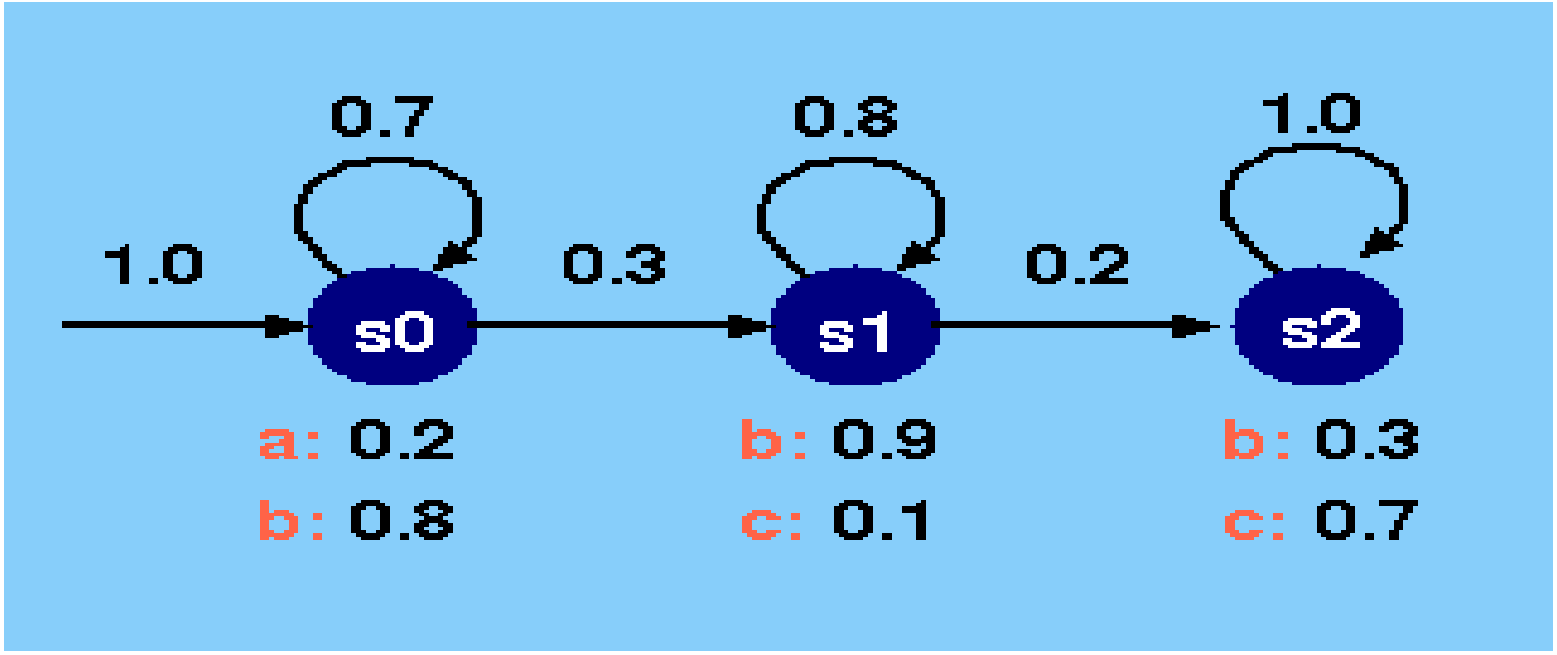


Estimação por máxima verossimilhança

Probabilidades
de **transições**
$t_{kl} = P(q_l \mid q_k)$
$\Sigma_l \, t_{kl} = 1$

$\hat{t}_{kl} = \#T_{kl} \, / \, \Sigma_l \, \#T_{kl}$

$\#T_{kl}$: nr de transições do estado k para o estado l



Estimação por máxima verossimilhança

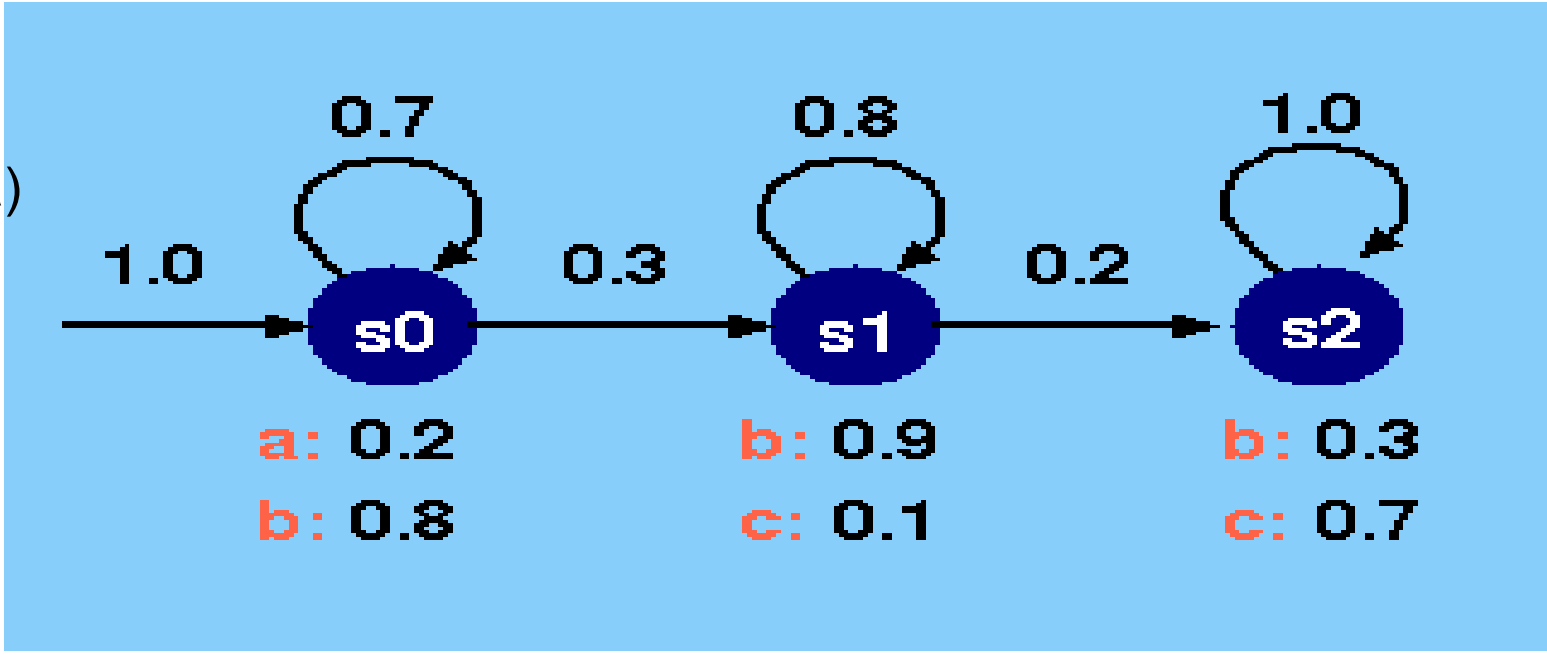
Probabilidades
de **emissões** no estado q_k

$e_k(a) = P(a \mid q_k)$

$\Sigma_{a \epsilon \Sigma}\, e_k(a) = 1$

$\hat{e}_k(a) = \#E_k(a) / \Sigma_{a \epsilon \Sigma} \#E_k(a)$

$\#E_k(a)$: nr de emissões do símbolo “a” no estado k



E se alguma contagem for 0?

Estimação por máxima ***posteriori (MAP)***

Probabilidades **iniciais** $\pi: Q \rightarrow [0,1]$

$\Sigma_k = \pi(q_k) = 1$

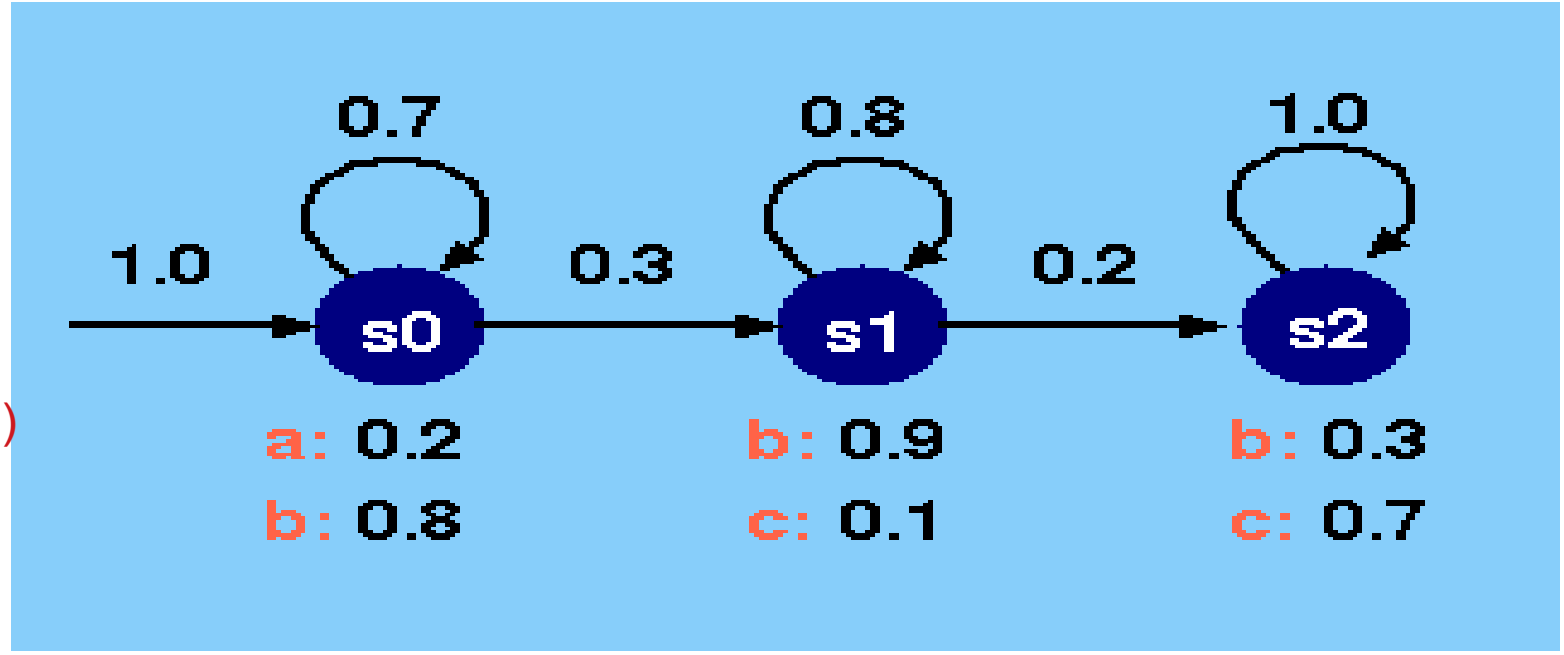
Simplificando a notação:

$\pi(q_k) = \pi_k$

$\hat{\pi}_k = \#\Pi_k / \Sigma_j \#\Pi_j$

$\#\Pi_k$: nr de inícios no estado $q_k + \alpha_k$

α_k : pseucontador (> 0)

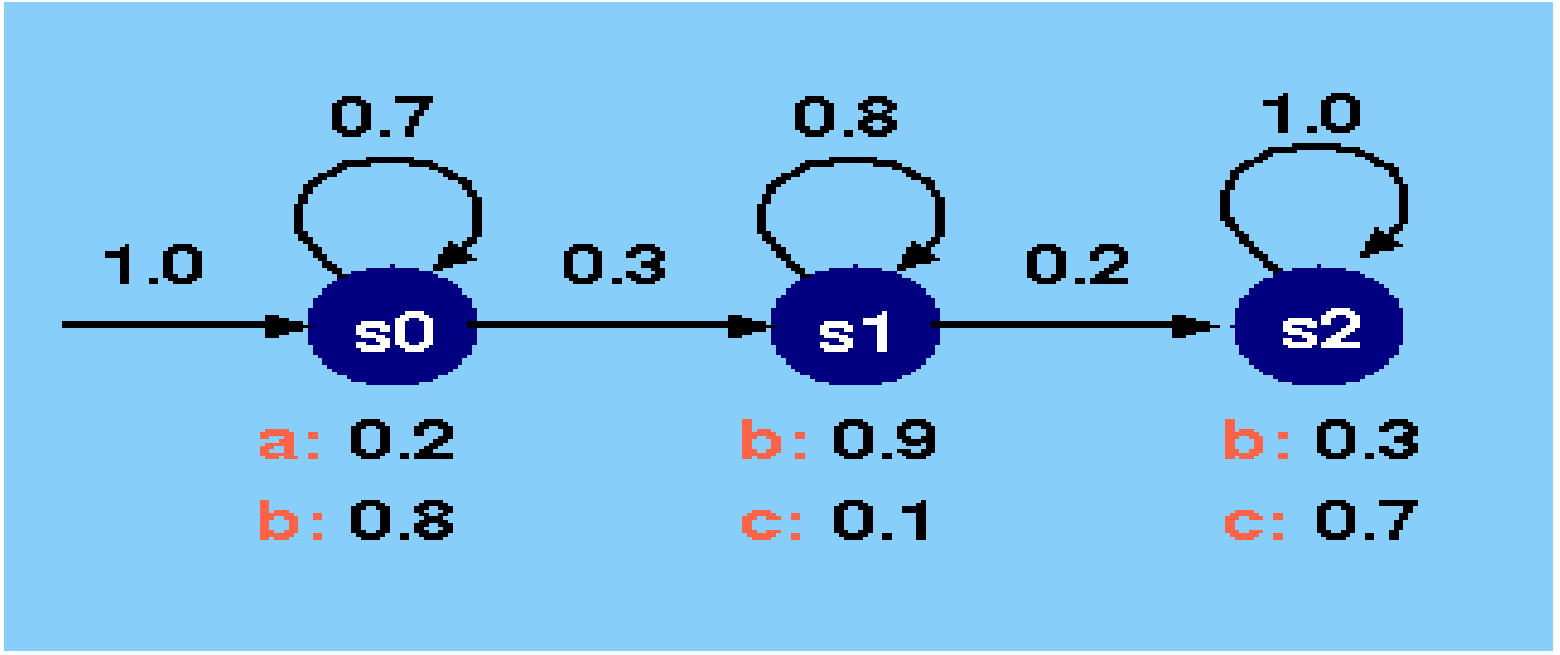


Estimação por máxima *posteriori* (MAP)

Probabilidades
de **transições**
$t_{kl} = P(q_k \mid q_l)$
$\Sigma_l \, t_{kl} = 1$

$\hat{t}_{kl} = \#T_{kl} \, / \, \Sigma_l \, \#T_{kl}$

$\#T_{kl}$: nr de transições do estado k para o estado l + α_{kl}
α_{kl} : pseucontador (> 0)



Estimação por máxima *posteriori* (MAP)

Probabilidades
de **emissões** no estado q_k

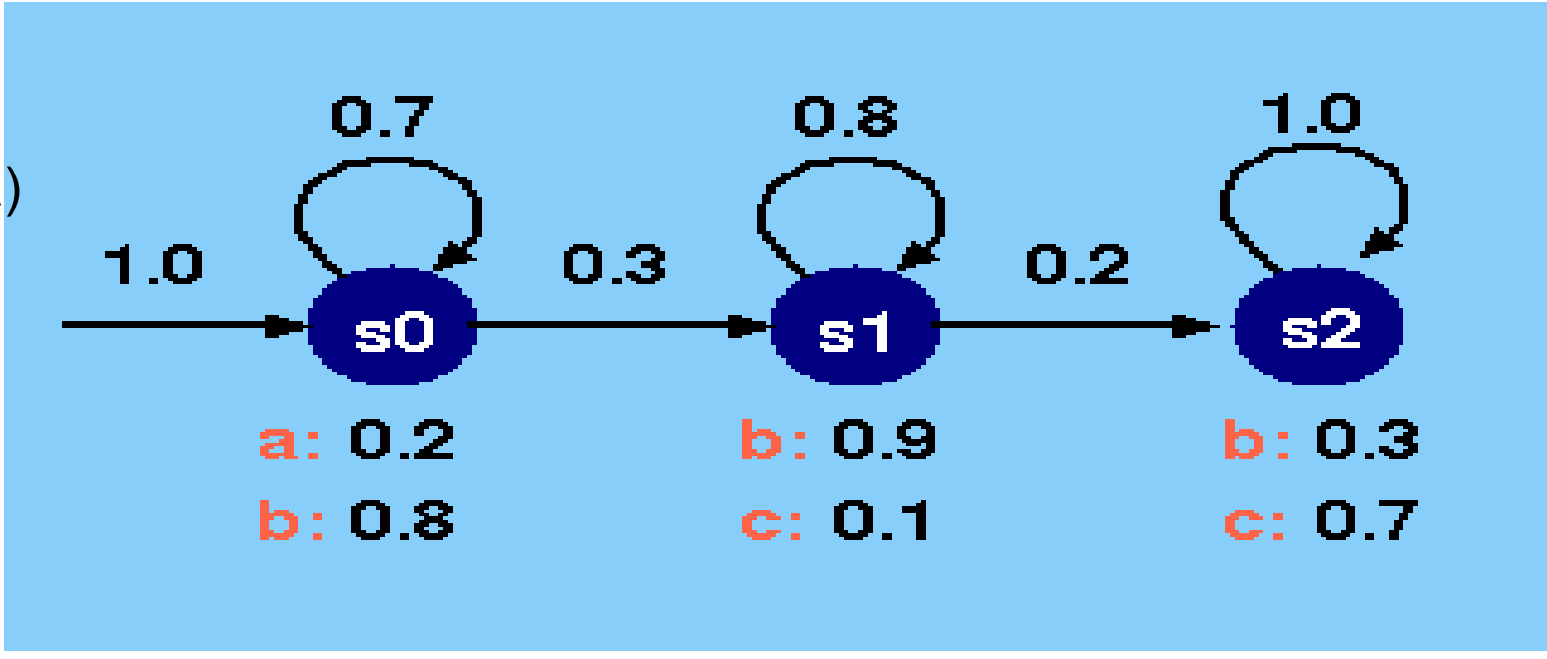
$e_k(a) = P(a \mid q_k)$

$\Sigma_{a \epsilon \Sigma}\, e_k(a) = 1$

$\hat{e}_k(a) = \#E_k(a) / \Sigma_{a \epsilon \Sigma} \#E_k(a)$

$\#E_k(a)$: nr de emissões do símbolo "a" no estado k $+ \alpha_{k(a)}$

$\alpha_{k(a)}$: pseucontador (> 0)



Estimação por máxima ***posteriori*** **(MAP)**

- Valores dos pseudocontadores: positivos, não necessariamente inteiros, representam informação *a priori*
- Quanto maior a somatória dos pseudocontadores, maior o peso da informação *a priori*, e mais dados são necessários para mudá-la
  - $\Sigma_k \alpha_k$, para estados iniciais
  - $\Sigma_l \alpha_{kl}$, para transições partindo do estado q_k
  - $\Sigma_a \alpha_{k(a)}$, para emissões no estado q_k
- Valores iguais → *priori* não informativa

Estimação dos parâmetros probabilísticos
Caso 2: caminhos NÃO conhecidos

- Aprendizado das probabilidades com base na amostra de treinamento $S = \{x^1, x^2, \ldots x^N\}$
- Para um dado conjunto de valores de parâmetros θ, a probabilidade conjunta $P(S \mid \theta)$:

$$P(x^1, x^2, \ldots x^N \mid \theta) = \prod_{j=1}^{N} P(x^j \mid \theta)$$

Precisamos estimar os parâmetros e “adivinhar” o melhor caminho ao mesmo tempo… de forma iterativa
Técnica EM (Expectation-Maximization)

Algoritmo Baum-Welch

- Ideia:

1) Valores iniciais das probabilidades (θ_0)

2) Cálculo dos caminhos mais prováveis de cada cadeia de S (utilizando o Viterbi, com base em θ_0) - Expectation

3) Recálculo das probabilidades com base nesses caminhos, utilizando o alg. MAP dos slides anteriores para o caso 1 (θ_1) - Maximization

4) Refaça os passos 2 e 3 até alcançar um critério de parada

Algoritmo Baum-Welch

- Como fazer isso de um jeito espertinho?
- A probabilidade de uma transição do estado q_k para um estado q_l (t_{kl}) em uma dada posição i de uma dada cadeia x é:

  $P(r_i = q_k, r_{i+1} = q_l \mid x, \theta) = f_k(i)\, t_{kl}\, e_l(x_{i+1})\, b_l(i+1) / P(x|\theta)$

- O número ESPERADO de vezes em que é realizada a transição do estado q_k para o estado q_l considerando TODAS as posições e TODAS as cadeias da amostra de treinamento é:

  $\#T_{kl} = \sum_j 1/P(x^j|\theta) \sum_i f_k^j(i)\, t_{kl}\, e_l(x_{i+1}^j)\, b_l^j(i+1)$

Algoritmo Baum-Welch

- Como fazer isso de um jeito espertinho?
- A probabilidade de emissão do símbolo *a* em um estado q_k ($e_{k(a)}$) em uma dada posição i de uma dada cadeia x é:

  $$P(x_i = a, r_i = q_k \mid x, \theta) = f_k(i)\, b_k(i) / P(x|\theta)$$

- O número ESPERADO de vezes em que o símbolo *a* é emitido no estado q_k considerando TODAS as posições e TODAS as cadeias da amostra de treinamento é:

  $$\#E_{k(a)} = \sum_j 1/P(x^j|\theta) \sum_{\{ i \mid x_i^j = a\}} f_k^j(i)\, b_k^j(i)$$

Algoritmo Baum-Welch

θ = valores iniciais das probabilidades

$\#\Pi_k = \alpha_k$, $\#T_{kl} = \alpha_{kl}$, $\#E_k(a) = \alpha_{k(a)}$ para todo k, l, a

Faça:

Para cada cadeia $x^j \in S=\{x^1, x^2, \ldots, x^N\}$

calcule forward f^j e backward b^j para x^j

$\#T_{kl} \mathrel{+}= 1/P(x^j|\theta) \sum_i f_k^j(i)\, t_{kl}\, e_l(x^j_{i+1})\, b_l^j(i+1)$

$\#E_{k(a)} \mathrel{+}= 1/P(x^j|\theta) \sum_{\{i \mid x^j_i = a\}} f_k^j(i)\, b_k^j(i)$

Recalcule θ (slides 25 a 27)

$P(S|\theta) = \Pi_{j=1} P(x^j|\theta)$

Enquanto não alcançar critério de parada

Algoritmo Baum-Welch

θ = valores iniciais das probabilidades

$\#\Pi_k = \alpha_k$, $\#T_{kl} = \alpha_{kl}$, $\#E_k(a) = \alpha_{k(a)}$ para todo k, l, a

Faça:

Para cada cadeia $x^j \in S=\{x^1, x^2, \ldots, x^N\}$

calcule forward f^j e backward b^j para x^j

$\#T_{kl} \mathrel{+}= 1/P(x^j|\theta) \sum_i f_k^j(i)\, t_{kl}\, e_l(x^j_{i+1})\, b_l^j(i+1)$

$\#E_{k(a)} \mathrel{+}= 1/P(x^j|\theta) \sum_{\{i \mid x^j_i = a\}} f_k^j(i)\, b_k^j(i)$

Recalcule θ (slides 25 a 27)

$P(S|\theta) = \Pi_{j=1} P(x^j|\theta)$

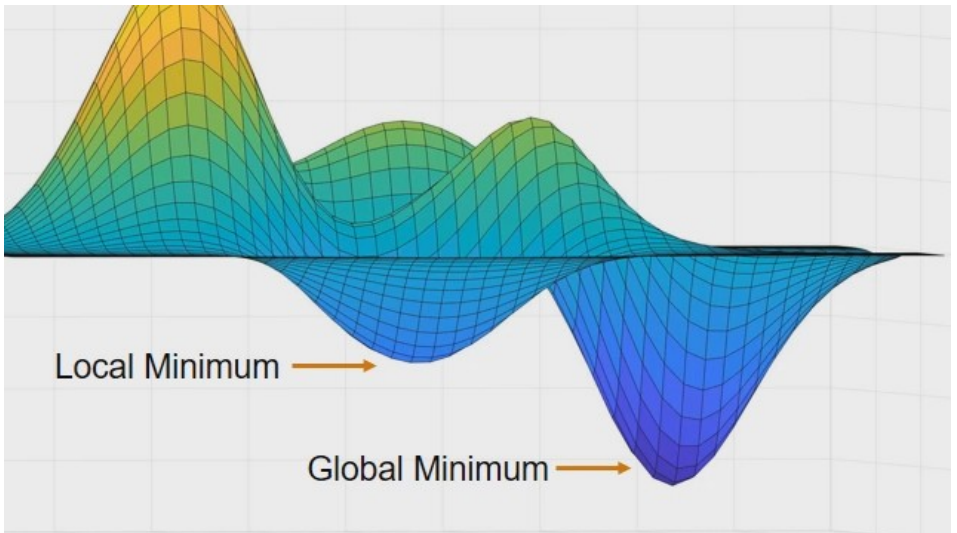
Pouca variação na verossimilhança de S
ou
número máximo de iterações

Enquanto não alcançar critério de parada

Mínimos locais

É bom lembrar que, como todo EM, sobre de mínimos locais… (máximos locais nesse caso)

Testar vários diferentes valores de inicialização e escolher o resultado que, por exemplo, maximizar a verossimilhança P(S | θ)



https://www.mathworks.com/products/global-optimization.html

Fim do vídeo 2

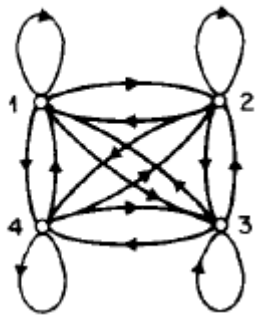
Treinamento de HMMs

Vídeo 3

Topologia de HMMs

Topologia

- Uma possível topologia é definir um certo número de estados completamente conectados (**modelo ergódico**)

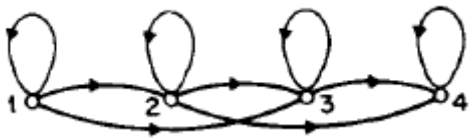


PROCEEDINGS OF THE IEEE, VOL. 77, NO. 2, FEBRUARY 1989

- Esses modelos porém podem não ser adequados para certas aplicações
  - Em que há uma “temporalidade” no modelo
  - Tendem a sofrer mais com os mínimos locais

Topologia

- **Modelos left-right** capturam “temporalidade”



PROCEEDINGS OF THE IEEE, VOL. 77, NO. 2, FEBRUARY 1989

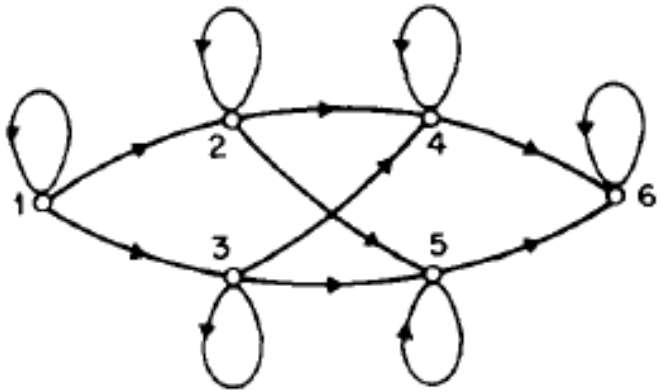
- Nesses modelos não há transições “para trás” ($t_{kl} = 0$ se $l < k$),

  $\pi_1 = 1$

  e normalmente não muito para frente ($t_{kl} = 0$ se $l-k > \Delta$)

Topologia

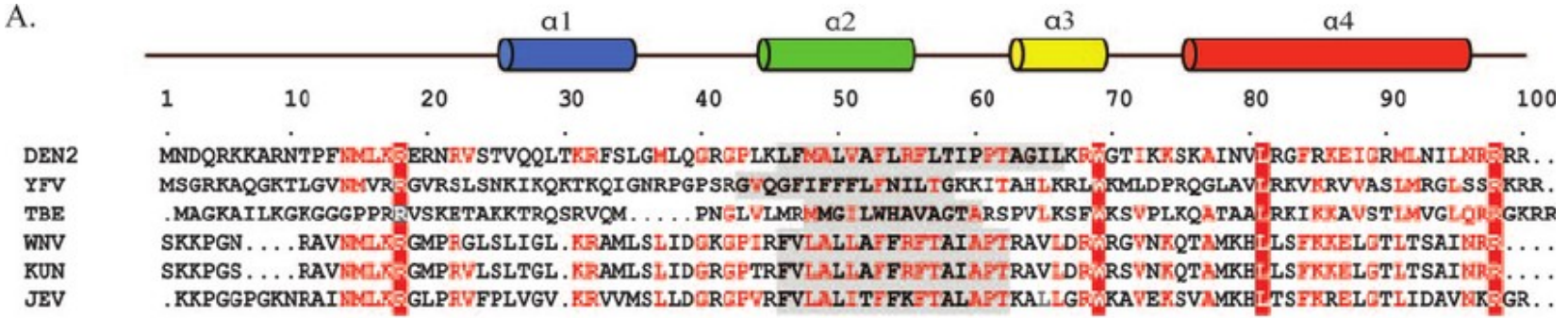
- E o que mais a criatividade permitir…
- Ex: 2 modelos left-right paralelos e cruzados:



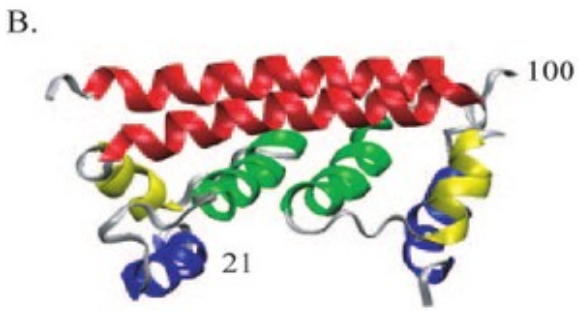
PROCEEDINGS OF THE IEEE, VOL. 77, NO. 2, FEBRUARY 1989

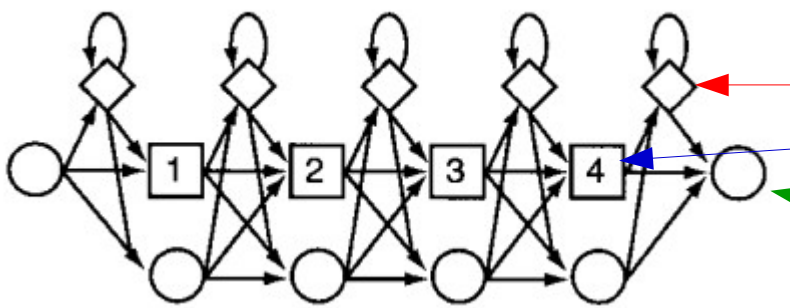
Ex: HMMER

- Profile HMM para domínios proteicos



DOI:
10.1128/JVI.02120-6





Estados de inserção
Estados de match
Estados de deleção (não emitem nada)

BIOINFORMATICS REVIEW
Vol. 14 no. 9 1998
Pages 755-763

Fim do vídeo 3

Topologia de HMMs

Vídeo 4

Estimação de desempenho

Professora:
Ariane Machado Lima

Contextualizando

- Reconhecimento de padrões: aprendizado de um classificador (binário ou multiclasse)

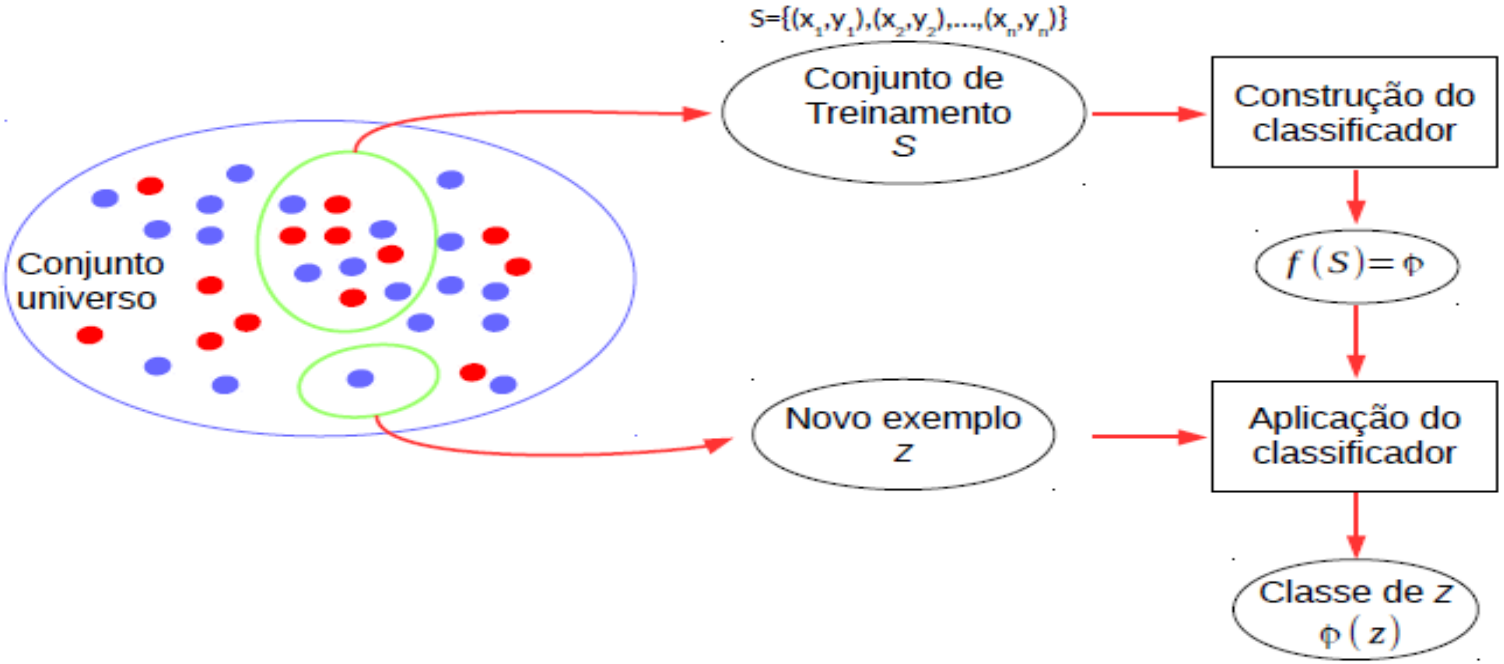
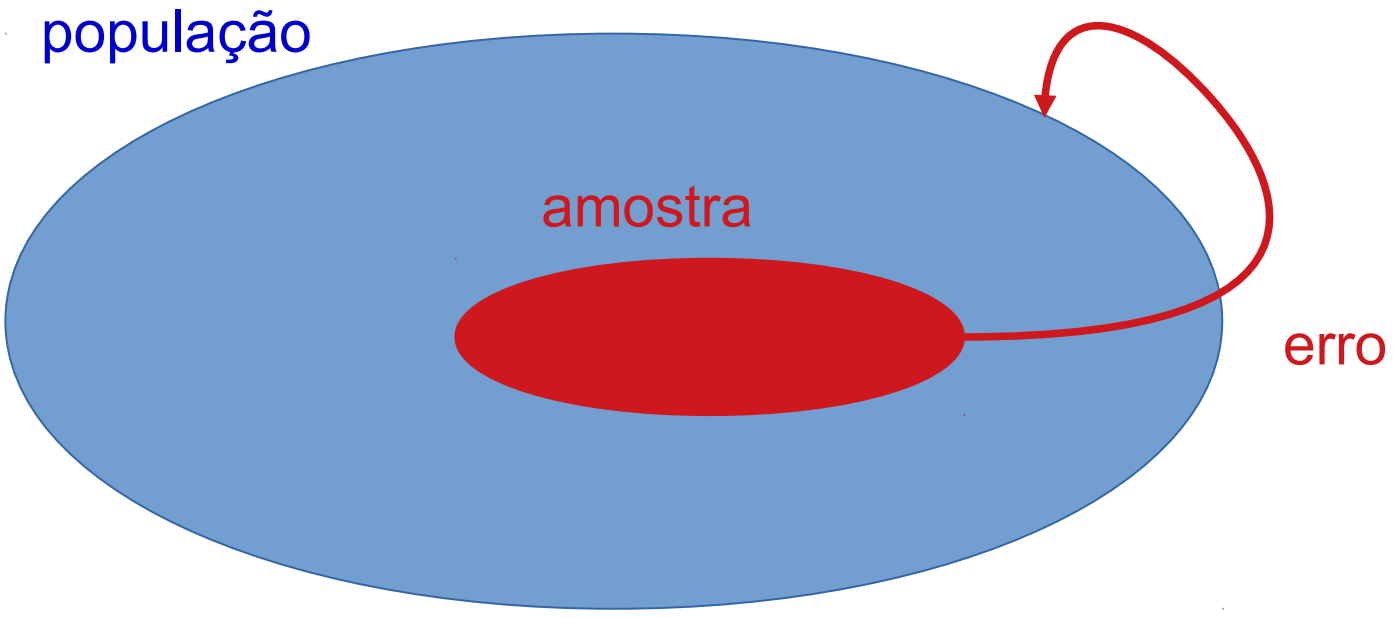


Figura: Representação do processo de aprendizado supervisionado

Aprendizado

Aprendizado a partir de uma amostra → HÁ ERRO

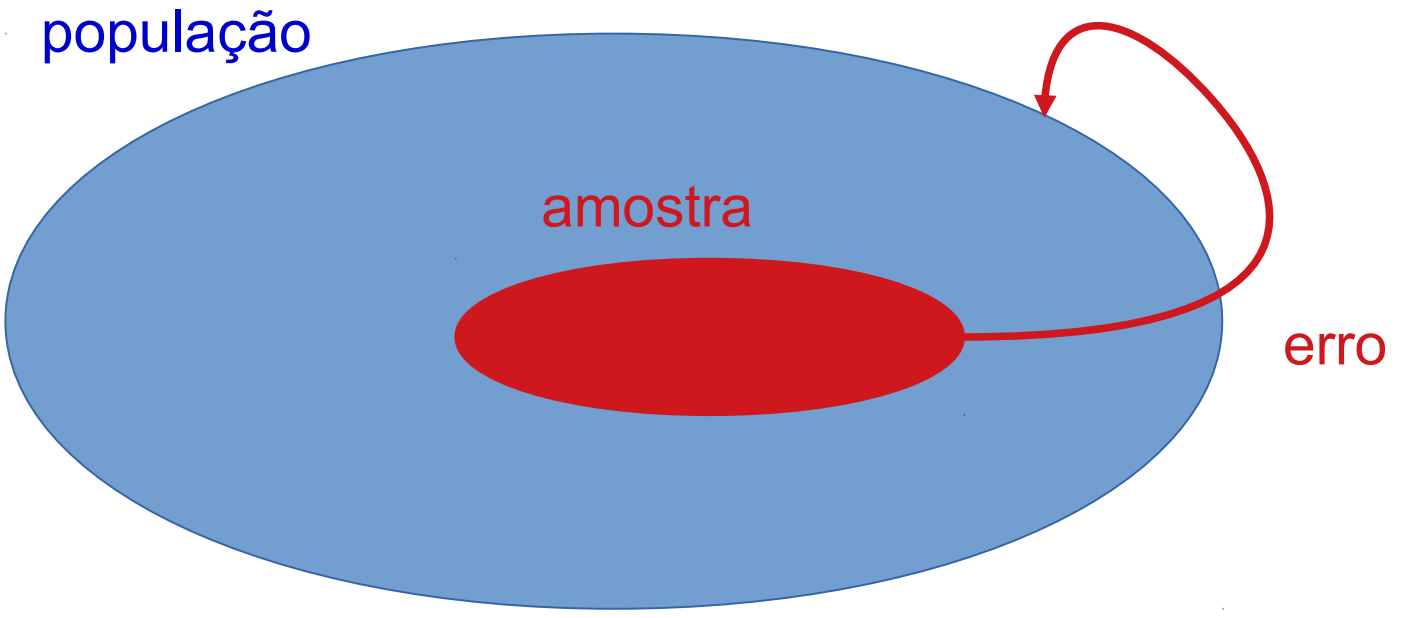


Como avaliar um classificador?

- Uma amostra para treinar e uma amostra para testar
- Os erros na amostra de teste são uma estimativa do erro

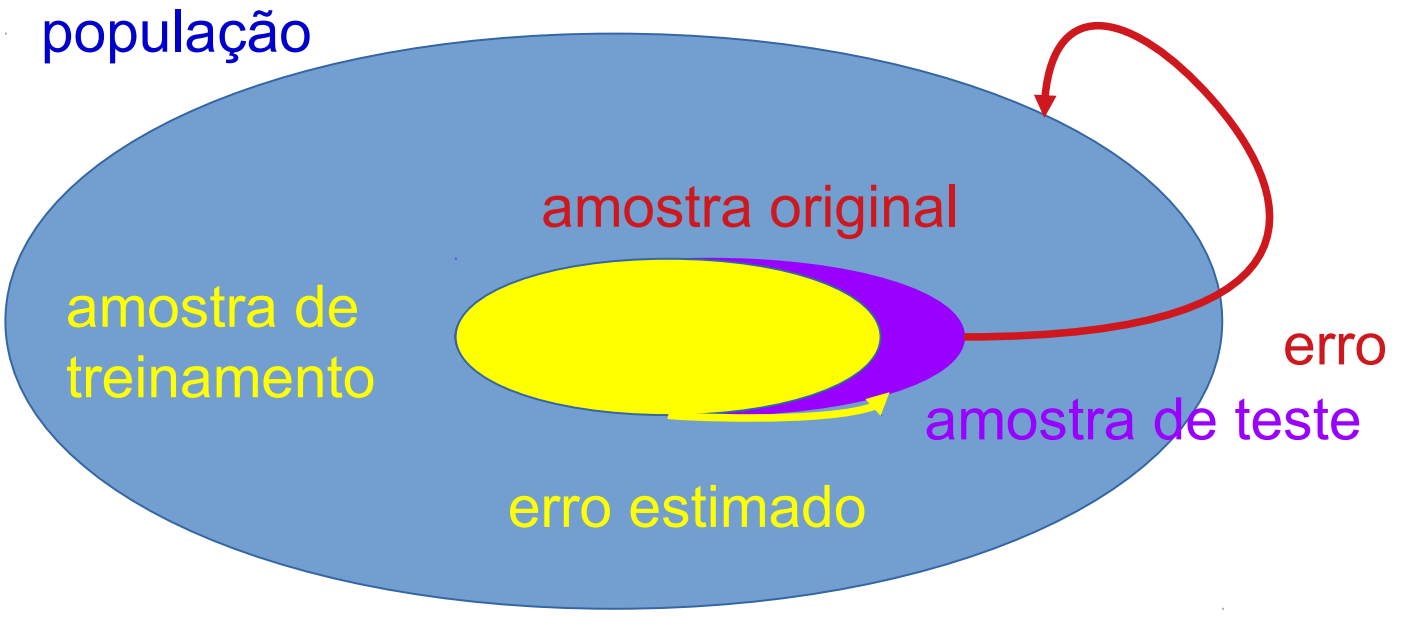
Aprendizado do erro

Aprendizado do erro a partir de amostras da amostra → HÁ ERRO acerca do erro!



Aprendizado do erro

Aprendizado do erro a partir de amostras da amostra → HÁ ERRO acerca do erro!



Estimação de erro de um classificador

- As amostras de treinamento e de teste deveriam ser:
  - Grandes
  - Independentes
  - Mas nem sempre conseguimos...
- Como utilizar uma dada amostra para isso?
  - Vários métodos de estimação de erro

Vídeo 5

Técnicas de estimação de desempenho

Métodos de estimação de erro de um classificador

- Resubstituição
- Holdout
- Leave-one-out
- k-fold cross-validation (validação cruzada k-vezes)
- Subamostragem aleatória
- Bootstrap

Resubstituição

- Toda a amostra original é usada para treinamento e depois para teste

- Problema?

Resubstituição

- Toda a amostra original é usada para treinamento e depois para teste
- Fornece uma estimativa otimista do erro
- Não revela se está havendo *overfitting*
- Quanto menor a amostra de treinamento, pior a estimativa
- Pior opção

Holdout

- Uma parte da amostra original é usada para treinamento e o restante para teste (não necessariamente 50%)

- Problema:

http://www.ebc.cat/2017/01/31/cross-validation-strategies/

Holdout

- Uma parte da amostra original é usada para treinamento e o restante para teste (não necessariamente 50%)

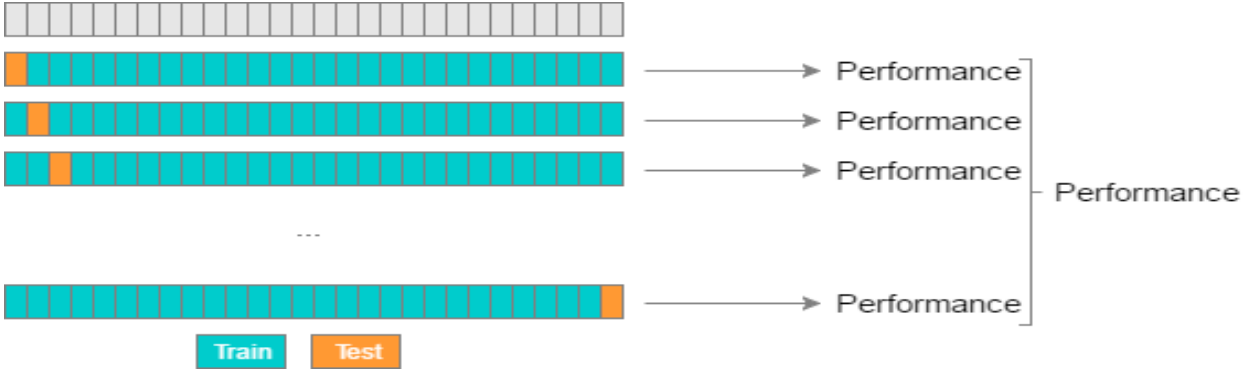
- Problema: diferentes divisões provavelmente darão diferentes estimativas

http://www.ebc.cat/2017/01/31/cross-validation-strategies/

Leave-one-out

- Se a amostra original tem n dados, treina com n-1 dados e testa no que restou
- Repetir o processo n vezes, cada vez deixando um dado de fora para teste
- A estimativa de erro é a média do erro de cada rodada

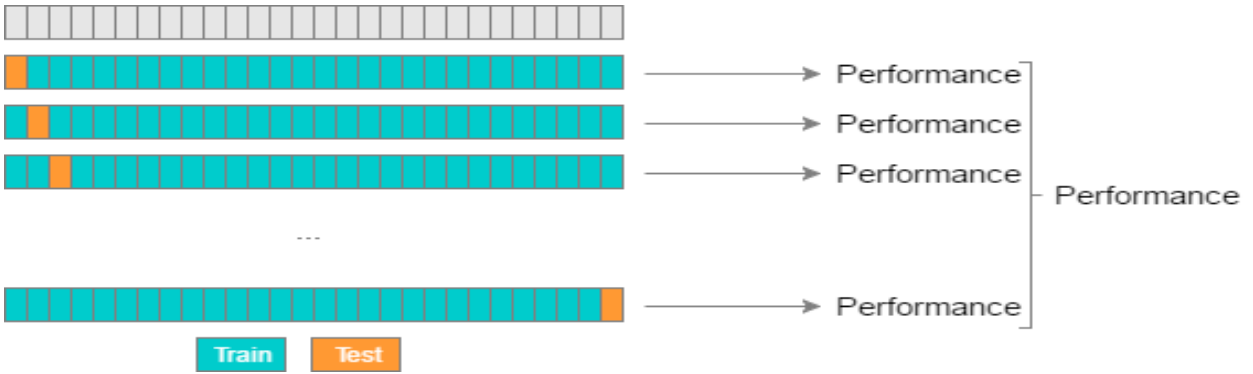
Obs: testa todas as combinações possíveis (exaustivo)



http://www.ebc.cat/2017/01/31/cross-validation-strategies/

Leave-one-out

- Estimativa menos enviesada
- Alta variância entre os n testes (erro = 0 ou 1)
- Alto custo computacional (n treinamentos e testes de classificadores)

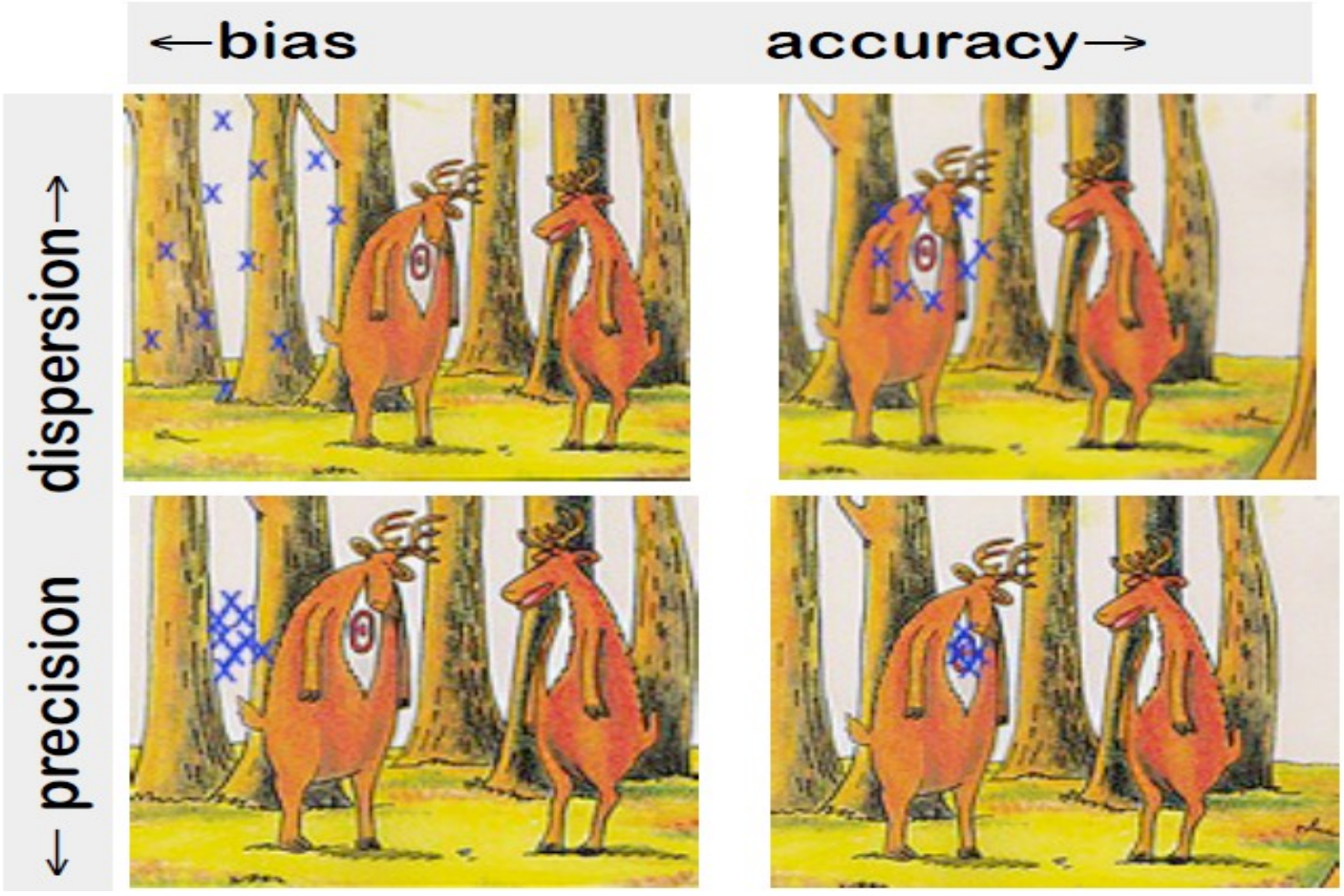


http://www.ebc.cat/2017/01/31/cross-validation-strategies/

Viés (bias)

x variância

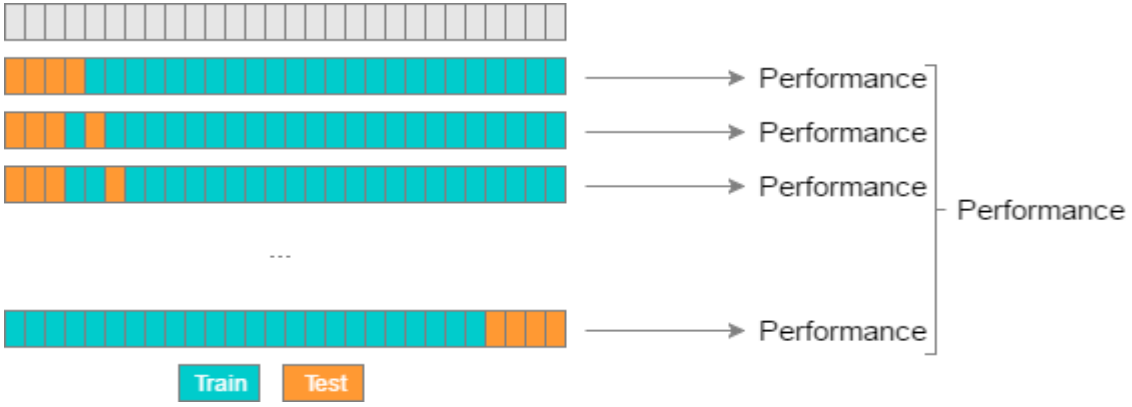
Viés: tendência (distorção) para um dado ponto; polarização.



https://bioconductor.org/help/course-materials/2010/EMBL2010/100609-multtestindepfilt-huber.pdf

Leave-p-out

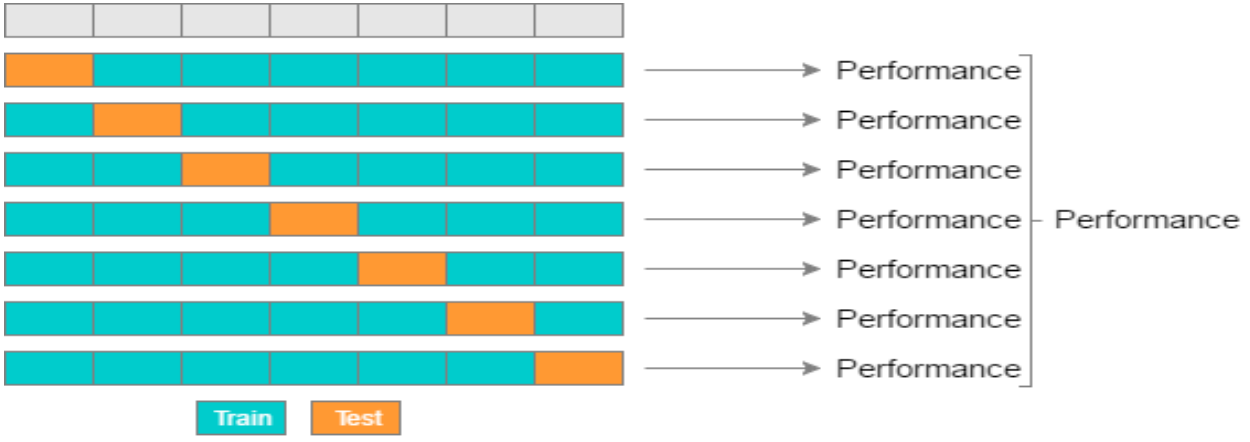
- Separa p intâncias para teste e treina com as demais
- Testa TODAs as combinações (também exaustivo)
- Computacionalmente mais caro ainda



http://www.ebc.cat/2017/01/31/cross-validation-strategies/

Validação cruzada k-vezes (k-fold cross-validation)

- Divida a amostra original em k (k < n) partes
- Treina com k-1 partes, testa com a que sobrou
- Repetir o processo k vezes, cada vez deixando uma parte diferente de fora do treinamento (para teste)



http://www.ebc.cat/2017/01/31/cross-validation-strategies/

Validação cruzada k-vezes (k-fold cross-validation)

- Divida a amostra original em k (k < n) partes
- Treina com k-1 partes, testa com a que sobrou
- Repetir o processo k vezes, cada vez deixando uma parte diferente de fora do treinamento (para teste)

Balanço entre holdout e leave-one-out:

Mais/menos? enviesada que o holdout

Mais/menos? custoso que o leave-one-out

Maior/menor? variância que a do leave-one-out

Validação cruzada k-vezes (k-fold cross-validation)

- Divida a amostra original em k (k < n) partes
- Treina com k-1 partes, testa com a que sobrou
- Repetir o processo k vezes, cada vez deixando uma parte diferente de fora do treinamento (para teste)

Balanço entre holdout e leave-one-out:

Menos enviesada que o holdout

Mais/menos? custoso que o leave-one-out

Maior/menor? variância que a do leave-one-out

Validação cruzada k-vezes (k-fold cross-validation)

- Divida a amostra original em k (k < n) partes
- Treina com k-1 partes, testa com a que sobrou
- Repetir o processo k vezes, cada vez deixando uma parte diferente de fora do treinamento (para teste)

Balanço entre holdout e leave-one-out:

Menos enviesada que o holdout

Menos custoso que o leave-one-out (k classificadores)

Maior/menor? variância que a do leave-one-out

Validação cruzada k-vezes (k-fold cross-validation)

- Divida a amostra original em k (k < n) partes
- Treina com k-1 partes, testa com a que sobrou
- Repetir o processo k vezes, cada vez deixando uma parte diferente de fora do treinamento (para teste)

Balanço entre holdout e leave-one-out:

Menos enviesada que o holdout

Menos custoso que o leave-one-out (k classificadores)

Menor variância que a do leave-one-out

Validação cruzada k-vezes (k-fold cross-validation)

- Divida a amostra original em k (k < n) partes
- Treina com k-1 partes, testa com a que sobrou
- Repetir o processo k vezes, cada vez deixando uma parte diferente de fora do treinamento (para teste)

Balanço entre holdout e leave-one-out:

Menos enviesada que o holdout

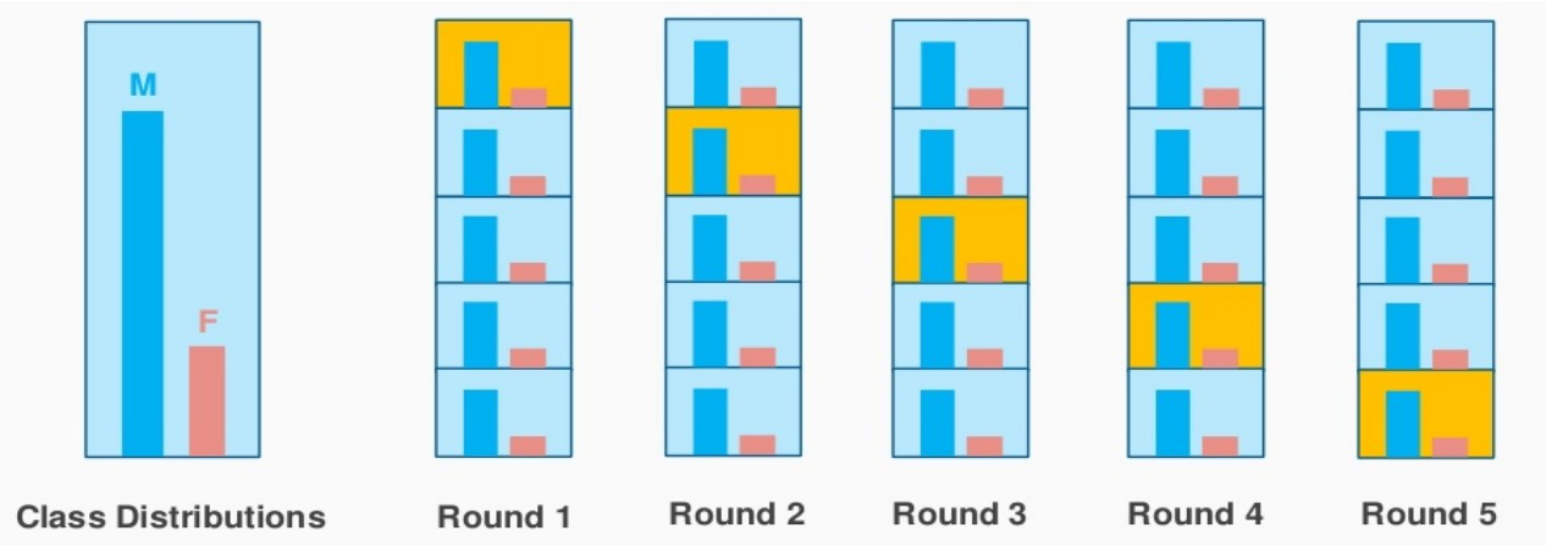
Menos custoso que o leave-one-out (k classificadores)

Menor variância que a do leave-one-out

Obs: não testa todas as combinações possíveis (não exaustivo)

Validação cruzada k-vezes ESTRATIFICADA

- Preserva em cada fold a proporção das classes presente no conjunto original



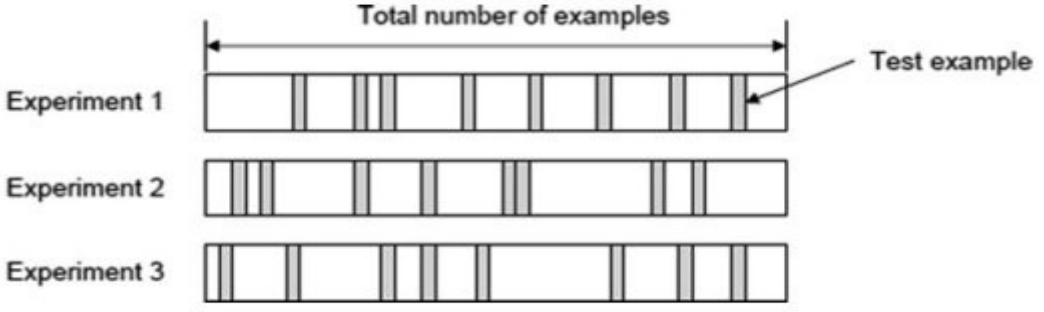
https://stats.stackexchange.com/questions/49540/understanding-stratified-cross-validation

Validações cruzadas repetidas

- Várias validações cruzadas, cada vez embaralhando a ordem das instâncias (e reestratificando-as se for o caso)

Subamostragem aleatória

- Também conhecida como múltiplos hold-outs ou validação cruzada Monte Carlo
- Semelhante à validação cruzada, mas as instâncias são sorteadas SEM reposição em cada holdout/fold (mas COM reposição a cada holdout)
- Você pode definir tamanhos das amostras e nr de holdouts
- Algumas instâncias podem nunca serem utilizadas, e outras várias vezes



DOI 10.13140/RG.2.2.18377.24167

Bootstrap

- Gera várias amostras (COM reposição) de tamanho de tamanho m, m <=> n
- Treina com uma e testa com outra

http://rasbt.github.io/mlxtend/user_guide/evaluate/bootstrap_point632_score/

Bootstrap

- Variância menor
- Computacionalmente caro
- Útil quando a amostra original é pequena

http://rasbt.github.io/mlxtend/user_guide/evaluate/bootstrap_point632_score/

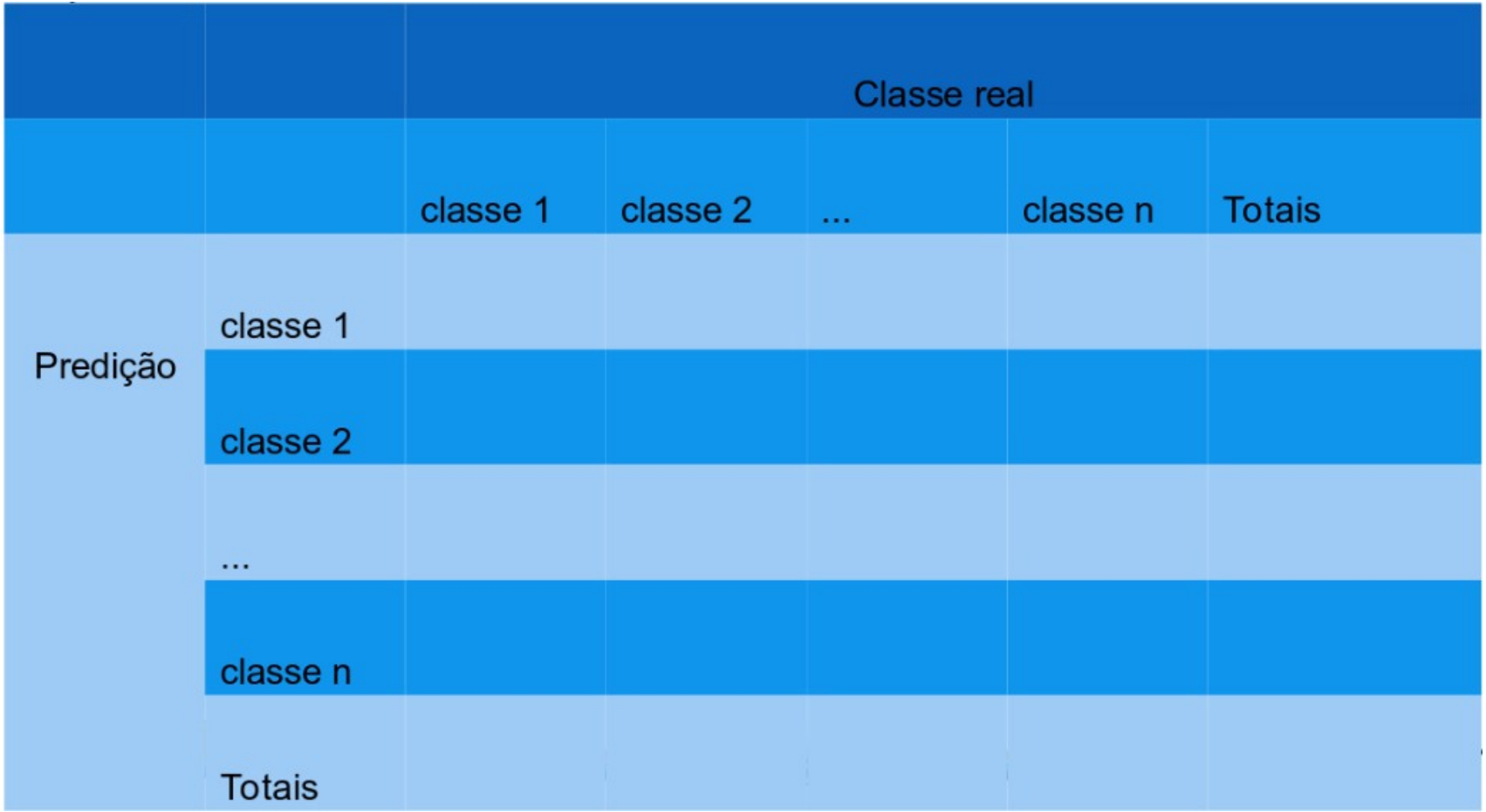
Vídeo 6

Medidas de desempenho

Medidas de desempenho

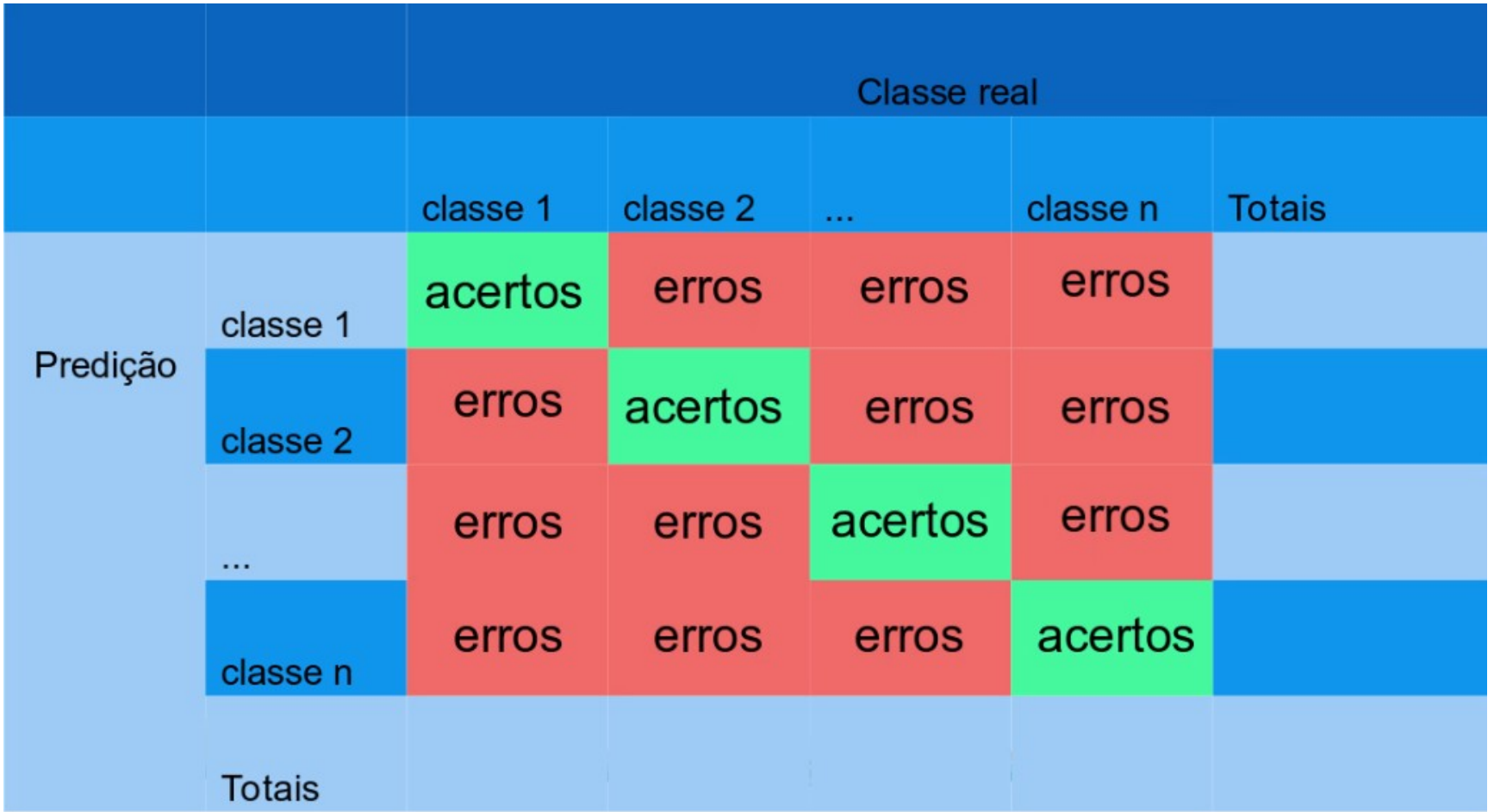
Matriz de confusão

M_{ij}: quanto elementos da classe j foram preditos como sendo da classe i



Matriz de confusão

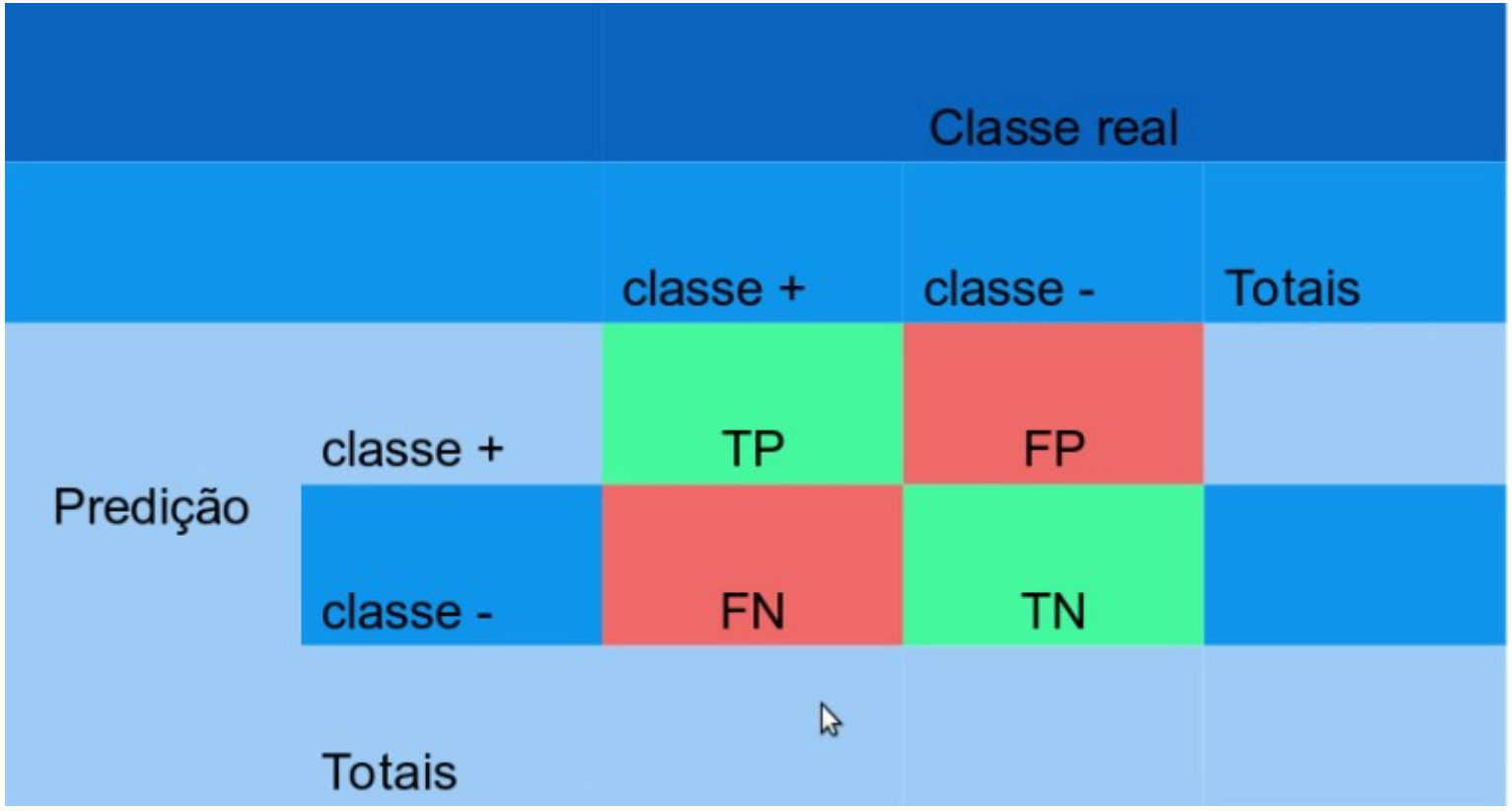
M_{ij}: quanto elementos da classe j foram preditos como sendo da classe i



Medidas de desempenho

- Classificação binária: considere uma classe positiva
  - Ex: peça defeituosa (+; P; positivo) e normal (-; N; negativo)
- Há dois tipos de erro:
  - Falso positivo (FP): o classificador diz que é + quando na verdade não é (é -)
  - Falso negativo (FN): o classificador diz que é – quando na verdade não é (é +)
- Há dois tipos de acerto:
  - Verdadeiro positivo (TP – *true positive*)
  - Verdadeiro negativo (TN – *true negative*)

Matriz de confusão (caso binário)



Medidas de desempenho

- Amostra de tamanho m
  - N objetos negativos
  - P objetos positivos
  - N+P = m
  - N = TN + FP
  - P = TP + FN

Acurácia: (TP+TN)/m
Erro: (FP+FN)/m
= 1-acurácia

Medidas de desempenho

- Amostra de tamanho m
  - N objetos negativos
  - P objetos positivos
  - N+P = m
  - N = TN + FP
  - P = TP + FN

Acurácia: (TP+TN)/m
Erro: (FP+FN)/m
= 1-acurácia

- Taxa de Falsa Aceitação (FAR): FP/m
- Taxa de Falsa Rejeição (FRR): FN/m
- Sensibilidade ou recall ou taxa de TP: TP/P
- Especificidade: TN/N
- taxa de FP = FP/N = 1-especificidade
- Precisão ou valor preditivo positivo (VPP): TP/(TP+FP)
- Valor preditivo negativo (VPN): TN/(TN+FN)

Medidas de desempenho

- Amostra de tamanho m
  - N objetos negativos
  - P objetos positivos
  - N+P = m
  - N = TN + FP
  - P = TP + FN
- Taxa de Falsa Aceitação (FAR): FP/m
- Taxa de Falsa Rejeição (FRR): FN/m
- Sensibilidade ou recall ou taxa de TP: TP/P
- Especificidade: TN/N
- taxa de FP = FP/N = 1-especificidade
- Precisão ou valor preditivo positivo (VPP): TP/(TP+FP)
- Valor preditivo negativo (VPN): TN/(TN+FN)

Acurácia: (TP+TN)/m
Erro: (FP+FN)/m
= 1-acurácia

É fácil ter sensibilidade = 1 !!!

Medidas de desempenho

- Amostra de tamanho m
  - N objetos negativos
  - P objetos positivos
  - N+P = m
  - N = TN + FP
  - P = TP + FN

Acurácia: (TP+TN)/m
Erro: (FP+FN)/m
= 1-acurácia

É fácil ter sensibilidade = 1 !!!

Basta dizer que tudo é positivo!

- Taxa de Falsa Aceitação (FAR): FP/m
- Taxa de Falsa Rejeição (FRR): FN/m
- Sensibilidade ou recall ou taxa de TP: TP/P
- Especificidade: TN/N
- taxa de FP = FP/N = 1-especificidade
- Precisão ou valor preditivo positivo (VPP): TP/(TP+FP)
- Valor preditivo negativo (VPN): TN/(TN+FN)

Medidas de desempenho

- Amostra de tamanho m
  - N objetos negativos
  - P objetos positivos
  - N+P = m
  - N = TN + FP
  - P = TP + FN
- Taxa de Falsa Aceitação (FAR): FP/m
- Taxa de Falsa Rejeição (FRR): FN/m
- Sensibilidade ou recall ou taxa de TP: TP/P
- Especificidade: TN/N
- taxa de FP = FP/N = 1-especificidade
- Precisão ou valor preditivo positivo (VPP): TP/(TP+FP)
- Valor preditivo negativo (VPN): TN/(TN+FN)

Acurácia: (TP+TN)/m
Erro: (FP+FN)/m
= 1-acurácia

É fácil ter especificidade = 1 !!!

Basta dizer que tudo é negativo!

Medidas de desempenho

- Amostra de tamanho m
  - N objetos negativos
  - P objetos positivos
  - N+P = m
  - N = TN + FP
  - P = TP + FN

Acurácia: (TP+TN)/m
Erro: (FP+FN)/m
= 1-acurácia

- Medida F: $\frac{2 * \text{precisão} * \text{revocação}}{\text{precisão} + \text{revovação}}$

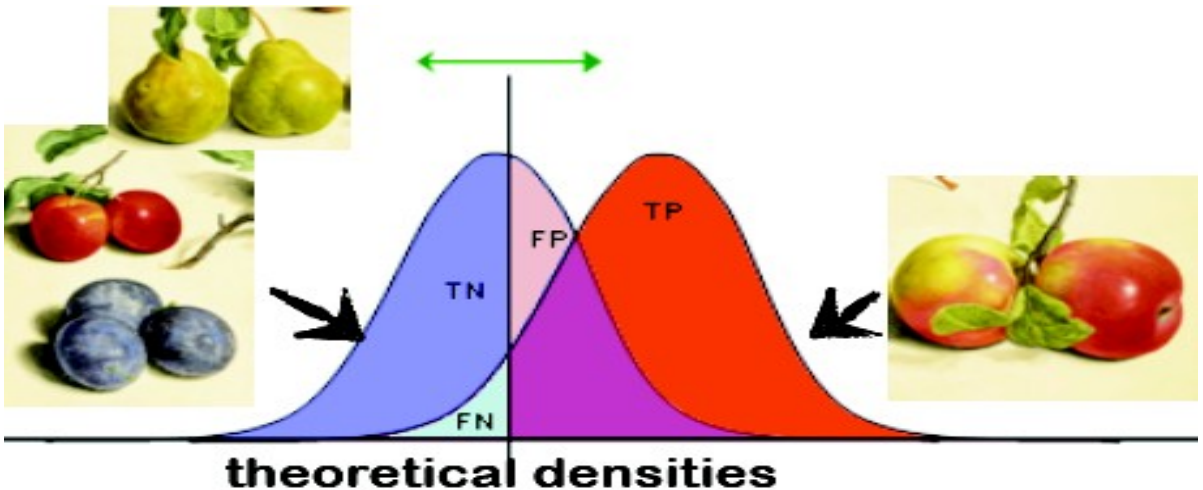
Vídeo 7

Curvas ROC

Curvas ROC

Medidas de desempenho

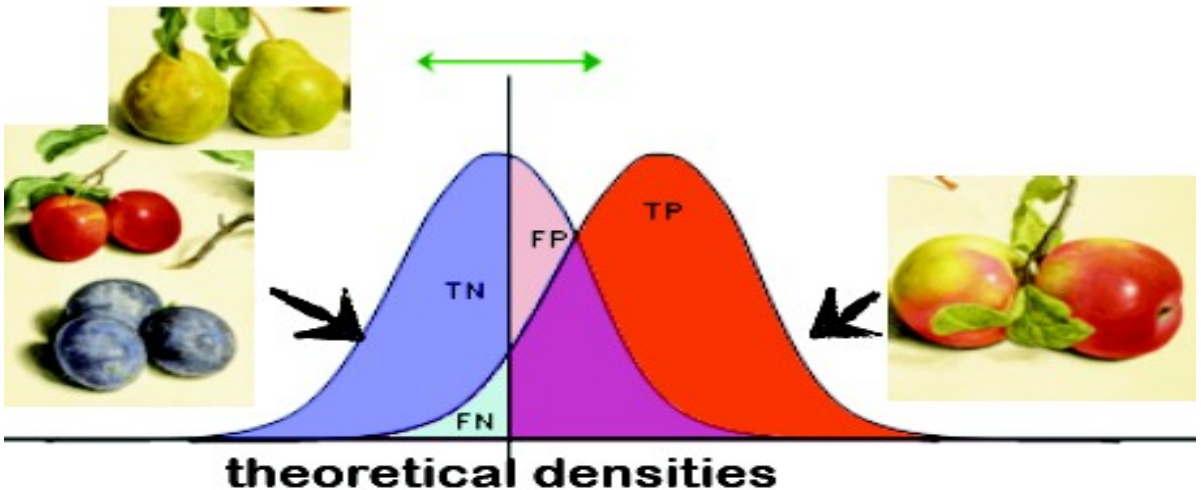
- Dependendo da aplicação, prioriza-se mais a sensibilidade (recall) ou a especificidade (precisão)
- Diferentes limiares de classificação modificam esses valores

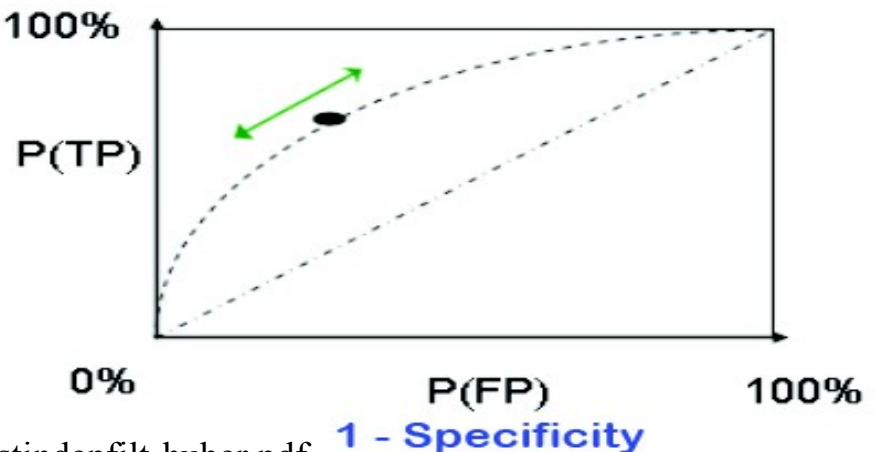


https://bioconductor.org/help/course-materials/2010/EMBL2010/100609-multtestindepfilt-huber.pdf

Medidas de desempenho

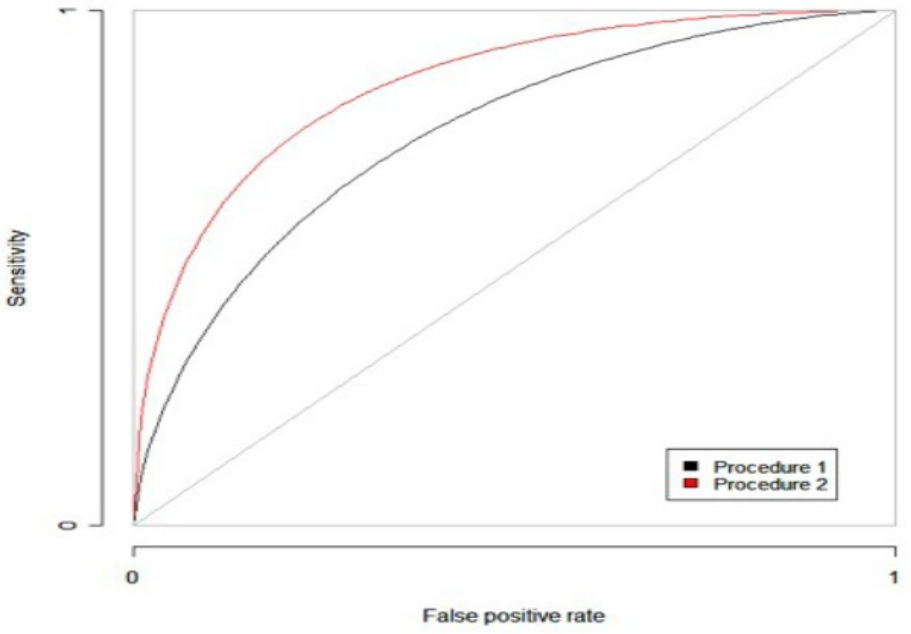
- Dependendo da aplicação, prioriza-se mais a sensibilidade (recall) ou a especificidade (precisão)
- Diferentes limiares de classificação modificam esses valores
- Quando uma sobe a outra desce
- Ex: maçã x outras frutas





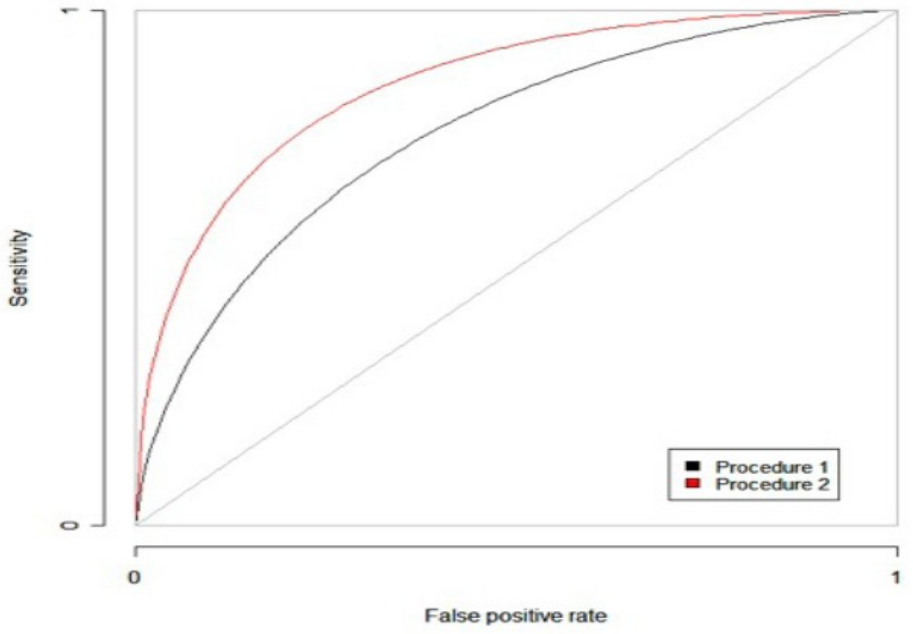
https://bioconductor.org/help/course-materials/2010/EMBL2010/100609-multtestindepfilt-huber.pdf

Curvas ROC – Receiver Operating Characteristic



- Ajuda a escolher um limiar
- Forma de comparar diferentes classificadores

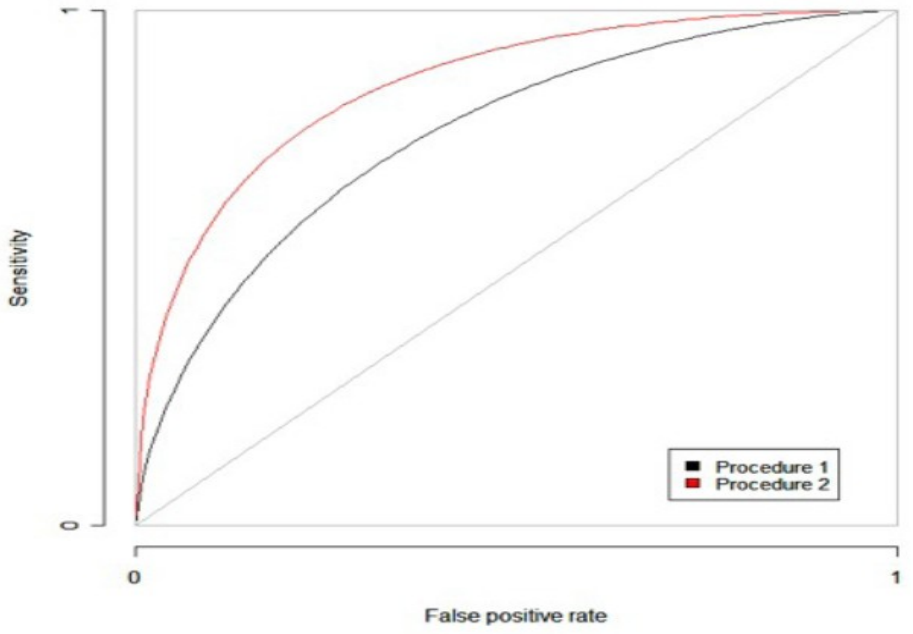
Curvas ROC – Receiver Operating Characteristic



- Qual dos dois algoritmos é melhor? Por quê?

- Ajuda a escolher um limiar
- Forma de comparar diferentes classificadores

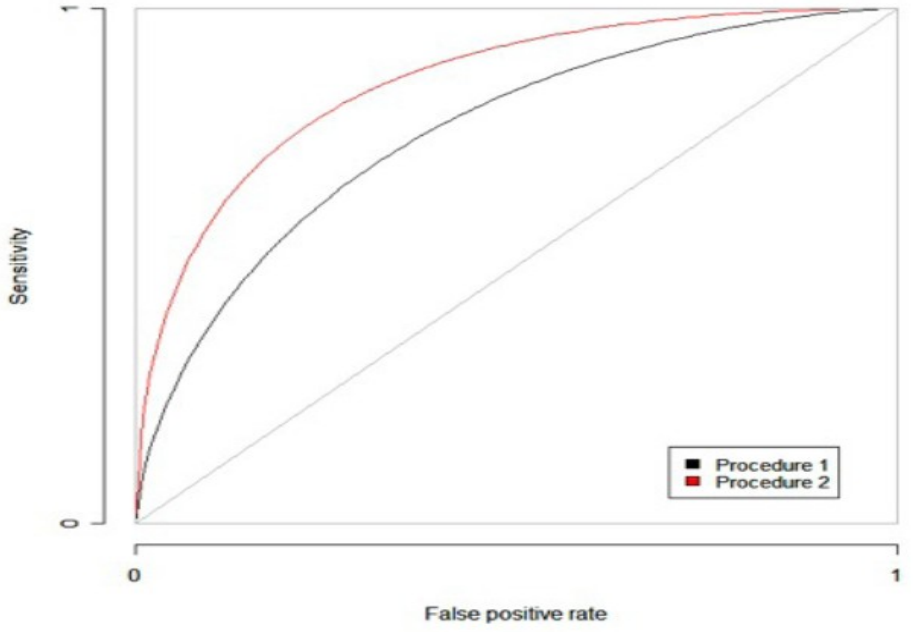
Curvas ROC – Receiver Operating Characteristic



- Qual dos dois algoritmos é melhor? Por quê?
  - 2 (linha vermelha)
  - Porque apresenta melhores taxas de TP e FP

- Ajuda a escolher um limiar
- Forma de comparar diferentes classificadores

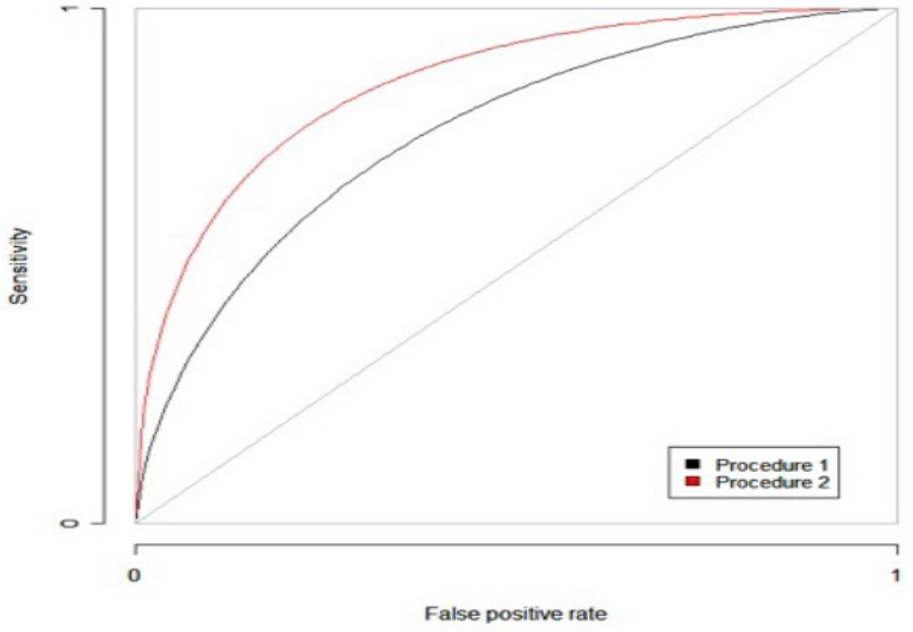
Curvas ROC – Receiver Operating Characteristic



- Qual dos dois algoritmos é melhor? Por quê?
  - 2 (linha vermelha)
  - Porque apresenta melhores taxas de TP e FP
- Como seria a curva para um classificador ideal?

- Ajuda a escolher um limiar
- Forma de comparar diferentes classificadores

Curvas ROC – Receiver Operating Characteristic



- Qual dos dois algoritmos é melhor? Por quê?
  - 2 (linha vermelha)
  - Porque apresenta melhores taxas de TP e FP
- Como seria a curva para um classificador ideal?
  - Ponto (0,1)

- Ajuda a escolher um limiar
- Forma de comparar diferentes classificadores

Como construir uma curva ROC

- Classificadores que só fornecem a classe: representam um único ponto
- Classificadores que fornecem uma probabilidade ou um score:
  - Teoricamente: basta variar o limiar de -∞ a +∞
  - Na prática: variar o limiar para cada probabilidade/score apresentado pelas instâncias de teste

Como construir uma curva ROC

Ex: 20 instâncias de teste (10 positivas e 10 negativas)

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

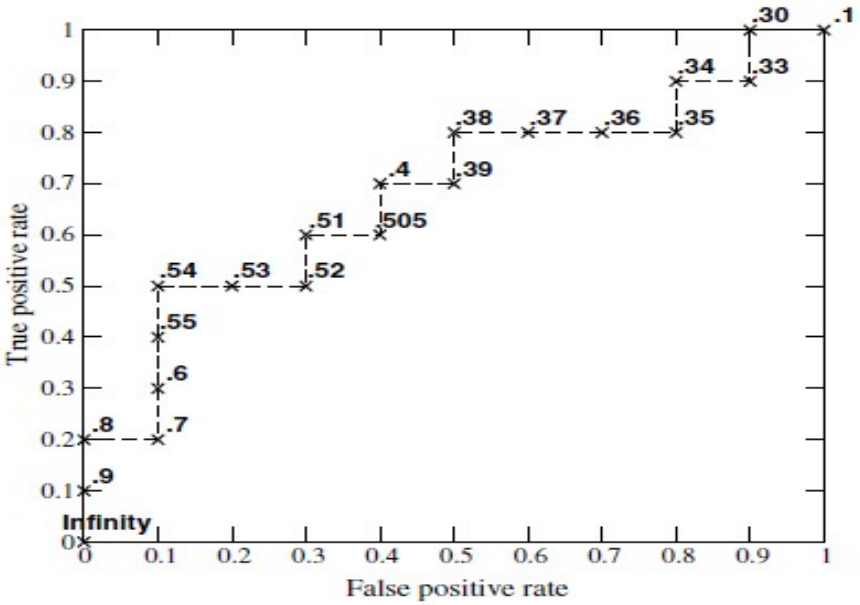
Apenas para fins didáticos, as intâncias positivas e negativas estão ordenadas decrescentemente pelo score

Pattern Recognition Letters 27 (2006) 861–874

Como construir uma curva ROC

Ex: 20 instâncias de teste (10 positivas e 10 negativas)

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



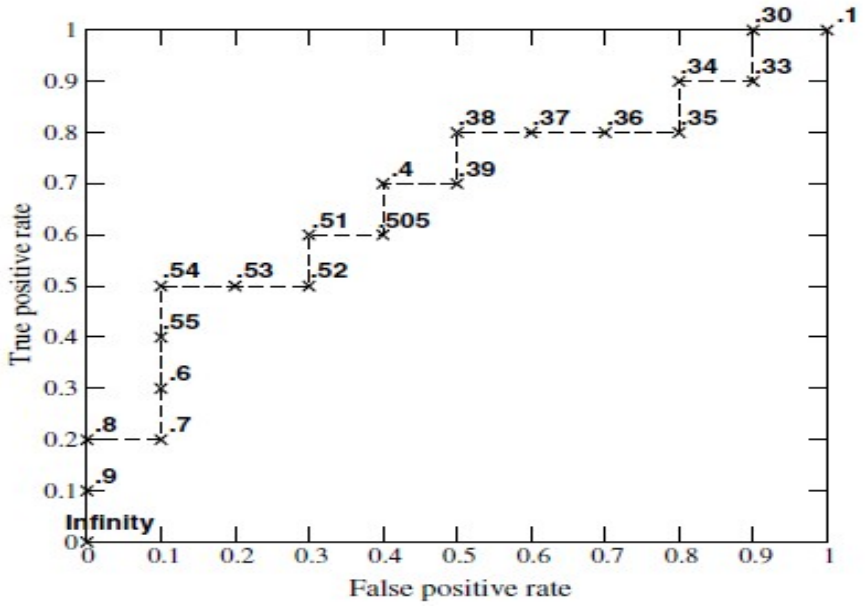
Apenas para fins didáticos, as intâncias positivas e negativas estão ordenadas decrescentemente pelo score

Cada valor distinto de score (acrescido do ponto (0,0)) corresponde a um possível limiar que resultará em um ponto da função degrau que define a curva ROC

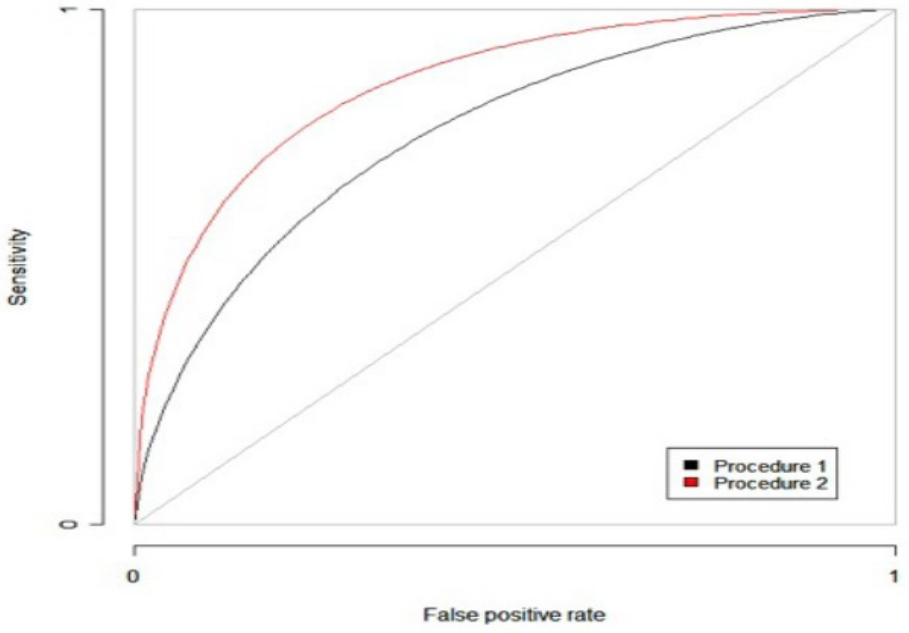
Pattern Recognition Letters 27 (2006) 861–874

Como construir uma curva ROC

A função degrau tenderá a uma curva de verdade à medida que o número de instâncias tender a infinito

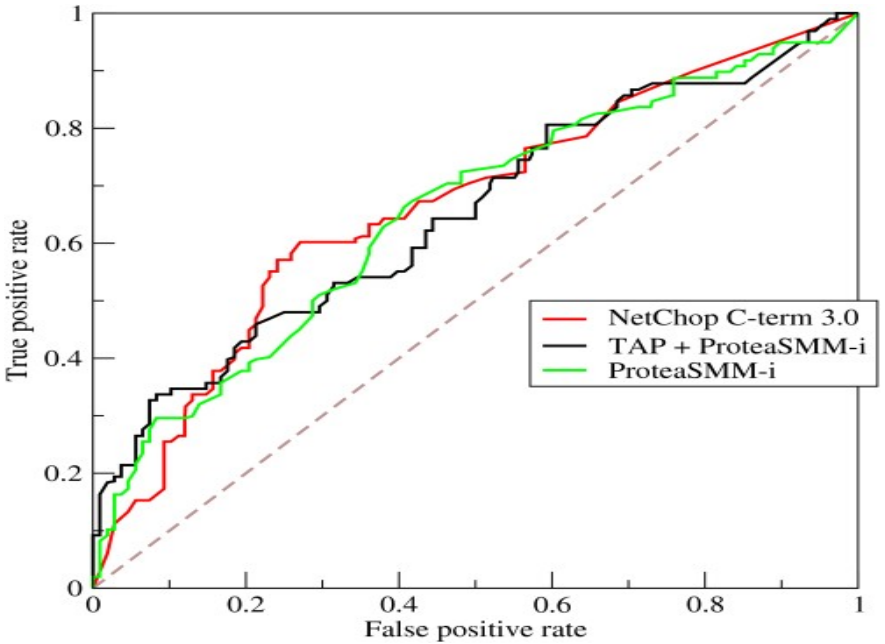


Pattern Recognition Letters 27 (2006) 861–874



Curvas ROC

Comparar visualmente dois ou mais classificadores pode não ser fácil

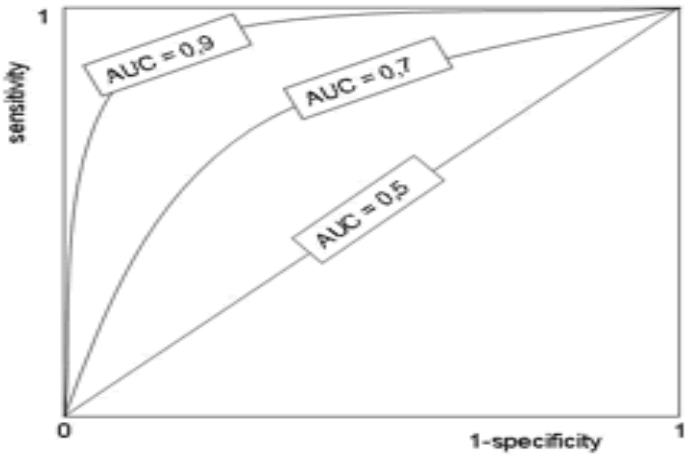


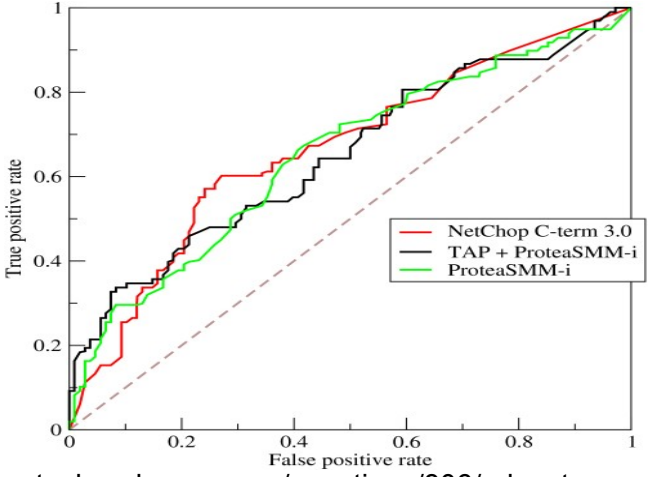
https://datascience.stackexchange.com/questions/806/advantages-of-auc-vs-standard-accuracy

AUC – Area Under the Curve

A área sob a curva ROC (AUC) é uma boa medida da qualidade do classificador

- quanto mais próximo do ideal (ponto (0,1)) maior a AUC





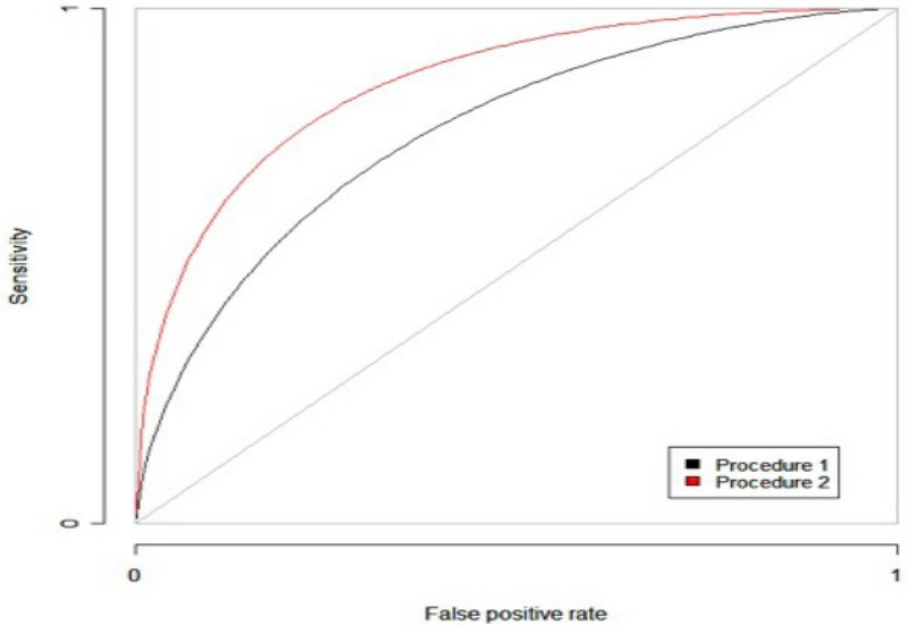
https://datascience.stackexchange.com/questions/806/advantages-of-auc-vs-standard-accuracy

AUC balanceada

E para aqueles classificadores cujo resultado é só a classificação?

- A amostra só a fornece um ponto
- A curva ROC ainda recebe o ponto (0,0)
- AUC_b = (sensibilidade + especificidade) / 2

Curvas ROC – Receiver Operating Characteristic

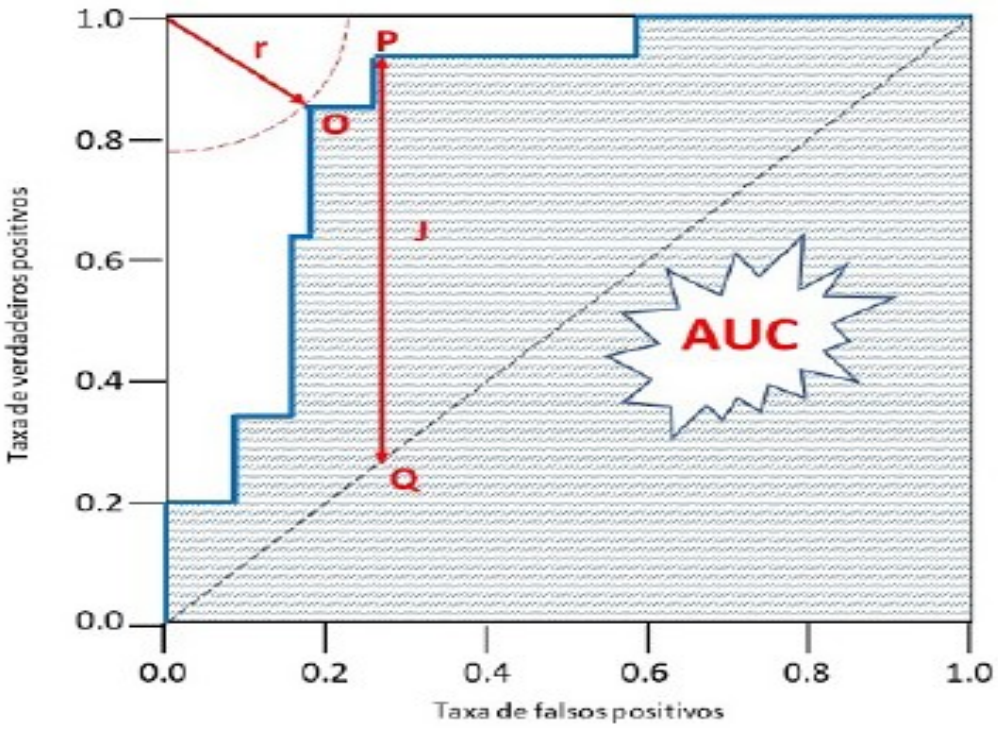


- A AUC te ajuda a avaliar um classificador para os vários limiares, mas na hora de usá-lo, qual limiar escolher?
- Você pode definir um balanço específico entre sensibilidade e especificidade
- Ou utilizar um critério específico

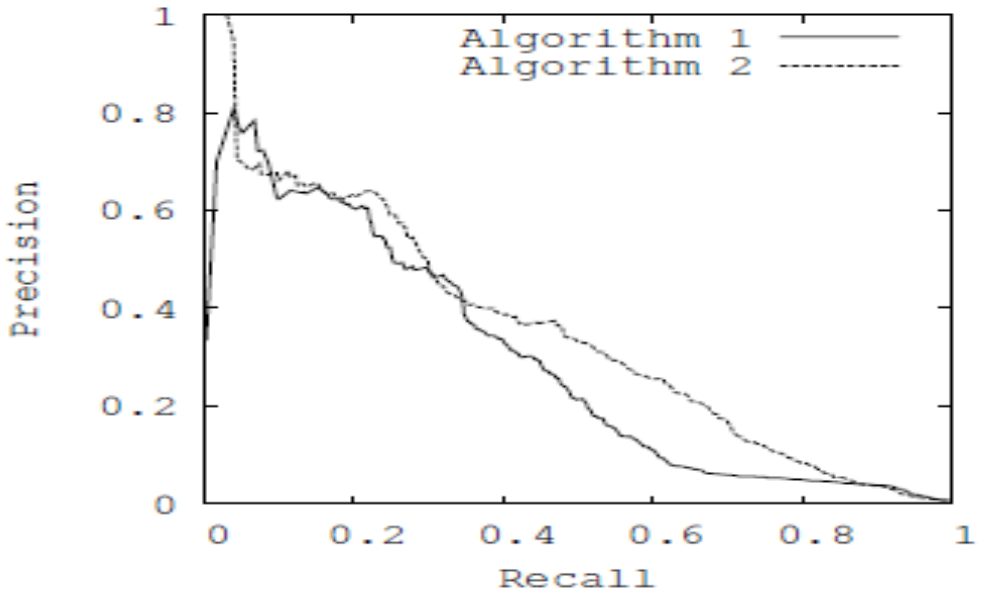
- Ajuda a escolher um limiar
- Forma de comparar diferentes classificadores

Curvas ROC – escolha de um limiar (optimal cutpoints)

- Alguns critérios:
  - Youden: escolher o que apresenta maior distância (no eixo y) da diagonal (que representa uma classificação totalmente ao acaso (na figura: ponto P)
  - 01: escolher o mais próximo do ponto (0,1) (na figura: ponto O)



Curvas “ROC-*like*”
Ex: curvas precision-recall

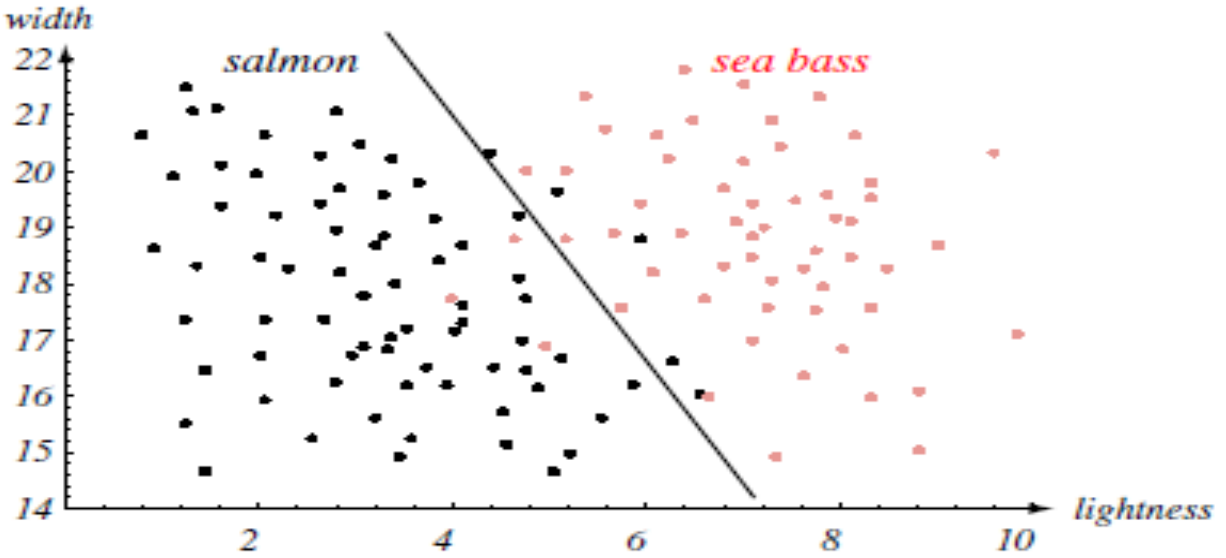


ftp://ftp.cs.wisc.edu/machine-learning/shavlik-group/davis.icml06.pdf

- Outras métricas (além da TPR e FPR) podem ser usadas
- Duas métricas “opostas”: no sentido de que quando uma sobre outra desce
- Ponto ótimo depende das métricas, assim como a melhor AUC

Taxa de rejeição

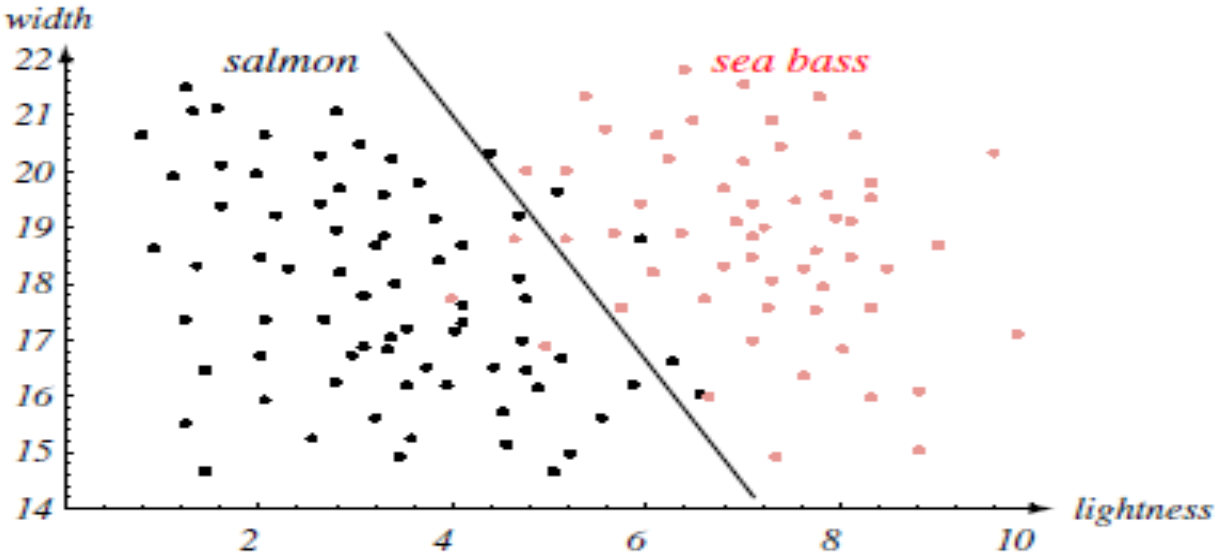
- O que fazer com os dados que caem muito próximos da fronteira de decisão?



[DUDA, HART & STORK, 2001]

Taxa de rejeição

- Uma alternativa é rejeitá-los (recusar-se a classificá-los)



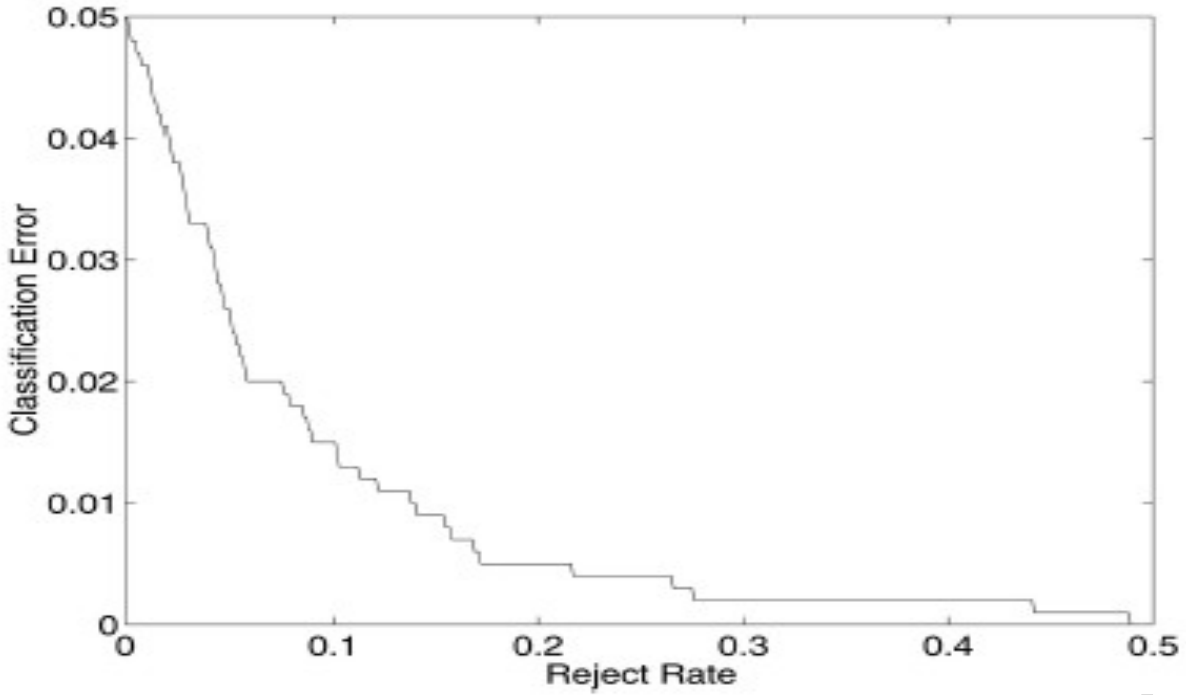
[DUDA, HART & STORK, 2001]

Taxa de rejeição

- Taxa de rejeição: razão do número de rejeitados (para classificação) sobre o total

- Quanto maior a taxa de rejeição, menor a taxa de erro sobre os que sobraram, e vice-versa

Curva taxa de rejeição x erro de classificação



[JAIN et al, 2000]

Referências (vídeos 1 a 3)

DURBIN, R.; EDDY, S. R.; KROGH, A. **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.** Cambridge University Press, 2002. Cap 3 a 6

MACHADO-LIMA, A.; KASHIWABARA, A. Y.; DURHAM, A. M. Decreasing the number of false positives in sequence classification. **BMC Genomics** 11 (Suppl 5):S10, 2010.

RABINER, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. **Proceedings of the IEEE**, v. 77, n. 2, p. 257-286 1989

Referências (Estimação de desempenho)

- FAWCETT, T. An Introduction to ROC Analysis. **Pattern Recognition Letters**, v. 27, p. 861-874, 2006.
- JAIN, A.K.; DUIN, R.P.W.; MAO, J. Statistical Pattern Recognition : A Review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 1, p. 4-37, 2000 (seções 2 e 7)
- Refaeilzadeh P., Tang L., Liu H. (2009) Cross-Validation. In: LIU L., ÖZSU M.T. (eds) **Encyclopedia of Database Systems**. Springer, Boston, MA
- SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond accuracy, f-score and ROC: A family of discriminant measures for performance evaluation. In: **Advances in Artificial Intelligence**, 2006. p. 1015–1021
- E outras citadas nos slides