Tema 05

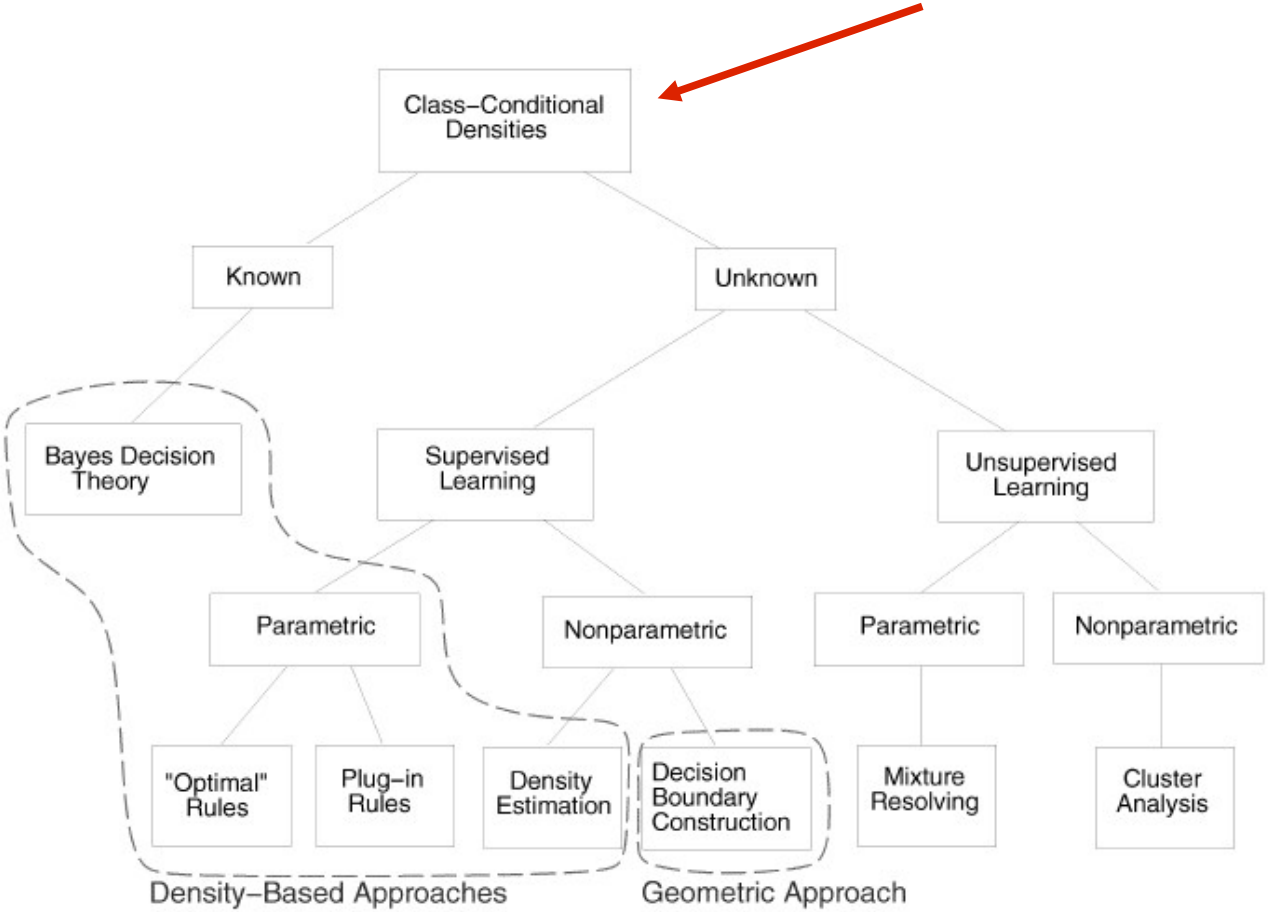
Estimação de Erro

Professora:
Ariane Machado Lima

Classificação

Vimos na aula anterior que, se você conhece as densidades condicionais das classes envolvidas (P(x|c)), o classificador Bayesiano deve ser usado pois ele é o que fornece MENOR erro (classificador ÓTIMO).

Técnicas de Classificação



[JAIN et al, 2000]

Problema

- Normalmente não conhecemos as probabilidades condicionais P(. |ci)
- Neste caso temos que estimá-las a partir de dados conhecidos, o que pode ser complexo

Alternativas

- Nestes casos, temos que optar por um classificador mais simples
- Por ex: escolher uma família de densidades conhecida, ou seja, com uma forma matemática conhecida e um número finito de parâmetros. Daí basta estimar os parâmetros (ex. Normal, Poisson, ...)
  -
-

Alternativas

- Nestes casos, temos que optar por um classificador mais simples
- Por ex: escolher uma família de densidades conhecida, ou seja, com uma forma matemática conhecida e um número finito de parâmetros. Daí basta estimar os parâmetros (ex. Normal, Poisson, ...)
  - CLASSIFICADOR PARAMÉTRICO
-

Alternativas

- Nestes casos, temos que optar por um classificador mais simples
- Por ex: escolher uma família de densidades conhecida, ou seja, com uma forma matemática conhecida e um número finito de parâmetros. Daí basta estimar os parâmetros (ex. Normal, Poisson, ...)
  - CLASSIFICADOR PARAMÉTRICO
- Quando não é possível escolher uma forma matemática então opta-se por um CLASSIFICADOR NÃO PARAMÉTRICO

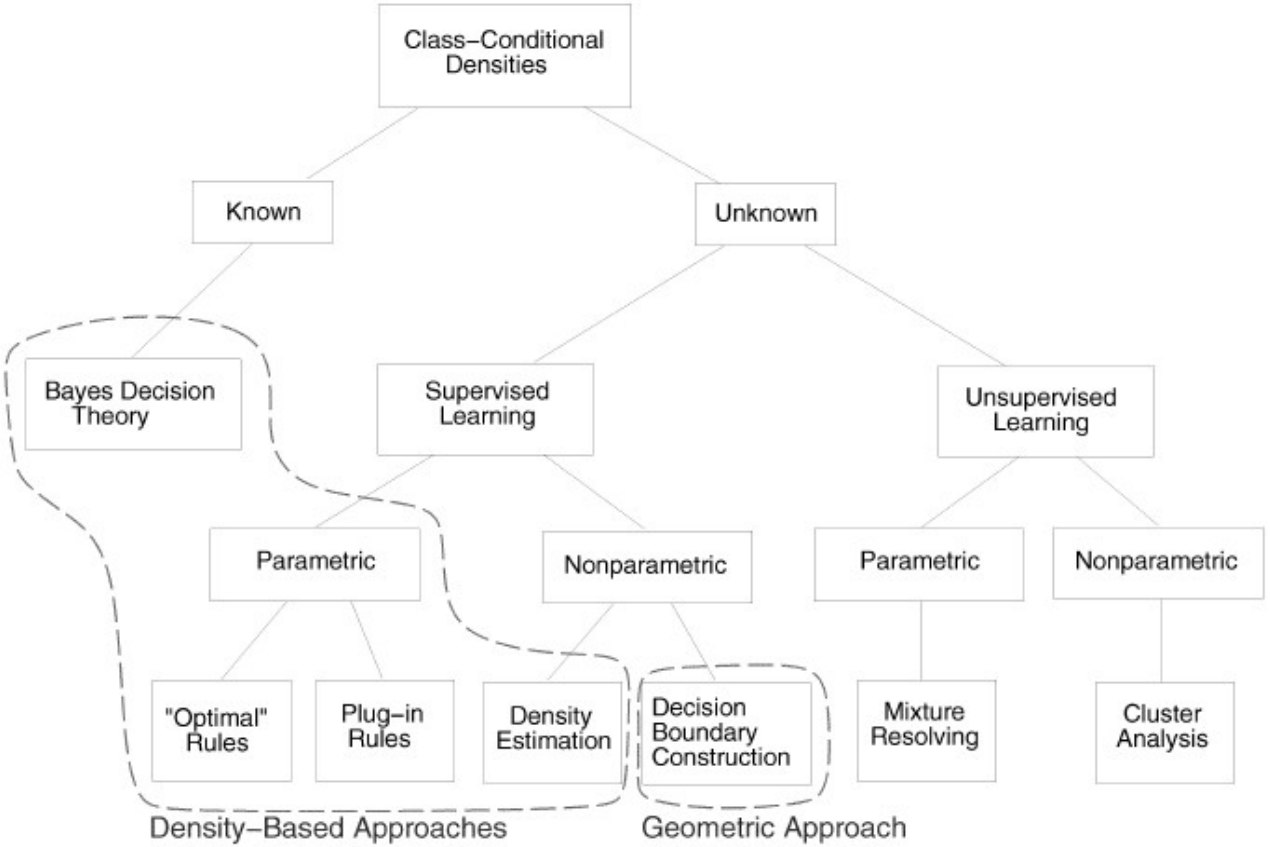
Métodos de construção de um classificador

- Aprendizado supervisionado: o modelo é aprendido a partir de amostras rotuladas (com sua classificação)
  - Amostra de treinamento:

  X = {(x,c)| x é um exemplo e c é sua classe}

- Aprendizado não supervisionado: a classificação é feita a partir de dados não rotulados

Métodos de Classificação



[JAIN et al, 2000]

Como avaliar um classificador?

Como avaliar um classificador?

- Taxa de erro
- Difícil obter uma expressão analítica
- Para regras de treinamento consistentes, a taxa de erro se aproxima do erro de Bayes à medida que cresce o tamanho da amostra de treinamento
- Na prática, como estimar o erro?
- Se eu tenho uma determinada amostra rotulada, tenho que treinar e estimar o erro. Como faço?

Estimação de erro de um classificador

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

Métodos de estimação de erro de um classificador

- Resubstituição
- Holdout
- Leave-one-out
- Rotação ou k-fold cross-validation (validação cruzada k-vezes)
- 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%)

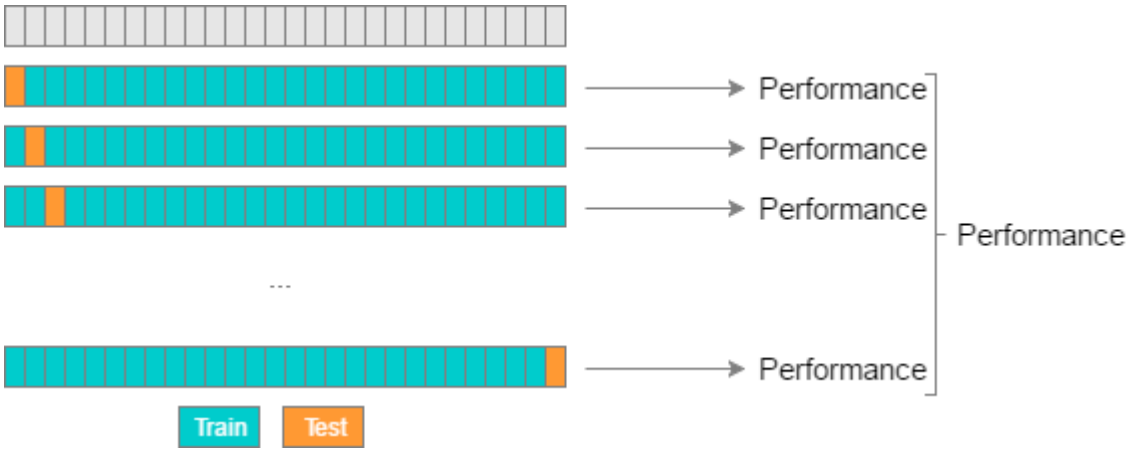
- Problema:

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

Leave-one-out

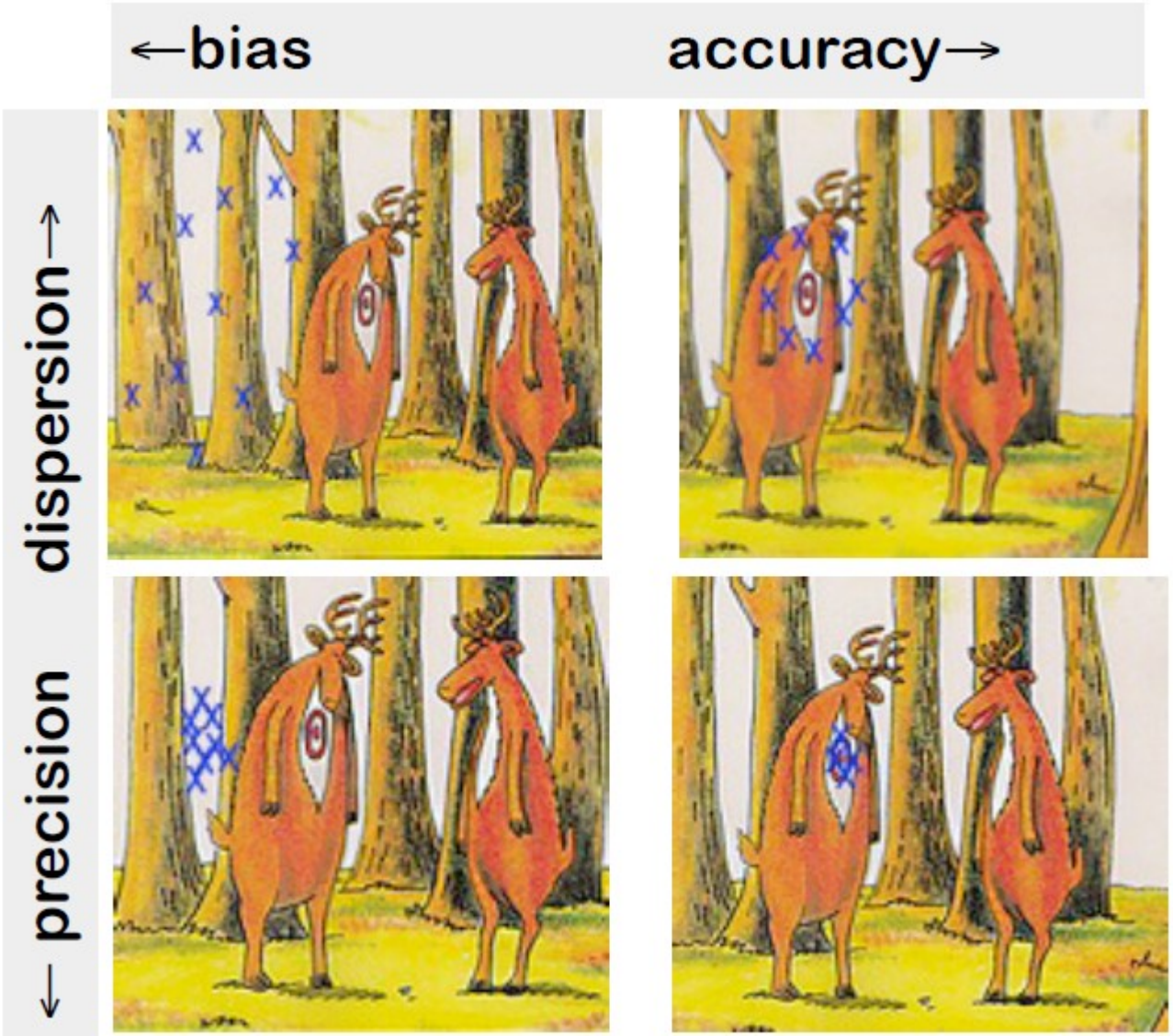
- Se a amostra original tem n dados, treina com n-1 dados e testa no que restou
- Repetir o processo n vezes, cada vez deixando um dado de fora para teste
- A estimativa de erro é a média do erro de cada rodada



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)

Viés (bias) x variância



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

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)

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

Bootstrap

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

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

Medidas de acurácia

Matriz de confusão

Mij: quanto elementos da classe j foram preditos como sendo da classe i

		Classe real				
		classe 1	classe 2	...	classe n	Totais
Predição	classe 1					
	classe 2					
	...					
	classe n					
	Totais					

Matriz de confusão

Mij: quanto elementos da classe j foram preditos como sendo da classe i

		Classe real				
		classe 1	classe 2	...	classe n	Totais
Predição	classe 1					
	classe 2					
	...					
	classe n					
	Totais					

Diagonal tem os acertos, e o resto os erros

Medidas de acurácia

- 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 (-)
  - Falso negativo (FN): o classificador diz que não é (-) quando na verdade é (+)
- Há dois tipos de acerto:
  - Verdadeiro positivo (TP – *true positive*)
  - Verdadeiro negativo (TN – *true negative*)

Matriz de confusão (caso binário)

		Classe real		
		classe +	classe -	Totais
Predição	classe +			
	classe -			
	Totais			

Matriz de confusão (caso binário)

		Classe real		
		classe +	classe -	Totais
Predição	classe +	TP	FP	
	classe -	FN	TN	
	Totais			

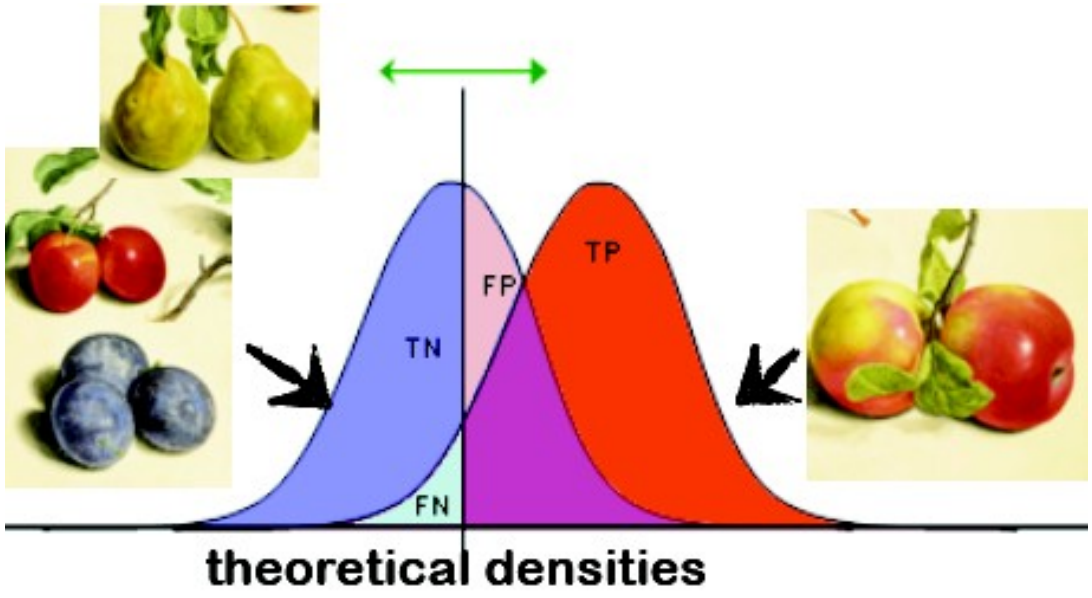
Medidas de acurácia

- 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 TP rate: TP/P
- Especificidade: TN/N
- FP rate = FP/N = 1-especificidade
- Precisão: TP/(TP+FP)

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

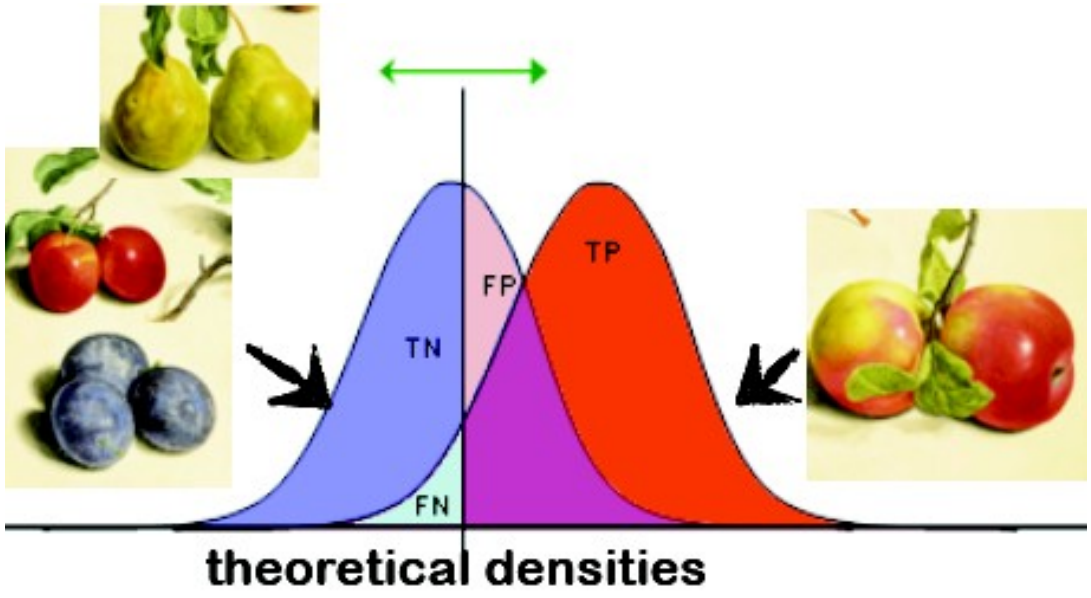
Medidas de acurácia

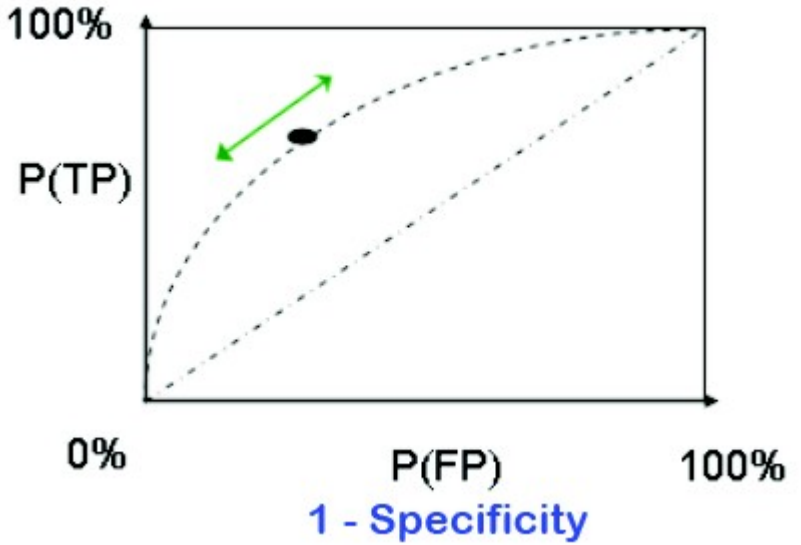
- Dependendo da aplicação, prioriza-se mais a sensibilidade (recall) ou a especificidade (precisão)
- Diferentes limiares de classificação modificam esses valores



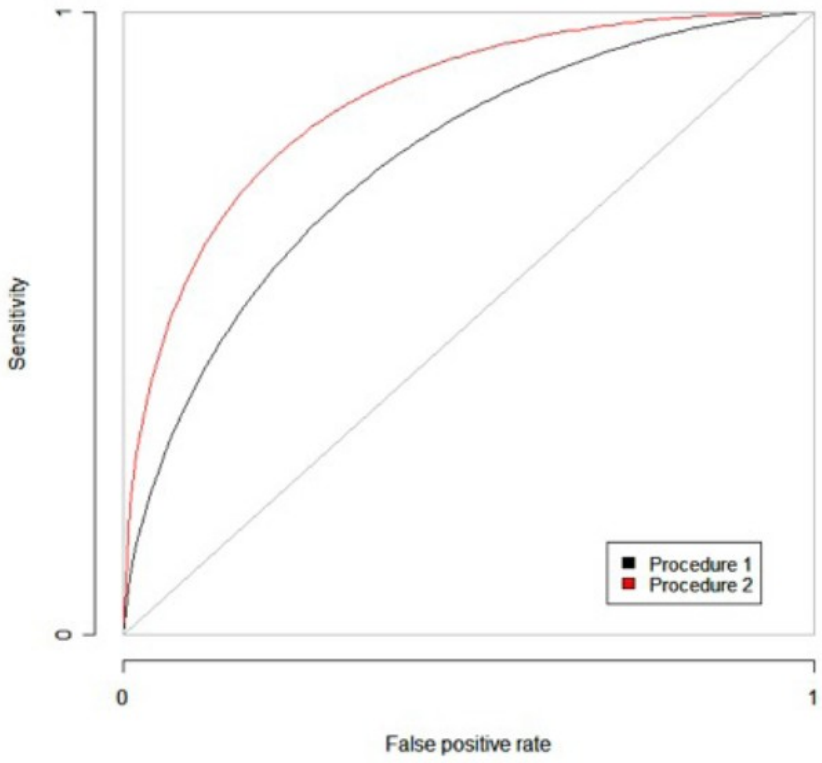
Medidas de acurácia

- 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



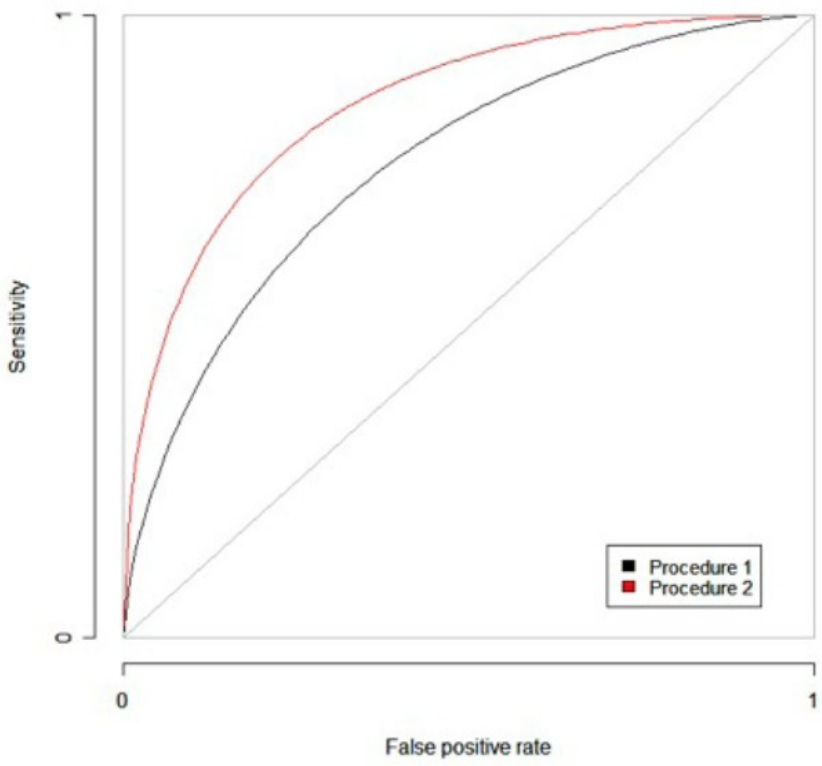


Curvas ROC – Receiver Operating Characteristic



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

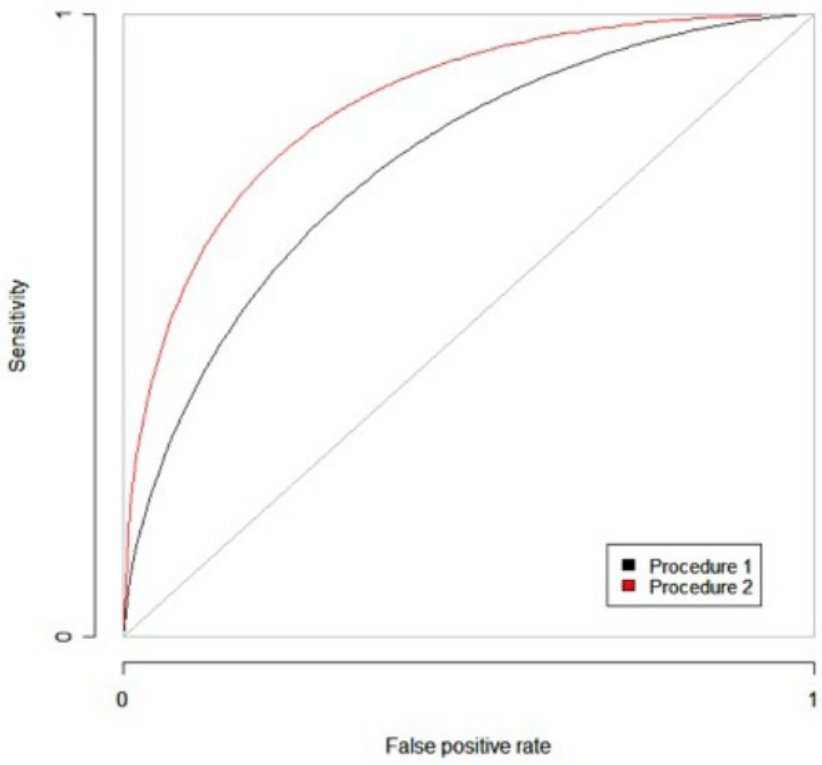
Curvas ROC – Receiver Operating Characteristic



- Qual dos dois algoritmos é melhor? Por quê?

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

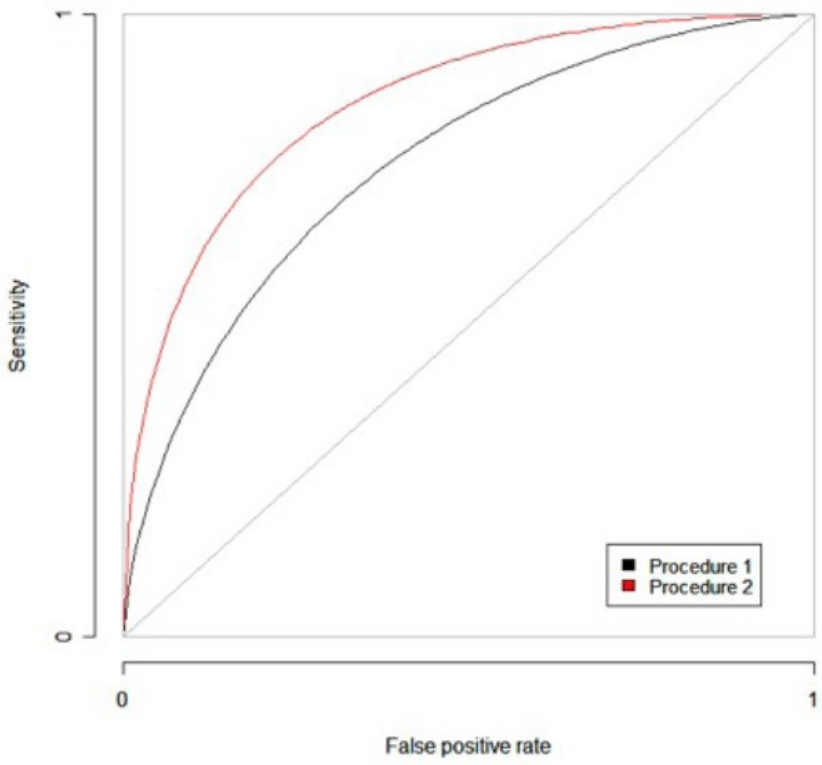
Curvas ROC – Receiver Operating Characteristic



- 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

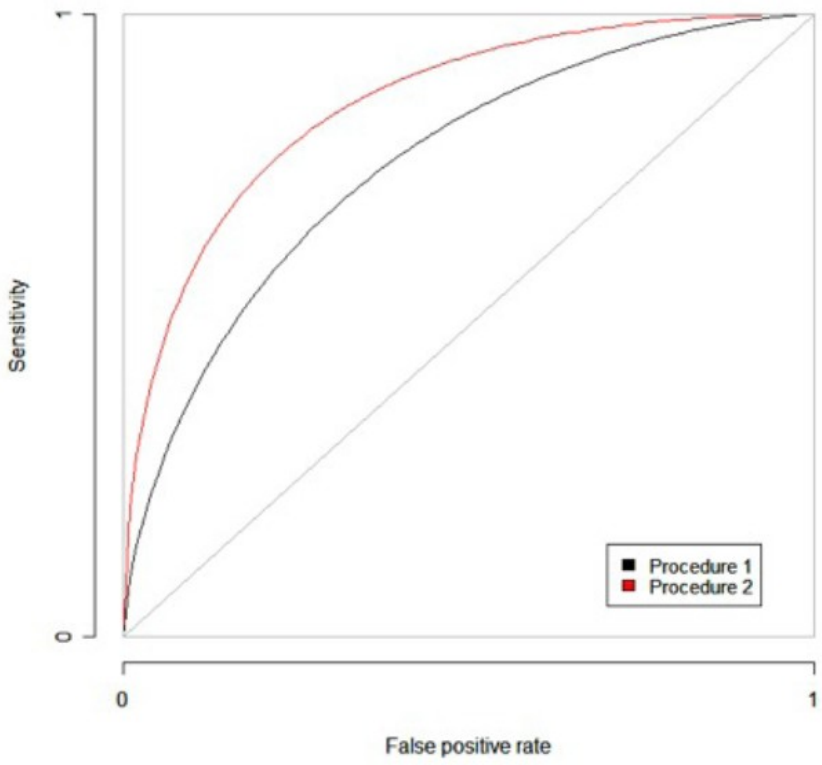
Curvas ROC – Receiver Operating Characteristic



- 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)

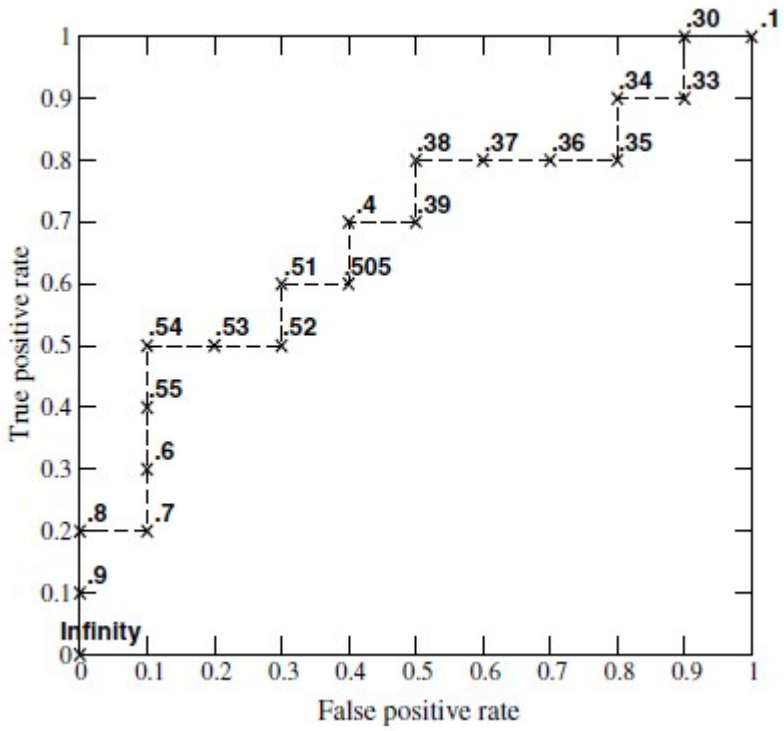
positivas			negativas		
Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

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

Como construir uma curva ROC

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

positivas			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

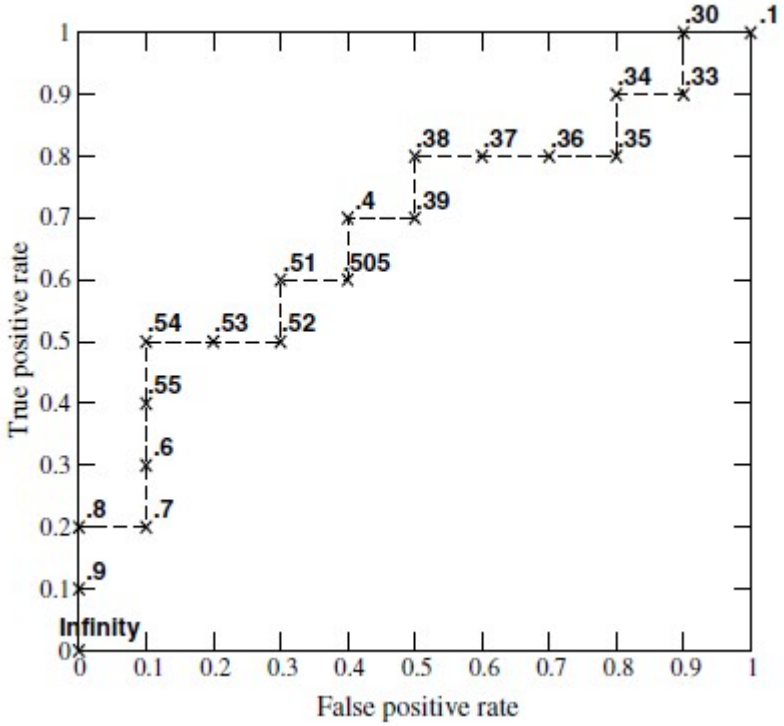


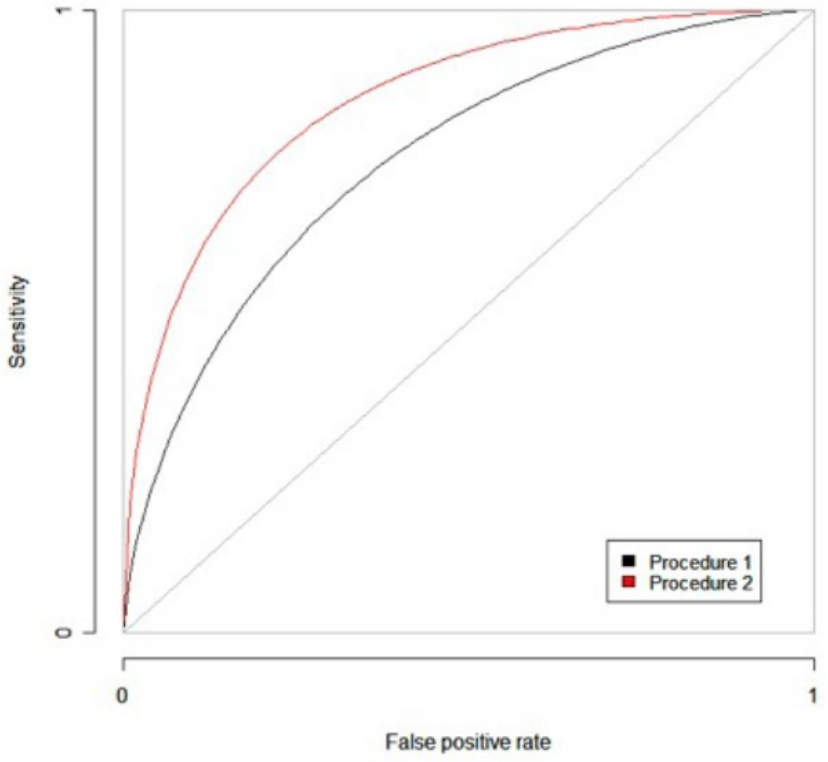
Apenas para fins didáticos, as intâncias positivas e negativas estão ordenadas decrescentemente pelo score

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

Como construir uma curva ROC

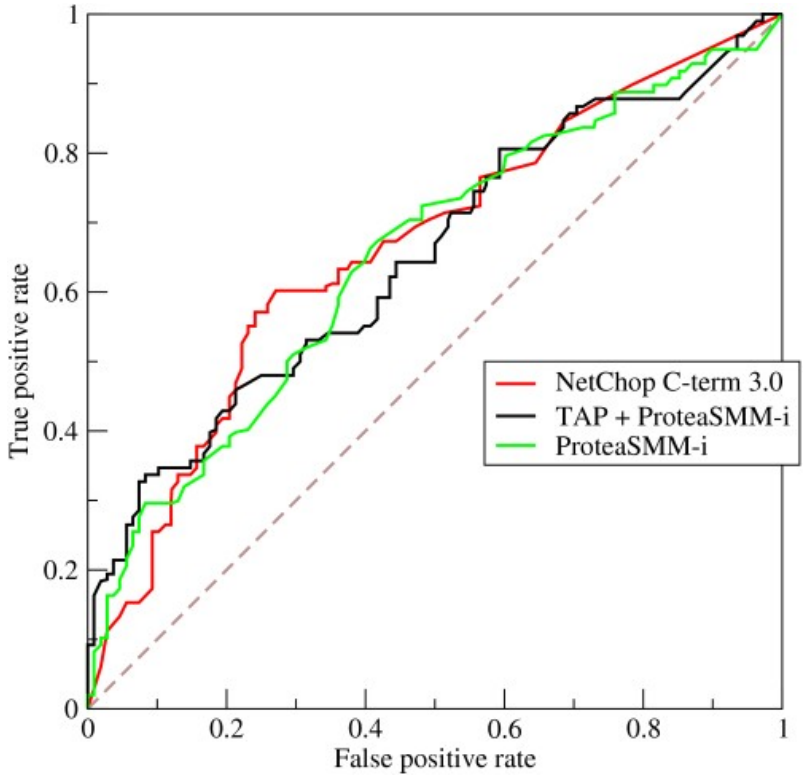
A função degrau tenderá a uma curva de verdade à medida que o número de instâncias tender a infinito





Curvas ROC

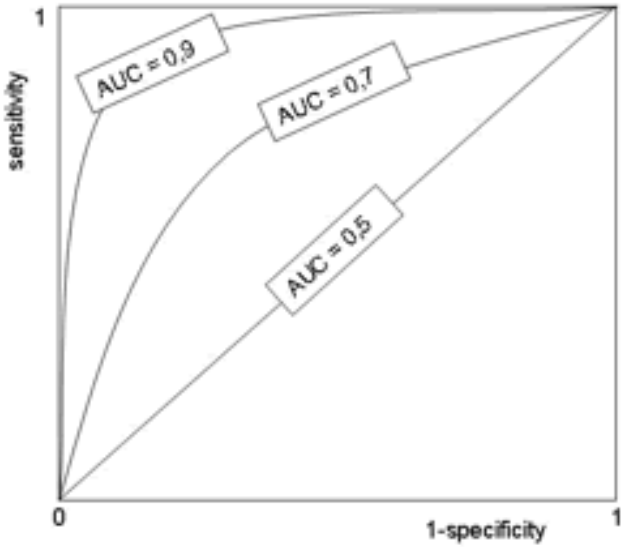
Comparar visualmente dois ou mais classificadores pode não ser fácil

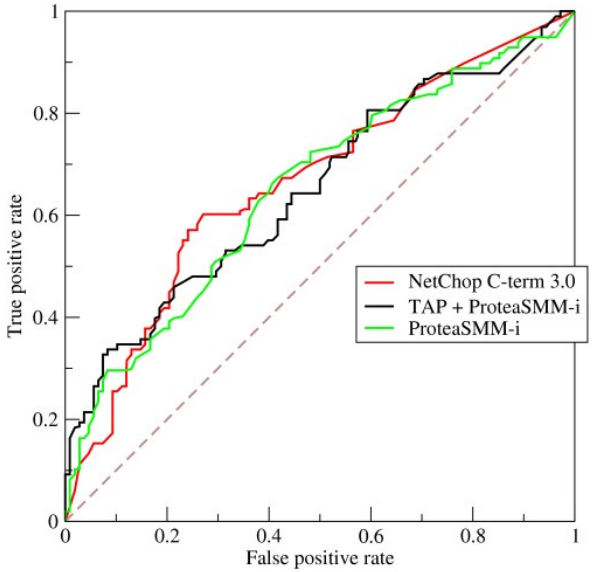


AUC – Area Under the Curve

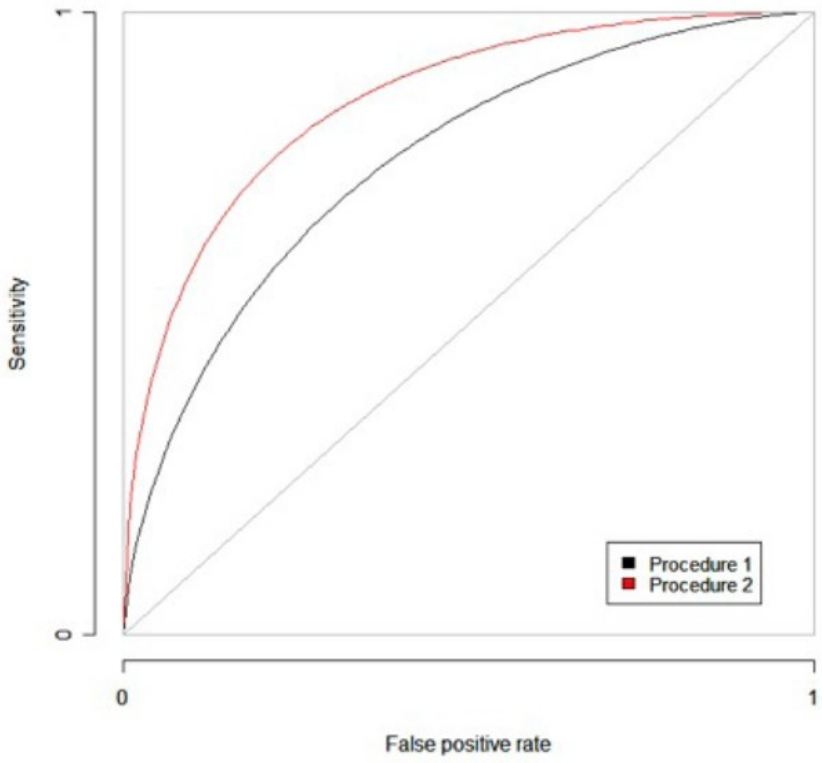
A área sob a curva ROC (AUC) é uma boa medida da qualidade do classificador

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





Curvas ROC – Receiver Operating Characteristic

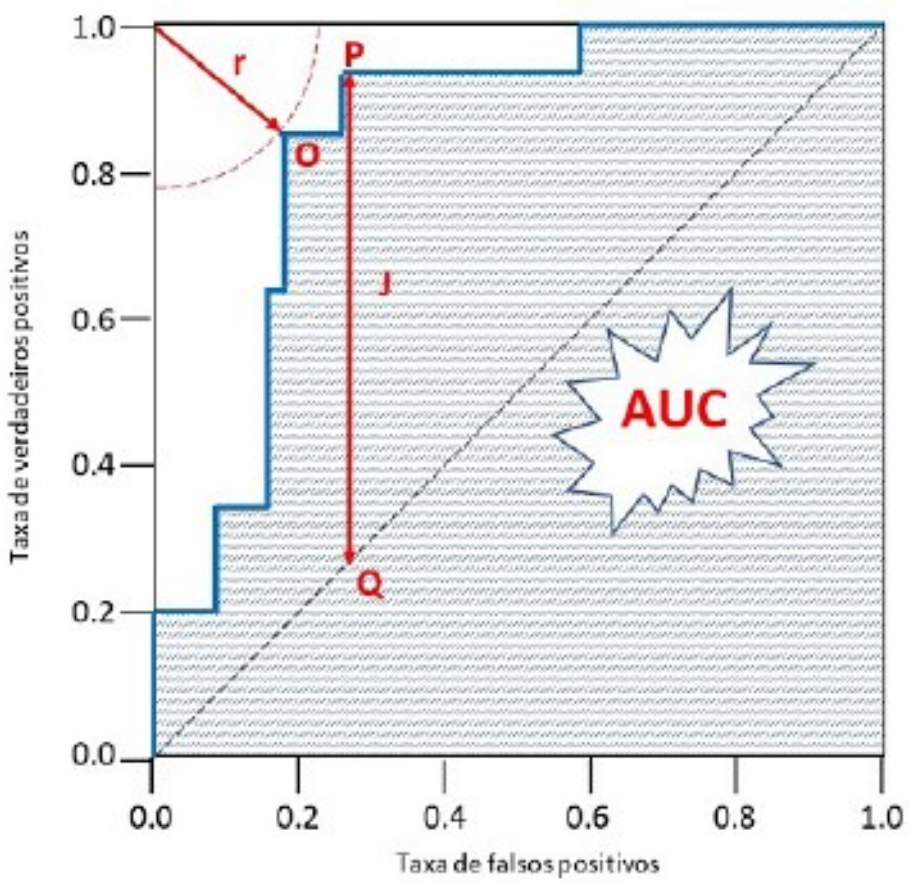


- A AUC te ajuda a avaliar um classificador para os vários limiares, mas na hora de usá-lo, qual limiar escolher?
- Você pode definir um balanço específico entre sensibilidade e especificidade
- Ou utilizar um critério específico

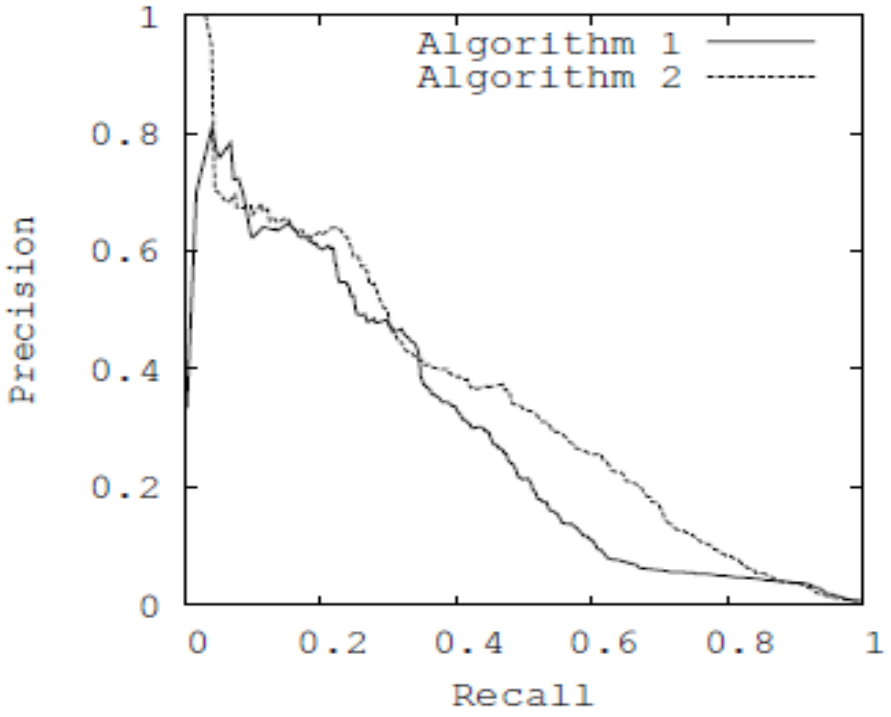
- Ajuda a escolher um limiar
- Forma de comparar diferentes classificadores

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

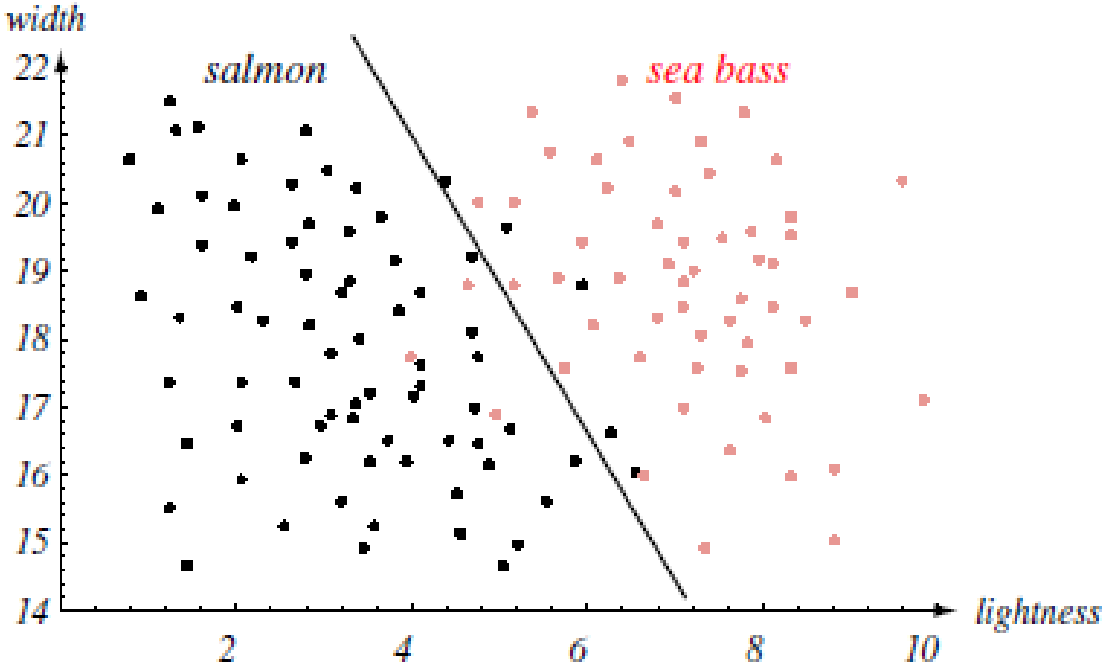


ftp://ftp.cs.wisc.edu/machine-learning/shavlik-group/davis.icml06.pdf

- 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

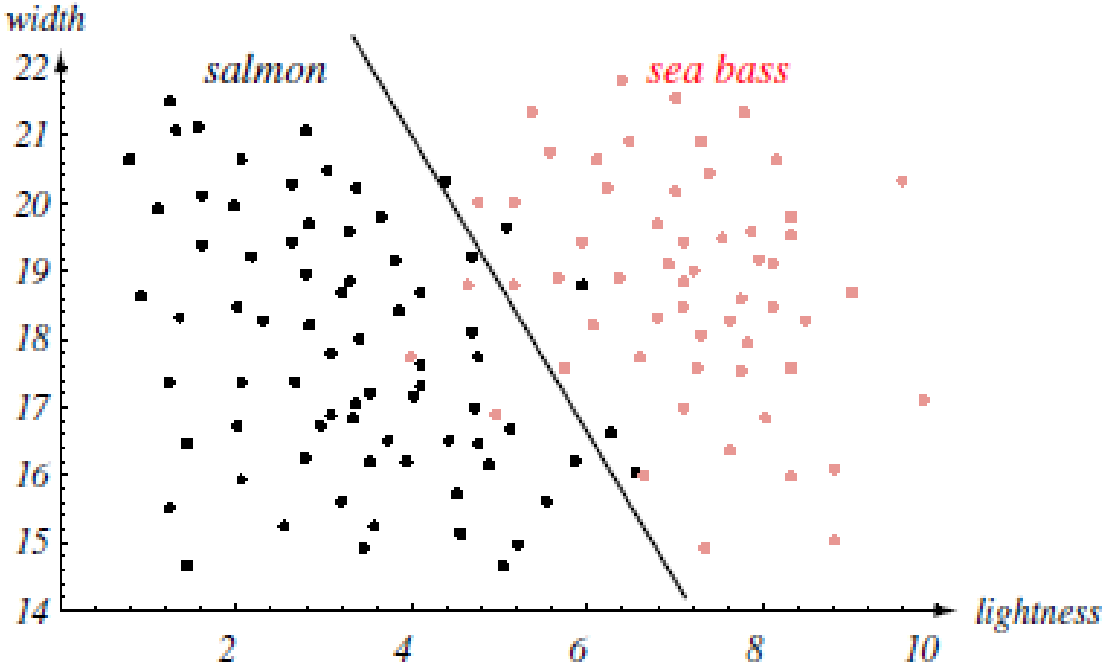
- O que fazer com os dados que caem muito próximos da fronteira de decisão?



[DUDA, HART & STORK, 2001]

Taxa de rejeição

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

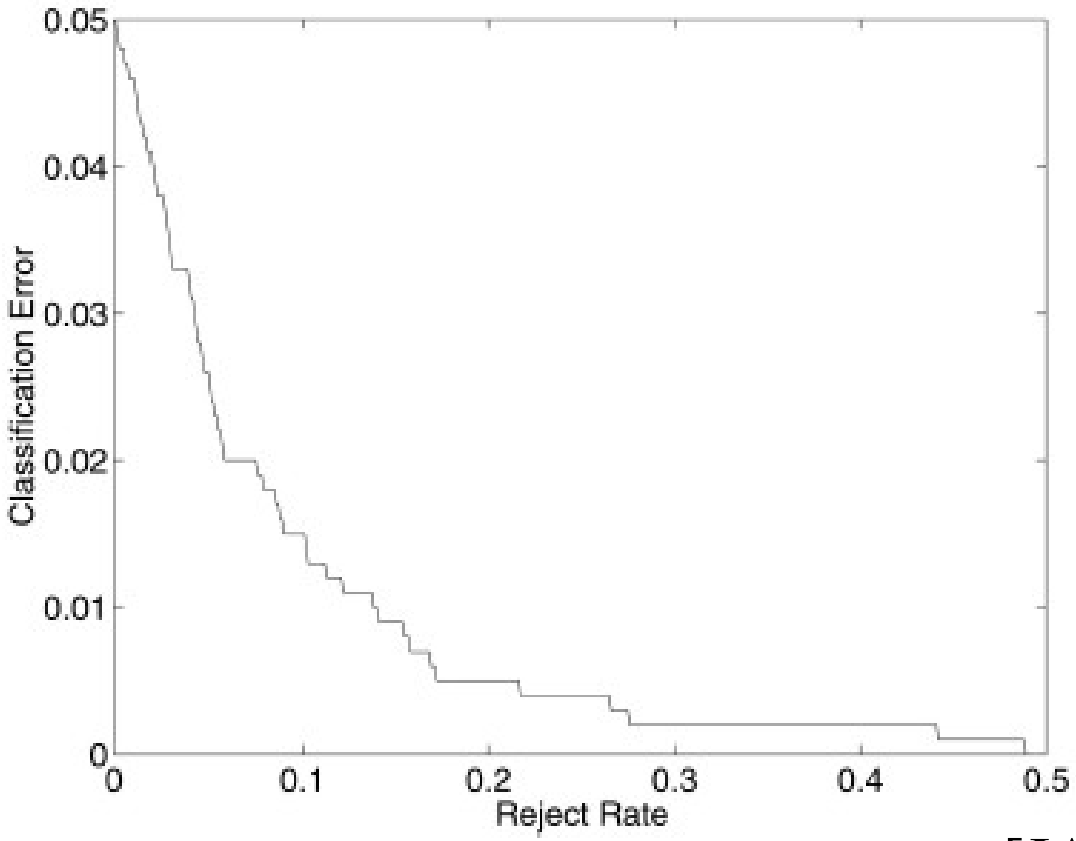


[DUDA, HART & STORK, 2001]

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]

Atividade 4

- Implementar uma função que FARÁ uma estimativa de erro de um classificador baseada em k-fold cross-validation
- Por enquanto, a função deve receber como parâmetros:
  - um conjunto de dados (amostra original)
  - o valor de k

  e deve dividir a amostra original em k subconjuntos de amostras (treinamento e teste)
- Futuramente, essa função vai estimar o erro de um classificador

Atividade 4

- Observações:
  - Cada parte deve ter tamanho diferente de outra em no máximo 1 elemento
  - Cada parte deve manter a proporção de cada classe
  - No final imprimir quantos elementos ficaram em cada parte (quantos de cada classe e quantos no total)

Referências

- DUDA, R.; HART, P.; STORK, D**. Pattern Classification** . John Willey, 2001 (Cap. 2.1 a 2.3)

- 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)