Estratégias para Avaliação de Classificadores

Clodoaldo A. M. Lima, Sarajane M. Peres

30 de maio de 2017

Material baseado em:
HAN, J. & KAMBER, M. Data Mining: Concepts and Techniques. 2nd. 2006
FAWCETT, T. An introduction to ROC analysis. Patt. Recog. Letters, 27, 2006, 861–874
PRATI, R. C.; Batista, G. E. A. P. A.; Monard, M. C. Curvas ROC para avaliação de classificadores. IEEE Latin America Transactions, v. 6, n.2, June, 2008.

Acurácia X Erro

Usar o conjunto de dados de treinamento (o qual derivou um modelo de classificação) para estimar a acurácia do modelo derivado pode produzir medidas (super)otimistas devido à (super)especialização do modelo. Então, é melhor usar um conjunto de teste, composto por dados que não foram usados no treinamento.

A acurácia de um classificador em um dado conjunto de teste é a porcentagem de tuplas do conjunto de teste que são corretamente classificadas pelo classificador. Essa medida pode também ser chamada de taxa de reconhecimento do classificador.

A taxa de erro, ou taxa de classificações erradas, de um classificador M é $1 - Acc(M)$ em que $Acc(M)$ é a acurácia de M.

Caso o conjunto de treinamento seja usado para estimar a taxa de erro de um modelo, esta medida recebe o nome de erro resubstituição.

Estratégias para Avaliação

Algumas estratégias melhoram a confiabilidade de medição da acurácia de um classificador.

- *Holdout* e amostragem randômica
- *Cross-validation* e *Leave-one-out*
- *Bootstrap*

Estratégias para Avaliação

Holdout

O conjunto de dados disponível para construção do modelo classificador é randomicamente particionado em dois conjuntos - o **conjunto de treinamento** e o **conjunto de teste**. Tipicamente, dois terços dos dados são alocados no conjunto de treinamento, e o restante fica para o conjunto de teste. O conjunto de treinamento é usado para derivar o modelo, cuja acurácia é estimada com o conjunto de teste. A estimativa da acurácia é pessimista porque somente parte do conjunto inicial de dados é usada para derivar o modelo.

Random Sampling

É uma variação do *holdout* na qual o método *holdout* é repetido k vezes. A acurácia total estimada é calculada como a média das acurácias obtidas em cada repetição (do *holdout*).

Estratégias para Avaliação

Cross-validation

No **k-fold cross-validation**, o conjunto de dados é randomicamente particionado em k subconjuntos mutuamente exclusivos e de tamanhos aproximadamente iguais, também chamados de *folds*, D_1, D_2, ..., D_k. Na iteração i, a partição D_i é reservada como conjunto de teste, e as partições restantes são coletivamente usadas para treinar (induzir) o modelo; a segunda iteração é treinada nos subconjuntos D_1, D_3, ..., D_k e testada na partição D_2; e assim por diante. A acurácia é estimada sobre o número de classificações corretas das k iterações, dividido pelo total de tuplas no conjunto de dados inicial.

Leave-one-out

Leave-one-out é um caso especial de *k-fold cross-validation* onde k é o número de tuplas no conjunto de dados.

Stratified cross-validation

No *cross-validation* estratificado, os *folds* são estratificados tal que a distribuição de classes das tuplas em cada *fold* seja aproximadamente a mesma que a distribuição das classes no conjunto de dados inicial.

Estratégias para Avaliação

Bootstrap

O método *boostrap* amostra as tuplas para o conjunto de treinamento usando reposição. Ou seja, cada vez que uma tupla é selecionada, ela não é retirada do conjunto inicial e tem a mesma chance de ser escolhida novamente.

.632 *Bootstrap*

Suponha um conjunto de dados de d tuplas. O conjunto de dados será amostrado d vezes, com reposição, gerando o conjunto de treinamento. É provável que alguns dos dados ocorra mais do que uma vez no conjunto de treinamento. As tuplas que não fazem parte do conjunto de treinamento comporão o conjunto de teste. Repetindo esse procedimento várias vezes, em média, 63.2% dos dados originais cairão no conjunto de treinamento, e 36.8% cairão no conjunto de teste. Aplicando o procedimento k vezes a acurácia será medida com

$$Acc(M) = \sum_{i=1}^{k}(0.632 * Acc(M_i)_{TestSet} + 0.368 * Acc(M_i)_{TrainSet})$$

Medidas de Avaliação

Matriz de confusão (ou matriz de contingência): tabulação cruzada entre as classes preditas pelo modelo e a classe real de cada exemplo.

Considerando contagem, ou frequência absoluta, tem-se:

	predito		
real	TP	FN	POS
	FP	TN	NEG
	PP	PN	N

- TP: true positive (verdadeiro positivo)
- FP: false positive (falso positivo)
- FN: false negative (falso negativo)
- TN: true negative (verdadeiro negativo)
- PP: predição positiva
- PN: predição negativa
- POS: positivos reais
- NEG: negativos reais
- N: número de elementos na amostra

Medidas de Avaliação

Dessa matriz, várias medidas de valor único podem ser extraídas. Medidas de valor único são úteis, mas perdem informação e podem levar a avaliações erradas. Cada medida derivada da matriz de confusão deve ser cuidadosamene interpretada de acordo com o domínio do problema e com as características das distribuições de classes nos conjuntos (de treinamento e/ou de teste).

Se dividirmos cada entrada da matriz pelo tamanho da amostra, cada entrada da matriz será a probabilidade conjunto da classe real do exemplo e da predição dada pelo exemplo. Em termos de probabilidades, a matriz é:

	Y	$\overline{Y}$	
X	$p(X,Y)$	$p(X,\overline{Y})$	$p(X)$
$\overline{X}$	$p(\overline{X},Y)$	$p(\overline{X},\overline{Y})$	$p(\overline{X})$
	$p(Y)$	$p(\overline{Y})$	1

em que X é a variável aleatória classe real do exemplo positivo; e Y é a variável aleatória classe predita do exemplo positivo

Medidas de Avaliação

Problema de classificação binária

Para avaliar um classificador binário, algumas medidas devem ser calculadas sobre os dados da matriz de confusão:

- sensibilidade ou taxa de verdadeiros positivos ou revocação: porcentagem de verdadeiros positivos dentre todos os exemplos cuja classe real é positiva $TPR = TP/(TP + FN)$
- taxa de falsos positivos: porcentagem de exemplos cuja classe real é negativa que são classificados como positivos $FPR = FP/(TN + FP)$
- especificidade ou taxa de verdadeiros negativos: proporção de verdadeiros negativos (rejeições corretas) entre os exemplos cuja classe real é negativa $SPC = TN/(FP + TN)$
- precisão (precision ou preditividade positiva): proporção de acertos dentre todos os exemplos preditos como positivos: $PPV = TP/(TP + FP)$
- preditividade negativa: proporção de rejeições corretas dentre os exemplos preditos como negativos: $NPV = TN/(TN + FN)$
- taxa de falsas descobertas: denota o número de falsos positivos dentre os exemplos classificados como positivos: $FDR = FP/(TP + FP)$
- acurácia: quantidade (ou taxa) de exemplos classificados corretamente;
- erro: quantidade (ou taxa) de exemplos classificados incorretamente.

Medidas de Avaliação

Confiança × Crença (ou Verosimilhança)

- A precisão, ou a preditividade positiva, é considerada uma medida de CONFIANÇA. A confiança pode ser interpretada como a probabildiade de que a classe seja positiva dado que a previsão feita pelo modelo é positiva.
- A revocação, sensibilidade ou taxa de verdadeiro positivo, é considerada uma medida de CRENÇA ou VEROSSIMILHANÇA. A crença, ou verossimilhança, é a probabilidade de uma predição em particular ser feita dado a ocorrência de uma observação. Ela indica quanto um modelo é capaz de discriminar os casos entre as possíveis classes.

Cada uma das medidas geradas pela matriz de confusão podem ser, ou não, úteis em determinados contextos.

Taxa de erro

Não é apropriada quando as classes são desbalanceadas: suponha que em um dado domínio o número de exemplos de uma das classes seja 99% do número total de observações. Nesse caso, é comum obter taxas baixas de erro, pois um modelo que sempre retorna a classe majoritária terá uma taxa de erro de 1%. No entanto, esse modelo não acerta nenhum dado da classe majoritária. Além disso, as taxas de erro assumem custos iguais para os erros em ambas as classes, o que pode não ser desejável.

Medidas de Avaliação

Um Problema de classificação binária

Considere um sistema de babá eletrônica que ao reconhecer o choro de uma criação, toca um alarme no dispositivo usado pelos pais. Considere ainda que a classe "choro do bebê" é a classe positiva e a "ausência do choro" é a classe negativa.

- sensibilidade (ou taxa de verdadeiros positivos ou revocação): um valor alto é um bom resultado para esse classificador, uma vez que indica que em grande parte das vezes que a criança chora, o sistema avisa os pais. Quanto mais alta, melhor.
- especificidade (ou taxa de verdadeinos negativos): um valor alto indica acerto de classificador, mas neste contexto não tem muito valor SE em detrimento de outra medida como a sensibilidade, pois acertar na situação de conforto/segurança da criança SE estiver errado na situação de desconforto/perigo, não resolve o problema.
- taxa de falsos positivos: um valor alto nesta medida indica que o classificador está acusando "choro" diante de situações que não representam choro. Obviamente que se trata de um erro do classificador, e seu desempenho deve ser melhorado. Mas no contexto, a consequência deste erro não é problemática.
- precisão (ou preditividade negativa): se esta taxa é alta, então o sistema esta acusando mais situações de choro corretamente do que acusa situações de choro erroneamente.
- taxa de falsas descobertas: se esta taxa é alta, então o sistema está acusando mais situações de choro erroneamente do que situações de choros corretas.

Medidas de Avaliação

F-Score

Considera medidas de precisão e revocação, ou seja, considera tanto a capacidade do classificador em reconhecer exemplos positivos dentre todos os exemplos positivos disponíveis quanto a capacidade de não considerar exemplos negativos como positivos.

$$F\_score = \frac{(TPR * PPV)}{(TPR + PPV)/2}$$

onde:

- TPR: taxa de verdadeiros positivos ou revocação (recall);
- PPV: preditividade positiva ou precisão (precision).

Das medidas de avaliação do classificador

Qualquer medida que tenha como objetivo reduzir a avaliação de um modelo de classificação a um único valor terá, em maior ou menor grau, uma perda de informação, e pode levar a interpretações errôneas.

Um outro problema ocorre quando o classificador retorna um valor contínuo como resposta, e é necessário discretizá-lo em uma classe. Isso é feito a partir da escolha arbitrária de um limiar. Cada possível limiar produz uma matriz de confusão diferente.

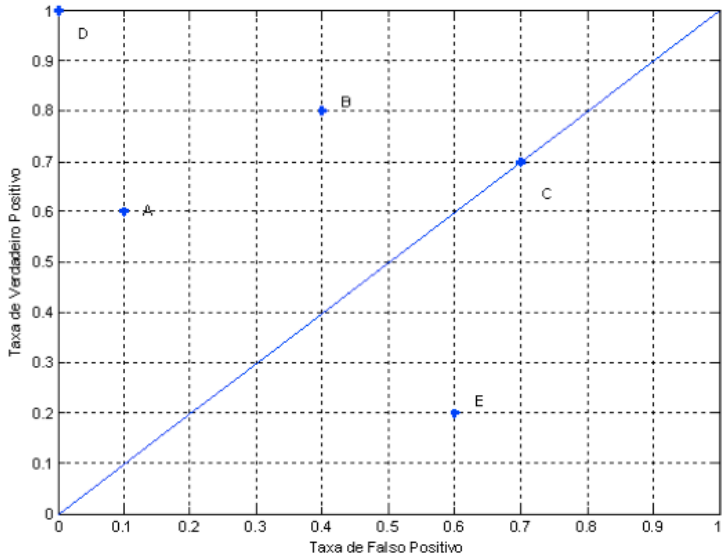
Análise ROC

Análise ROC - *Receiver Operating Characteristic* - é útil quando o domínio sob análise apresenta desproporcionalidade entre as classes, ou quando é necessário considerar diferentes custos/benefícios para diferentes erros/acertos de classificação. A análise ROC pode ainda ser útil no refinamento de modelos classificadores.

Análise ROC

Gráficos ROC - Receiver Operating Characteristic

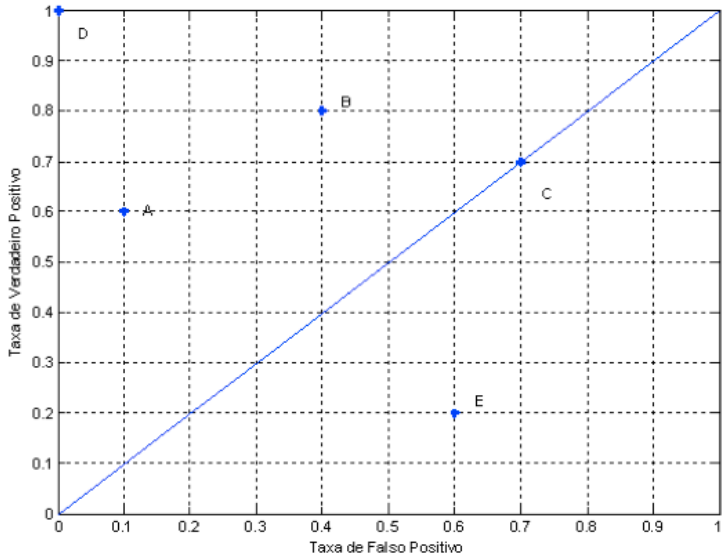
Representa cada classificador por meio de um ponto em um gráfico que contém a taxa de verdadeiros positivos (revocação ou detecção) no eixo Y e a taxa de falsos positivos (*fall-out* ou alarmes falsos) no eixo X.



Algumas considerações:

- o ponto (0,0) representa o classificador que nunca classifica um exemplo como positivo;
- o ponto (1,1) representa um classificador que sempre classifica um exemplo como positivo;
- o ponto (0,1) é o modelo perfeito;
- o ponto (1,0) é o modelo que sempre faz predições erradas;

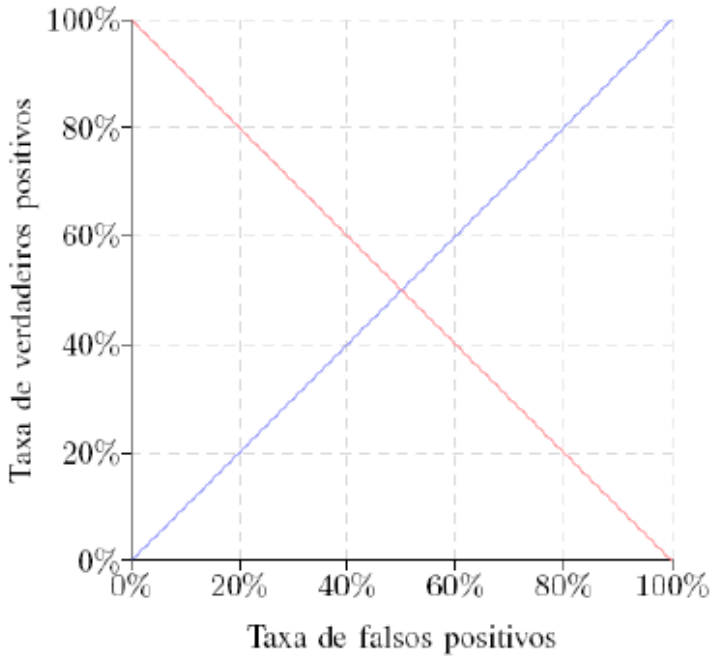
Análise ROC



- modelos próximos ao canto inferior esquerdo são *conservadores*: eles fazem uma classificação positiva somente se têm grande segurança na classificação. Como consequências, cometem poucos erros de falsos positivos, e têm baixas taxas de verdadeiros positivos;
- modelos próximos ao canto superior direito são *liberais*: eles predizem a classe positiva com maior frequência, de tal maneira que classificam a maioria dos exemplos positivos corretamente, mas também possuem altas taxas de falsos positivos.

Análise ROC

Modelos na diagonal ascendente são os modelos estocásticos: nela, cada ponto (p,p) pode ser obtido pela previsão da classe positiva com probabilidade p e da classe negativa com probabilidade $1 - p$. Acima da diagonal ascendente estão os modelos de desempenho melhor que o modelo aleatório.

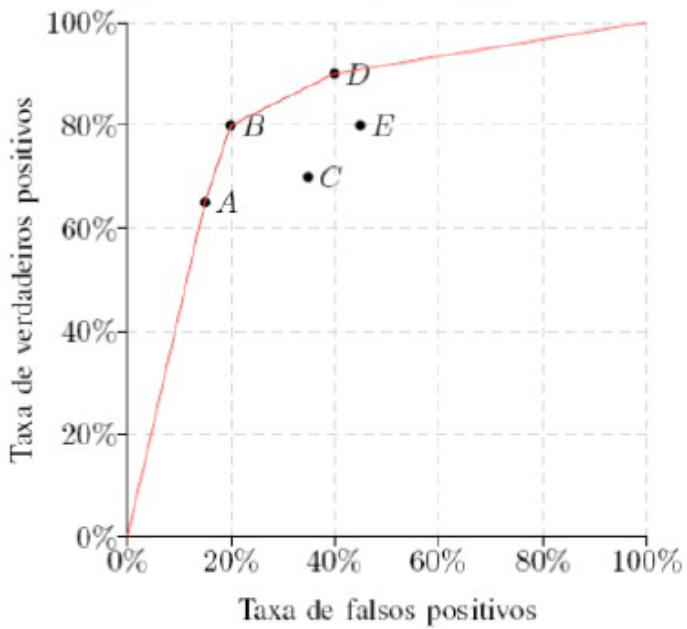


Na diagonal descendente estão os modelos de desempenho igualmente bons em ambas as classes. À esquerda dessa linha estão os modelos com desempenho melhor para a classe negativa em detrimento da positiva. E à direita estão aqueles com desempenho melhor para a classe positiva.

Análise ROC

Escolha de modelos

Os classificadores que se encontram no *Convex Hull* e que mais se aproximam do ponto (0,100), são os modelos que podem ser considerados ótimos, dada uma certa **condição operacional**. Os demais podem ser descartados. Uma condição operação pode ser: proporção de exemplos entre classes; custos/benefícios de classificação.



Em geral, um ponto no espaço ROC é melhor do que o outro se e somente se ele está acima e à esqueda do outro ponto (tem uma maior taxa de verdadeiros positivos e uma menor taxa de falsos positivos).

Análise ROC

Condição operacional

Uma condição operacional é representada por meio da inclinação de um linha no espaço ROC: a linha de isodesempenho. Nela, todos os pontos têm uma característica em comum: a taxa de erro é a mesma.

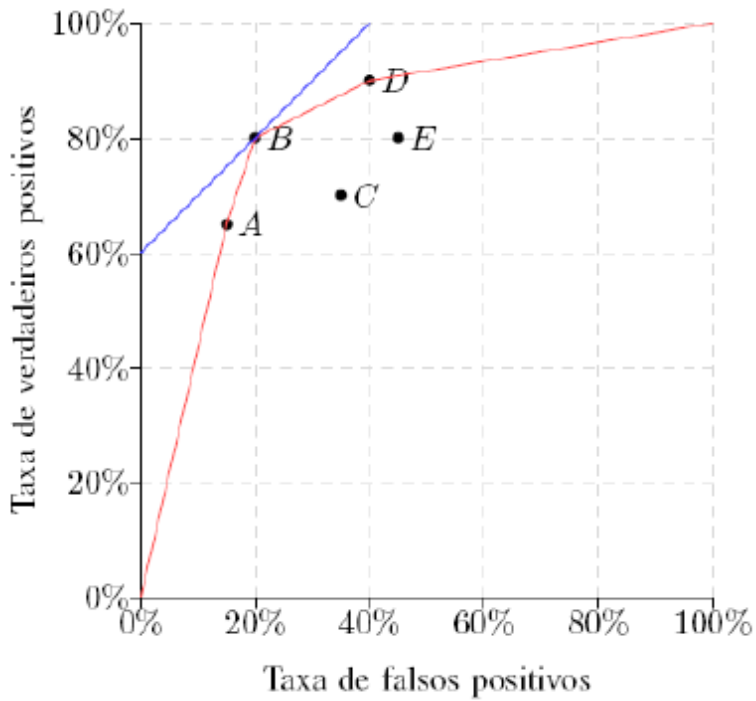
A inclinação da linha está relacionada a quanto um erro é relativamente mais importante do que outro, e o modelo ótimo para uma dada condição operacional deve estar em uma linha com a inclinação que representa a condição, e o mais próximo possível do ponto (0,1).

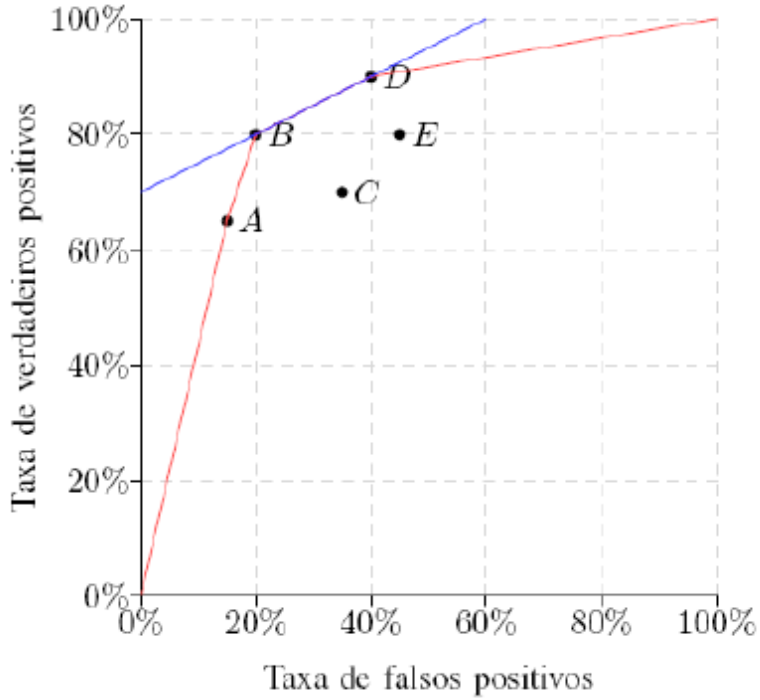
Dois pontos no espaço ROC, $(FP_1,\ TP_1)$ e $(FP_2,\ TP_2)$ tem o mesmo desempenho se

$$\frac{TP_2 - TP_1}{FP_2 - FP_1} = m$$

onde m é a inclinação da linha de isodesempenho.

Análise ROC

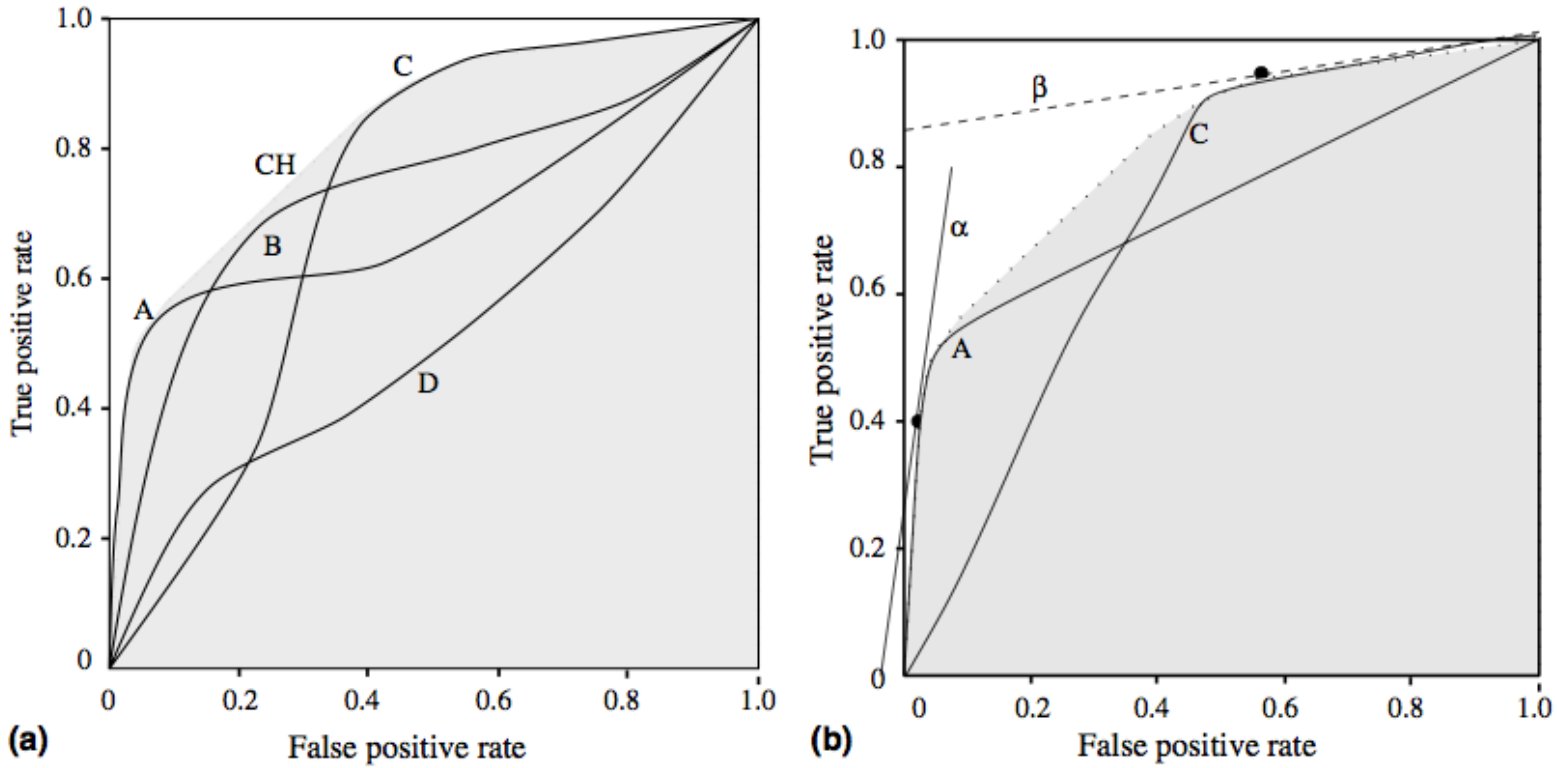




- Se inclinação da linha de isodesempenho é igual a 1, então se essa linha representa a condição operacional real, a proporção de exemplos entre as classes (ou o custo de classificar erroneamente um exemplo positivo ou negativo) é a mesma.
- Se a inclinação da linha de isodesempenho é igual a $0,5$, e essa linha representa a condição operacional real, a classe positiva será duas vezes mais populosa (ou o custo de classificar erroneamente um exemplo da classe positiva será duas vezes maior) que a classe negativa.

Análise ROC

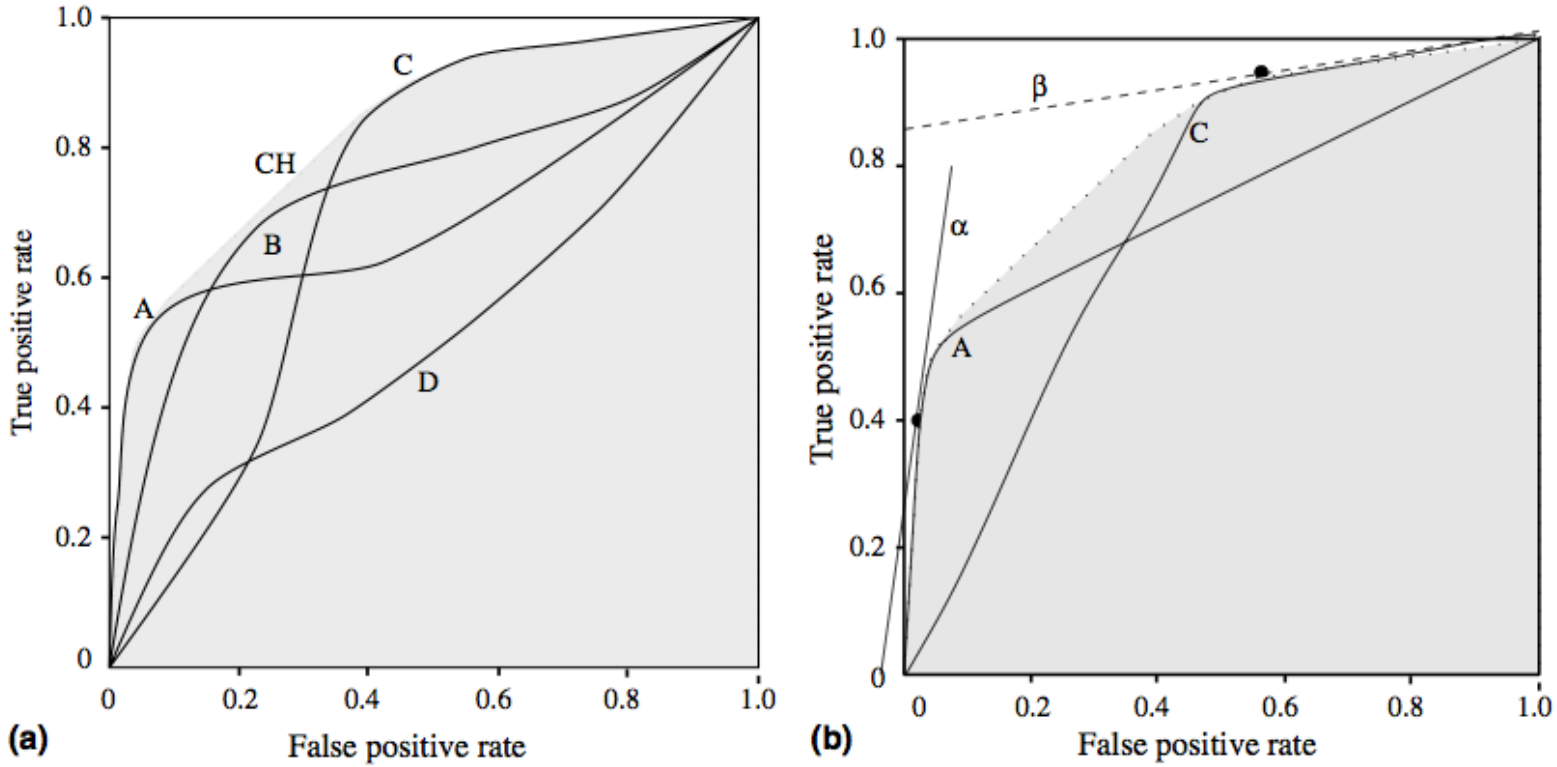
Todos os classificadores que caem sobre a linha de iso-desempenho têm o mesmo custo esperado. Linhas mais a noroeste são melhores pois dizem respeito à classificadores com menor custo esperado.



(a) O *convex hull ROC* (CH) identifica os classificadores potencialmente ótimos (A e C). B e D podem ser descartados.

Análise ROC

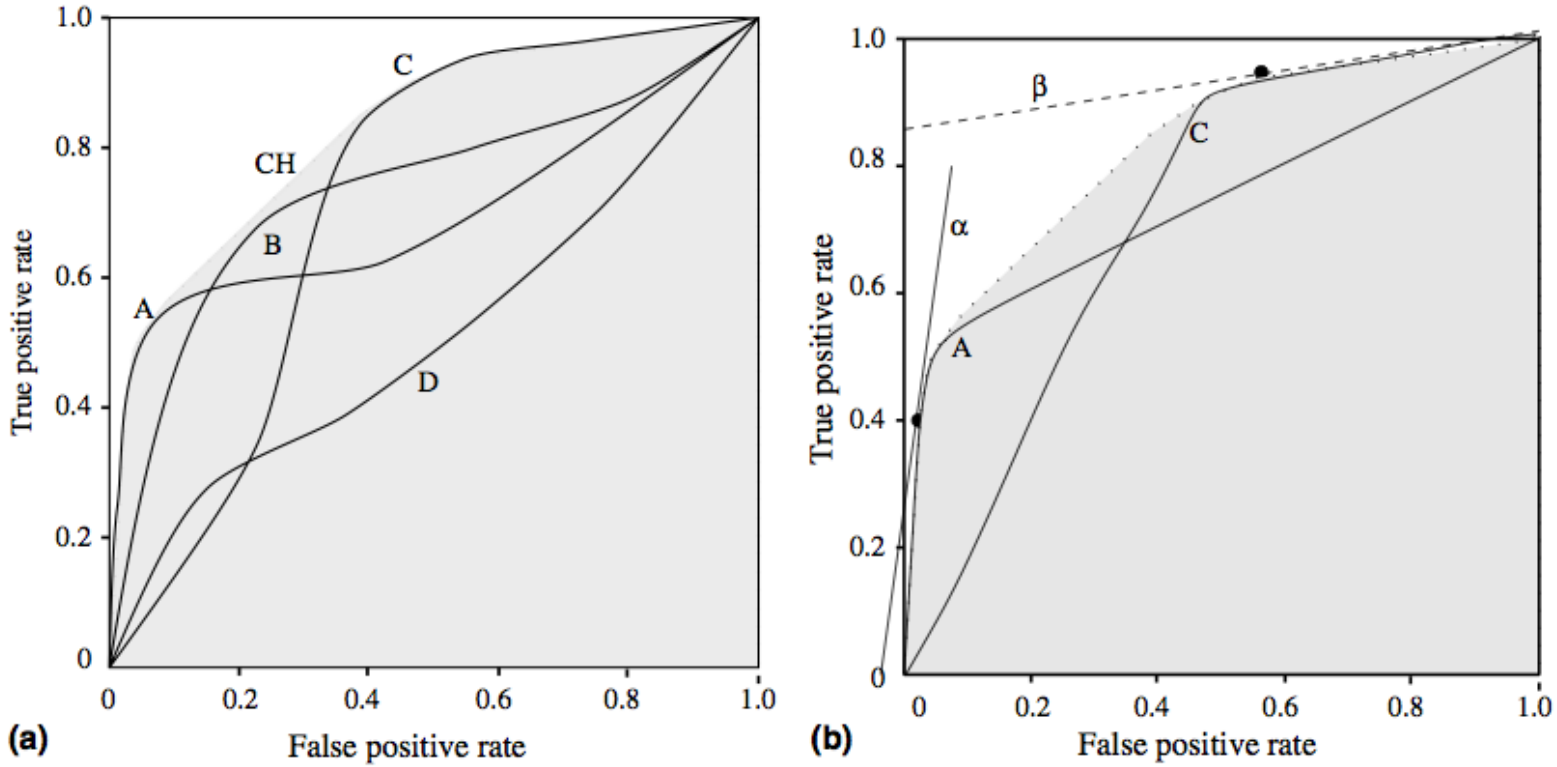
(b) Linhas α e β mostram o classificador ótimo sob diferentes condições operacionais.



Cenário α: negativos ultrapassam positivos em 10 para 1, mas falsos positivos e falsos negativos tem custo igual. A inclinação da linha é $m = 10$.

Análise ROC

(b) Linhas α e β mostram o classificador ótimo sob diferentes condições operacionais.



Cenário β: positivos e negativos são balanceados, mas falso negativo é 10 vezes mais caro que o falso positivo. A inclinação da linha é $m = \frac{1}{10}$.

Análise ROC

Gráficos ROC - Receiver Operating Characteristic

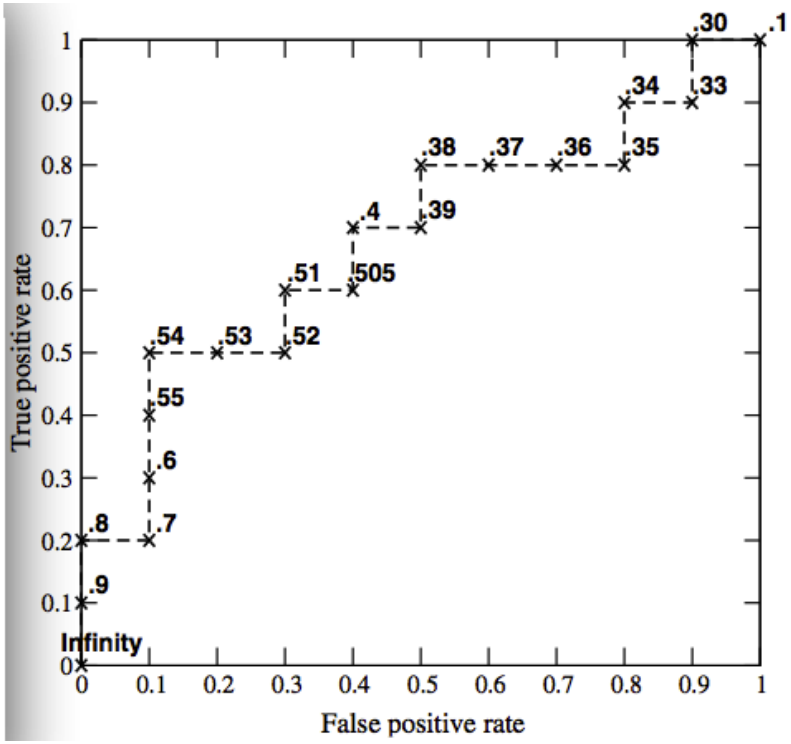
A avaliação de classificadores que produzem um valor contínuo (ou ordinal) pode ser realizada por meio da ordenação dos exemplos - simulando a escolha de vários limiares de decisão de classificação.
Para isso varia-se o limiar em todo o seu espectro, desde o valor mais restritivo até o valor mais liberal e representa-se o desempenho do sistema por uma curva no espaço ROC - a **curva ROC**.

Ordena-se todos os casos de teste de acordo com o valor contínuo predito pelo modelo. A partir desse conjunto ordenado, para cada caso desse conjunto e seguindo-se esta ordem: dê um passo de tamanho $\frac{1}{POS}$ na direção do eixo Y se o exemplo for positivo; dê um passo de tamanho $\frac{1}{NEG}$ na direção do eixo X se o exemplo for negativo.

- em que POS é o número de exemplos positivos e NEG é o número de exemplos negativos.

Análise ROC

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



- o limiar ∞ produz o ponto (0,0): ninguém é classificado como positivo (verdadeiro ou falso);
- diminuindo o limiar para 0,9, a primeira instância positiva é classificada como positiva (ponto (0; 0,1).
- ...

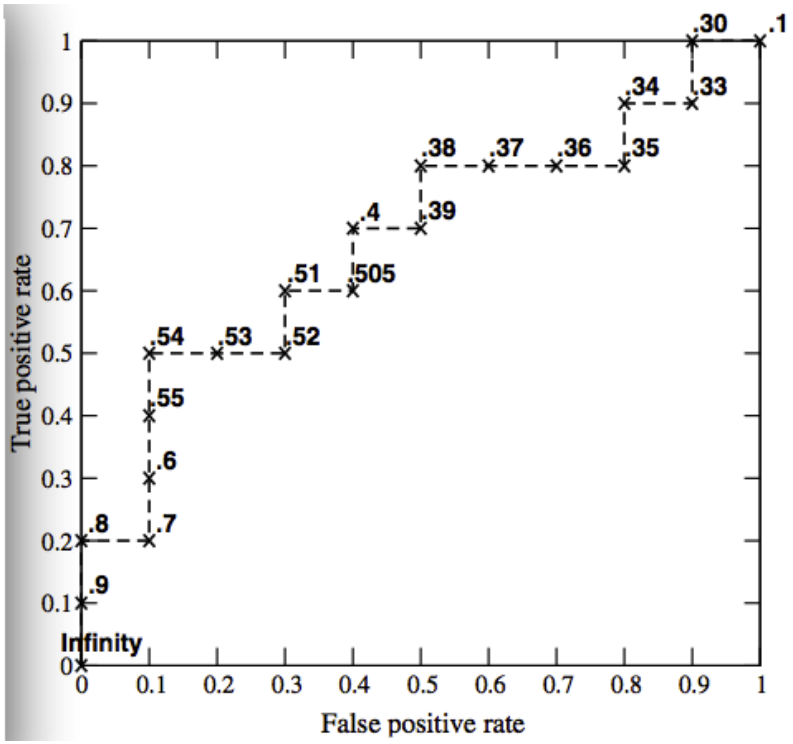
Análise ROC

Diminuir o limiar é equivalente a mover do mais conservador para o mais liberal.

Analisando o ponto (0,1;0,5)

Esse ponto produz uma acurácia de 70%: são 14 pontos sendo classificados corretamente. 50% de acerto na classe positiva e 90% de acerto na classe negativa.

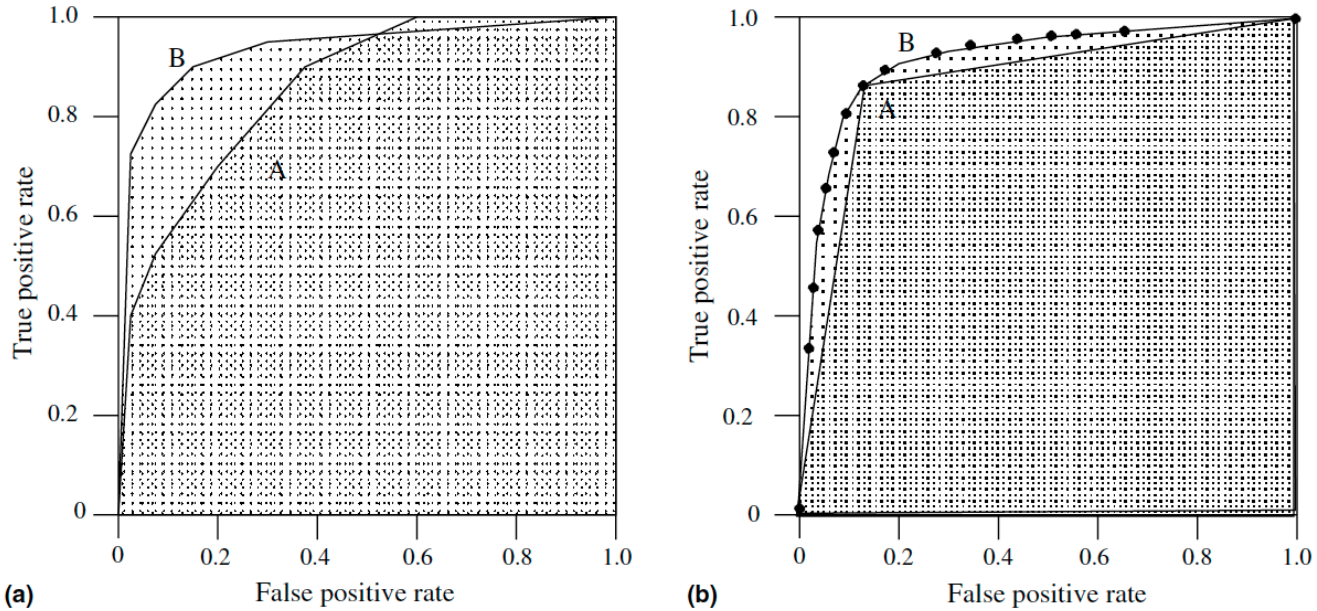
Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



Análise ROC

Area Under Curve - AUC

A área abaixo da curva ROC (AUC - *Area Under Curve*): reduz a curva a um escalar. Quanto maior a área, melhor o desempenho médio do classificador.



a) Em média, B é melhor do que A; b) A é um classificador discreto; B é um classificador probabilístico (ou de respostas contínuas).

Análise ROC

Vantagem das curvas ROC

As curvas ROC são insensíveis a mudanças nas distribuições das classes.

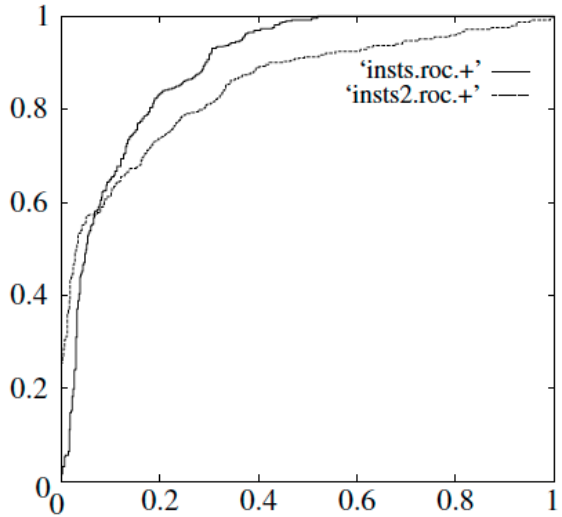
A distribuição das classes na matriz de confusão (a proporção de positivo para negativo) é o relacionamento da coluna da esquerda com a coluna da direita, ou da linha de cima com a linha de baixo, a depender da convenção da matriz.

Qualquer métrica que usa valores de ambas as colunas (ou ambas as linhas) é sensível à distribuição das classes.

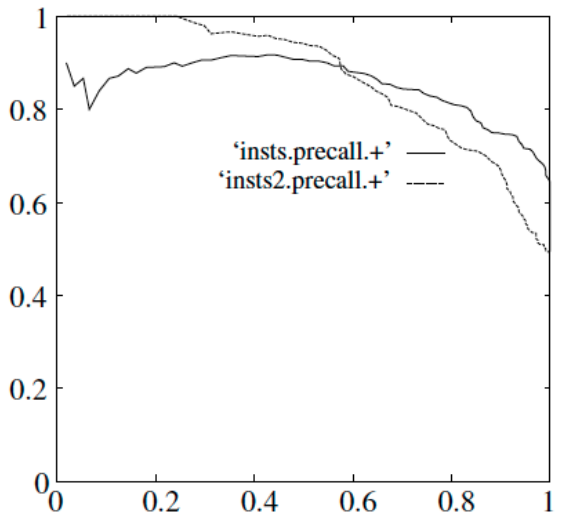
Observe que cada dimensão da curva ROC é baseada em uma coluna (ou linha), então não depende da distribuição.

Análise ROC

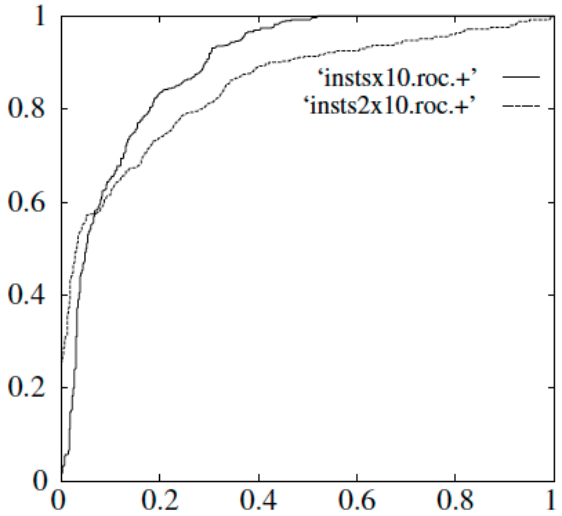
Comparação de Curvas ROC com Curvas *Precision* × *Recall*

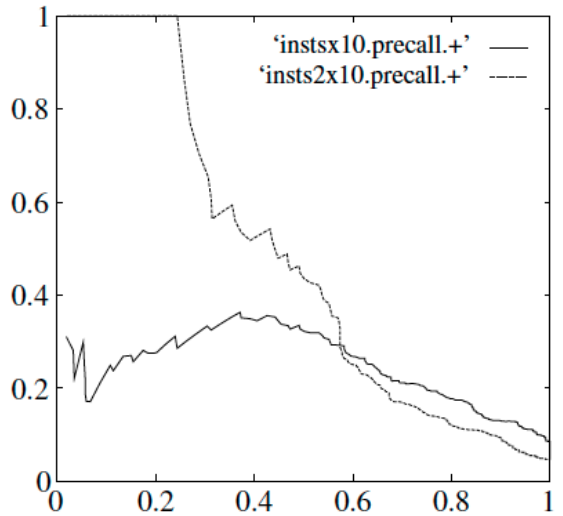


(a)



(b)





Aula 05 – Estratégias para Avaliação de Classificadores

- Clodoaldo A. M. Lima - c.lima@usp.br
- Sarajane M. Peres - sarajane@usp.br

Disciplina de Mineração de Dados
Programa de Pós Graduação em Sistemas de Informação - PPgSI
Escola de Artes, Ciências e Humanidades - EACH
Universidade de São Paulo - USP