Inteligência Artificial – ACH2016

Aula24 – Estimando o Erro e Comparação de Modelos

Norton Trevisan Roman
(norton@usp.br)

17 de junho de 2019

Estimando o Erro

Erro de treino e teste

- Quando analisamos nossos dados, tipicamente fazemos suposições
  - Cada exemplo é independente dos demais
  - Conjunto de treino e teste são distribuídos de forma idêntica

Estimando o Erro

Erro de treino e teste

- Quando analisamos nossos dados, tipicamente fazemos suposições
  - Cada exemplo é independente dos demais
  - Conjunto de treino e teste são distribuídos de forma idêntica
- Sob essas circunstâncias, o erro esperado de um modelo selecionado aleatoriamente é igual tanto no conjunto de treino quanto no de teste
  - Lembre que o modelo não foi treinado, mas selecionado aleatoriamente

Estimando o Erro

Erro de treino e teste

- Selecionar modelos aleatoriamente, contudo, não é o que fazemos
  - Definimos um modelo e o ajustamos (treinamos) no conjunto de treino, testando no conjunto de teste
  - O erro esperado no teste é então maior que no treino

Estimando o Erro

Erro de treino e teste

- Selecionar modelos aleatoriamente, contudo, não é o que fazemos
  - Definimos um modelo e o ajustamos (treinamos) no conjunto de treino, testando no conjunto de teste
  - O erro esperado no teste é então maior que no treino
- No entanto, um bom algoritmo deveria
  - Apresentar baixo erro no treino
  - Apresentar uma diferença pequena entre os erros de treino e teste

Estimando o Erro

Estimativa de Ponto

- É a tentativa de dar a melhor predição para alguma quantidade de interesse
  - Como a classificação de um determinado ponto, por exemplo

Estimando o Erro

Estimativa de Ponto

- É a tentativa de dar a melhor predição para alguma quantidade de interesse
  - Como a classificação de um determinado ponto, por exemplo
- Normalmente desconhecemos o valor real desse parâmetro para um ponto ainda não visto
  - Estimamos assim esse valor, com base no padrão observado durante o treinamento
  - Temos assim um $\hat{\theta}$, estimativa do parâmetro θ para um dado ponto

Estimando o Erro

Estimativa de Ponto

- Seja $\{\vec{x}_1, \ldots, \vec{x}_m\}$ um conjunto de m pontos independentes e identicamente distribuídos
  - Um **Estimador** para esse conjunto é uma variável aleatória $\hat{\theta}$ usada para estimar algum parâmetro θ da população
  - É uma função $\hat{\theta} = g(\vec{x}_1, \ldots, \vec{x}_m)$ dos dados

Estimando o Erro

Estimativa de Ponto

- Seja $\{\vec{x}_1, \ldots, \vec{x}_m\}$ um conjunto de m pontos independentes e identicamente distribuídos
  - Um **Estimador** para esse conjunto é uma variável aleatória $\hat{\theta}$ usada para estimar algum parâmetro θ da população
  - É uma função $\hat{\theta} = g(\vec{x}_1, \ldots, \vec{x}_m)$ dos dados
- Assim, cada modelo M_i aprendido para predizer uma classe θ pode ser visto como um estimador $\hat{\theta}_i$
  - M_i, por si só, é uma estimativa da função real que gera os dados (e cuja existência assumimos)

Estimando o Erro

Estimador de Ponto: Viés

- Definimos o viés de um estimador como sendo

$$\textit{Viés}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

  - O viés mede assim o desvio esperado do estimador em relação ao valor real do parâmetro θ

Estimando o Erro

Estimador de Ponto: Viés

- Definimos o viés de um estimador como sendo

$$\mathit{Viés}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Valor esperado do estimador $\hat{\theta}$ (sua média)

  - O viés mede assim o desvio esperado do estimador em relação ao valor real do parâmetro θ

Estimando o Erro

Estimador de Ponto: Viés

- Definimos o viés de um estimador como sendo

$$Viés(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta \qquad \mathbb{E}(\hat{\theta}) = \sum_{i=1}^{n} \hat{\theta}_i P(\hat{\Theta} = \hat{\theta}_i)$$

  - O viés mede assim o desvio esperado do estimador em relação ao valor real do parâmetro θ

Estimando o Erro

Estimador de Ponto: Viés

- Definimos o viés de um estimador como sendo

$$\textit{Viés}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Valor real do parâmetro que estimamos com $\hat{\theta}$

  - O viés mede assim o desvio esperado do estimador em relação ao valor real do parâmetro θ

Estimando o Erro

Estimador de Ponto: Viés

- Definimos o viés de um estimador como sendo

$$\textit{Viés}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

  - O viés mede assim o desvio esperado do estimador em relação ao valor real do parâmetro θ
- Um bom estimador deveria dar um valor perto de θ
  - Ou seja, um bom estimador $\hat{\theta}$ é aquele em que $\textit{Viés}(\hat{\theta}) \approx 0$
  - Se $\textit{Viés}(\hat{\theta}) = 0$ trata-se de um **estimador sem viés**

Estimando o Erro

Estimador de Ponto: Variância

- Mede o quanto de variação no estimador é esperado com diferentes amostras de treino

$$\begin{aligned} Var(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2) \\ &= \mathbb{E}(\hat{\theta}^2) - \mathbb{E}^2(\hat{\theta}) \end{aligned}$$

Estimando o Erro

Estimador de Ponto: Variância

- Mede o quanto de variação no estimador é esperado com diferentes amostras de treino

$$
\begin{aligned}
Var(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2) \\
&= \mathbb{E}(\hat{\theta}^2) - \mathbb{E}^2(\hat{\theta})
\end{aligned}
$$

Estimando o Erro

Estimador de Ponto: Variância

- Mede o quanto de variação no estimador é esperado com diferentes amostras de treino

Demonstração no final dos slides

$$
\begin{aligned}
Var(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2) \\
&= \mathbb{E}(\hat{\theta}^2) - \mathbb{E}^2(\hat{\theta})
\end{aligned}
$$

Estimando o Erro

Estimador de Ponto: Variância

- Mede o quanto de variação no estimador é esperado com diferentes amostras de treino

Demonstração no final dos slides

$$
\begin{aligned}
Var(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2) \\
&= \mathbb{E}(\hat{\theta}^2) - \mathbb{E}^2(\hat{\theta})
\end{aligned}
$$

- Mede como esperamos que $\hat{\theta}$ varie, em relação a seu valor esperado, ao treinarmos em novas amostras independentes dos dados
  - O tanto que $\hat{\theta}$ varia na medida em que mudamos os conjuntos de treino

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Da mesma forma que gostaríamos de ter um estimador com pouco viés, também gostaríamos que apresentasse baixa variância

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Da mesma forma que gostaríamos de ter um estimador com pouco viés, também gostaríamos que apresentasse baixa variância
  - Queremos então que se ajuste bem aos dados de treino (baixo viés)

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Da mesma forma que gostaríamos de ter um estimador com pouco viés, também gostaríamos que apresentasse baixa variância
  - Queremos então que se ajuste bem aos dados de treino (baixo viés)
  - E também a diferentes conjuntos de treino (baixa variância)

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Da mesma forma que gostaríamos de ter um estimador com pouco viés, também gostaríamos que apresentasse baixa variância
  - Queremos então que se ajuste bem aos dados de treino (baixo viés)
  - E também a diferentes conjuntos de treino (baixa variância)
- Note que uma grande variância é um indicativo do comportamento em dados não vistos
  - Se treinar com dados distintos leva a $\hat{\theta}$s bastante distintos, o $\hat{\theta}$ de um treino pode não corresponder ao do teste

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Como podemos conciliar isso?
  - A maneira mais comum é usando validação cruzada
  - Ou então adotando uma medida de erro que leve em conta tanto viés quanto variância

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Como podemos conciliar isso?
  - A maneira mais comum é usando validação cruzada
  - Ou então adotando uma medida de erro que leve em conta tanto viés quanto variância
- Essa medida é o erro quadrático médio

$$\begin{aligned} EQM(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \theta)^2) \\ &= Viés^2(\hat{\theta}) + Var(\hat{\theta}) \end{aligned}$$

que mede o desvio total esperado entre o estimador $\hat{\theta}$ e o valor real do parâmetro θ

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Como podemos conciliar isso?
  - A maneira mais comum é usando validação cruzada
  - Ou então adotando uma medida de erro que leve em conta tanto viés quanto variância
- Essa medida é o erro quadrático médio

$$\begin{aligned} EQM(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \theta)^2) \\ &= Viés^2(\hat{\theta}) + Var(\hat{\theta}) \end{aligned}$$

que mede o desvio total esperado entre o estimador $\hat{\theta}$ e o valor real do parâmetro θ

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Como podemos conciliar isso?
  - A maneira mais comum é usando validação cruzada
  - Ou então adotando uma medida de erro que leve em conta tanto viés quanto variância
- Essa medida é o erro quadrático médio

Demonstração no final dos slides

$$\begin{aligned} EQM(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \theta)^2) \\ &= Viés^2(\hat{\theta}) + Var(\hat{\theta}) \end{aligned}$$

que mede o desvio total esperado entre o estimador $\hat{\theta}$ e o valor real do parâmetro θ

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- $EQM(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2)$
  - Note que a variância $Var(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2)$ nada mais é que o erro quadrático esperado ao usarmos uma única observação de $\hat{\theta}$ para estimar sua média $\mathbb{E}(\hat{\theta})$

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- $EQM(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2)$
  - Note que a variância $Var(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2)$ nada mais é que o erro quadrático esperado ao usarmos uma única observação de $\hat{\theta}$ para estimar sua média $\mathbb{E}(\hat{\theta})$
- Se pudéssemos reduzir o erro indefinidamente, então teríamos:
  - $Viés(\hat{\theta}) = 0$ e $Var(\hat{\theta}) = 0$
  - Estaríamos com o modelo real, calibrado em infinitos dados
  - Esse, contudo, não é o caso

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Temos dados finitos, e geralmente com ruído
  - Ou seja, $\theta = \theta_{real} + \varepsilon$ (θ não é de fato o valor real, há um ruído ε), e assim $EQM(\hat{\theta}) > 0$

Estimando o Erro

Estimador de Ponto: Viés $\times$ Variância

- Temos dados finitos, e geralmente com ruído
  - Ou seja, $\theta = \theta_{real} + \varepsilon$ (θ não é de fato o valor real, há um ruído ε), e assim $EQM(\hat{\theta}) > 0$
- Ressurge assim nosso dilema:
  - Suponha que $EQM(\hat{\theta}) = Viés^2(\hat{\theta}) + Var(\hat{\theta}) > 0$ é mínimo
  - Então, ao reduzirmos $Viés(\hat{\theta})$, necessariamente aumentamos $Var(\hat{\theta})$ para compensar $\rightarrow$ *overfitting*
  - E, ao reduzirmos $Var(\hat{\theta})$, necessariamente aumentamos $Viés(\hat{\theta})$ para compensar $\rightarrow$ *underfitting*

Estimando o Erro

Estimador de Ponto: Capacidade

- E como reduzimos o viés de um modelo?
  - Aumentando sua **capacidade** (ou complexidade)

Estimando o Erro

Estimador de Ponto: Capacidade

- E como reduzimos o viés de um modelo?
  - Aumentando sua **capacidade** (ou complexidade)
- Informalmente, capacidade é a habilidade de um modelo de se ajustar a uma variedade de funções

Estimando o Erro

Estimador de Ponto: Capacidade

- E como reduzimos o viés de um modelo?
  - Aumentando sua **capacidade** (ou complexidade)
- Informalmente, capacidade é a habilidade de um modelo de se ajustar a uma variedade de funções
  - Modelos com alta capacidade reduzem o erro de treino
    - Em compensação, podem se ajustar até ao ruído lá existente (*overfitting*)

Estimando o Erro

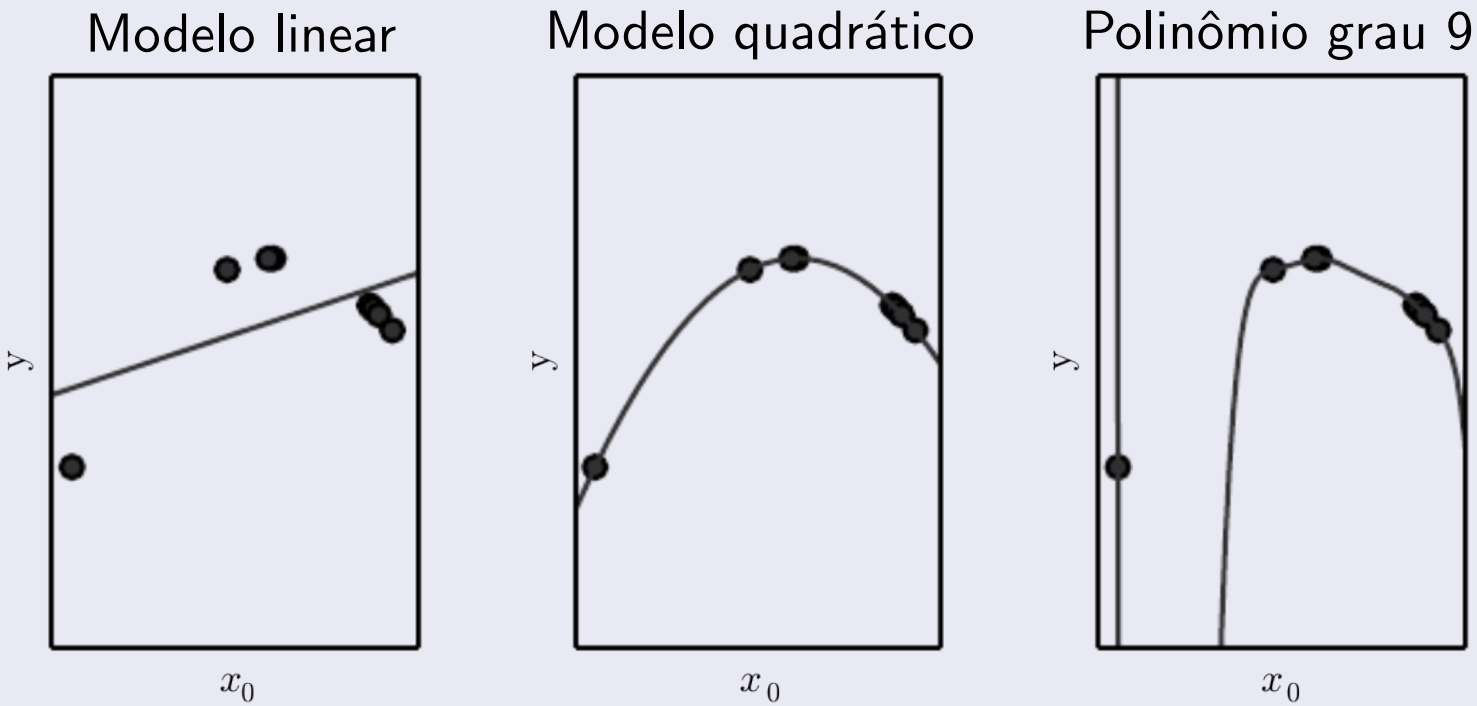
Estimador de Ponto: Capacidade

- E como reduzimos o viés de um modelo?
  - Aumentando sua **capacidade** (ou complexidade)
- Informalmente, capacidade é a habilidade de um modelo de se ajustar a uma variedade de funções
  - Modelos com alta capacidade reduzem o erro de treino
    - Em compensação, podem se ajustar até ao ruído lá existente (*overfitting*)
  - Modelos com baixa capacidade podem ter dificuldades em se ajustar aos dados de treino (*underfitting*)
    - Em compensação a diferença com dados de teste não será grande

Estimando o Erro

Estimador de Ponto: Capacidade

- Ex: Função real quadrática

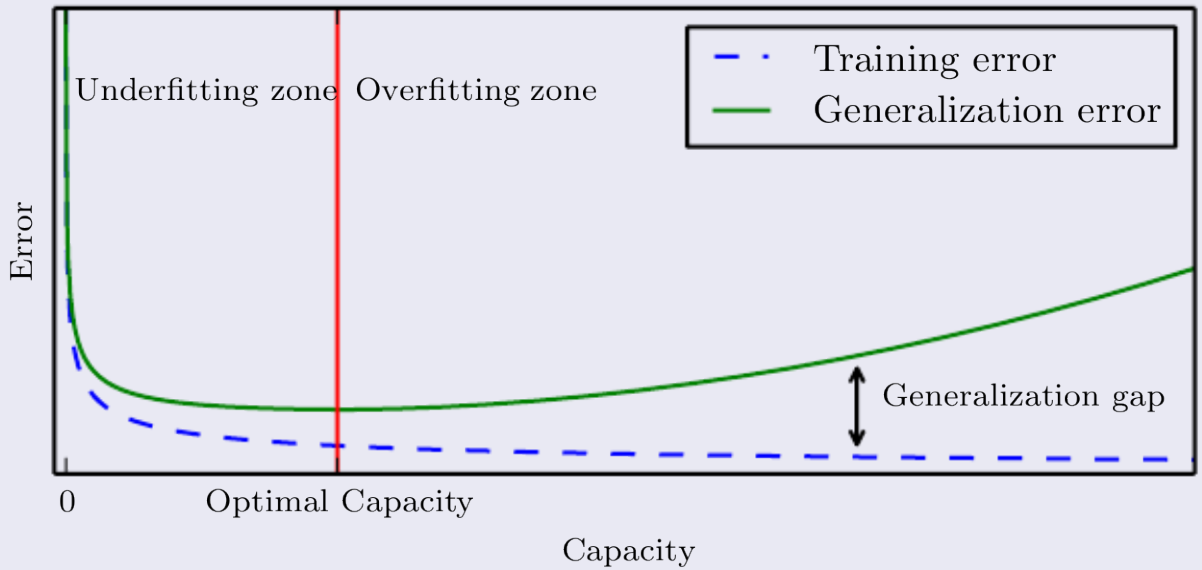


Fonte: Adaptado de DL. Goodfellow.

Estimando o Erro

Estimador de Ponto: Capacidade

- O truque está em achar a capacidade ideal
  - Devemos testar várias capacidades, identificando o ponto em que a distância entre erro de treino e de generalização começa a ficar muito grande



Fonte: DL. Goodfellow.

Medidas Alternativas de Erro

Estimando o Erro

Medidas alternativas

- Considere o seguinte cenário
  - Uma determinada doença fatal atinge 0,1% da população
  - Queremos desenvolver um sistema que, a partir de determinados atributos, diga se uma pessoa tem ou não alta probabilidade de ter essa doença

Estimando o Erro

Medidas alternativas

- Considere o seguinte cenário
  - Uma determinada doença fatal atinge 0,1% da população
  - Queremos desenvolver um sistema que, a partir de determinados atributos, diga se uma pessoa tem ou não alta probabilidade de ter essa doença
- Agora considere um sistema que, independentemente de quem entre no consultório, diz que essa pessoa não possui essa doença

Estimando o Erro

Medidas alternativas

- Considere o seguinte cenário
  - Uma determinada doença fatal atinge 0,1% da população
  - Queremos desenvolver um sistema que, a partir de determinados atributos, diga se uma pessoa tem ou não alta probabilidade de ter essa doença
- Agora considere um sistema que, independentemente de quem entre no consultório, diz que essa pessoa não possui essa doença
  - Taxa de erro de 0,1% – Maravilhoso!

Estimando o Erro

Medidas alternativas

- Qual o problema?
  - A taxa de erro supõe custos de erro idênticos para cada classe
  - Não importa se os erros ficaram bem distribuídos entre as classes ou se se concentraram em uma única delas

Estimando o Erro

Medidas alternativas

- Qual o problema?
  - A taxa de erro supõe custos de erro idênticos para cada classe
  - Não importa se os erros ficaram bem distribuídos entre as classes ou se se concentraram em uma única delas
- Em algumas situações, contudo, o custo de errar em uma classe é maior que o de errar em outra
  - É melhor termos um alarme falso (apontarmos como doentes pessoas sem a doença) do que não detectarmos um caso real (apontarmos como sadios quem possui a doença)

Estimando o Erro

Medidas alternativas

- Dada uma instância, as possíveis saídas de um classificador binário são:
  - A instância é positiva e foi classificada como positiva (verdadeiro positivo)
  - A instância é positiva e foi classificada como negativa (falso negativo)
  - A instância é negativa e foi classificada como negativa (verdadeiro negativo)
  - A instância é negativa e foi classificada como positiva (falso positivo)

Estimando o Erro

Medidas alternativas

- Dado um classificador e um conjunto de instâncias, essas 4 possibilidades podem ser organizadas em uma tabela

Estimando o Erro

Medidas alternativas

- Dado um classificador e um conjunto de instâncias, essas 4 possibilidades podem ser organizadas em uma tabela
  - A **Matriz de Confusão**, ou **Tabela de Contingência**

Hypothesized class	True class: p	True class: n
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	P	N

Fonte: [6]

Estimando o Erro

Medidas alternativas

- Valores na diagonal principal representam as decisões corretas
  - Os da secundária representam os erros – a confusão entre as classes
  - Daí o nome...

Hypothesized class	True class: p	True class: n
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	P	N

Fonte: [6]

Estimando o Erro

Medidas alternativas

- Podemos então calcular diversas métricas

Hypothesized class	True class: p	True class: n
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	**P**	**N**

Fonte: [6]

Estimando o Erro

Medidas alternativas

- Podemos então calcular diversas métricas
- **Taxa de verdadeiros positivos**
  - $t_{VP} = \frac{TP}{P}$

    Também chamada de **abrangência** ou **revocação** (*recall*)
  - De todos os positivos existentes, quantos classificamos como tal

Hypothesized class	True class: p	True class: n
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	**P**	**N**

Fonte: [6]

Estimando o Erro

Medidas alternativas

- **Taxa de falsos positivos**
  - $t_{FP} = \frac{FP}{N}$

    Também chamada de **taxa de alarmes falsos**
  - De todos os negativos, quantos classificamos incorretamente

Hypothesized class	True class: p	True class: n
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	P	N

Fonte: [6]

Estimando o Erro

Medidas alternativas

- **Especificidade**
  - $esp = 1 - t_{FP}$

		True class: p	True class: n
Hypothesized class: Y		True Positives	False Positives
Hypothesized class: N		False Negatives	True Negatives
Column totals:		**P**	**N**

Fonte: [6]

Estimando o Erro

Medidas alternativas

- **Especificidade**
  - $esp = 1 - t_{FP}$
- **Precisão**
  - $pr = \frac{TP}{TP + FP}$
  - De todos os que dissemos serem positivos, quantos realmente o eram

Hypothesized class	True class: p	True class: n
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	P	N

Fonte: [6]

Estimando o Erro

Medidas alternativas

- **Acurácia**
  - $ac = \frac{TP + TN}{P + N}$
  - De todos os dados que temos, quantos classificamos corretamente
  - Corresponde à taxa de sucesso que usamos até agora

Hypothesized class	True class: p	True class: n
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	P	N

Fonte: [6]

Estimando o Erro

Medidas alternativas

- **Medida-f** (*f-measure*)
  - $F = 2 \times \frac{pr \times t_{VP}}{pr + t_{VP}}$
  - Média harmônica de precisão e abrangência (taxa de verdadeiros positivos)
  - Também chamada de medida F_1, por conta de pr e t_{VP} estarem com pesos iguais

Hypothesized class	True class: p	True class: n
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	P	N

Fonte: [6]

Comparando Algoritmos

Medidas alternativas

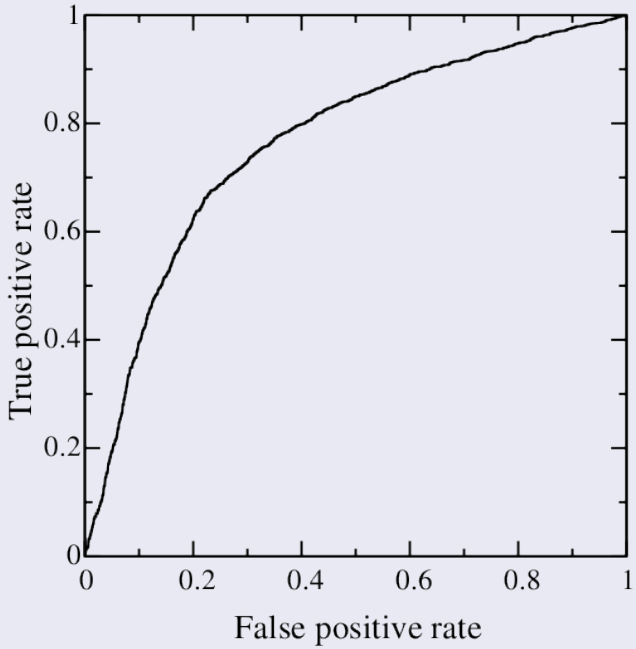
Gráficos ROC

- Um modo de usar algumas dessas métricas é criarmos um gráfico ROC
  - *Receiver operating characteristics*

Medidas alternativas

Gráficos ROC

- Um modo de usar algumas dessas métricas é criarmos um gráfico ROC
  - *Receiver operating characteristics*
- Gráfico de $t_{VP} \times t_{FP}$
  - Mostra a relação entre as taxas de acertos e de alarmes falsos
  - Relação entre benefícios e custos
- Ilustração bidimensional do desempenho de um classificador



Fonte: [6]

Medidas alternativas

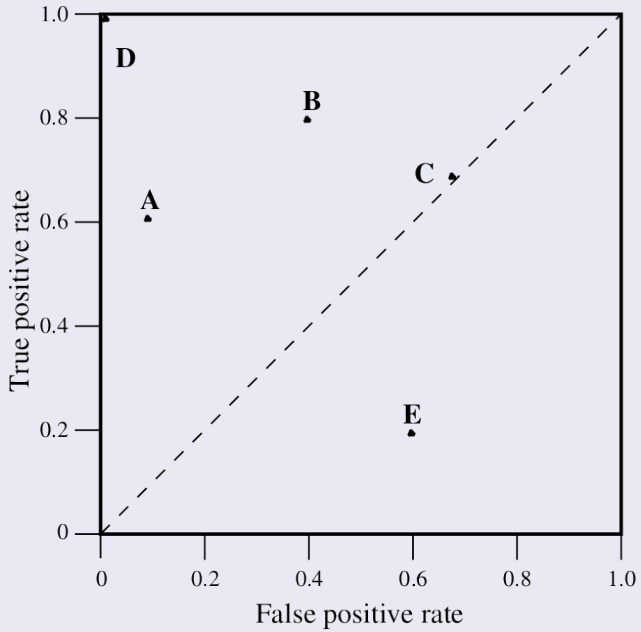
Gráficos ROC

- E como criamos um gráfico ROC?
  - Depende do tipo de classificador

Medidas alternativas

Gráficos ROC

- E como criamos um gráfico ROC?
  - Depende do tipo de classificador
- Classificadores discretos:
  - São os que produzem um rótulo de classe apenas
    - Ex: árvores de decisão
  - Representados por um único ponto no gráfico
    - Pois produzem uma única matriz de confusão para o conjunto de teste

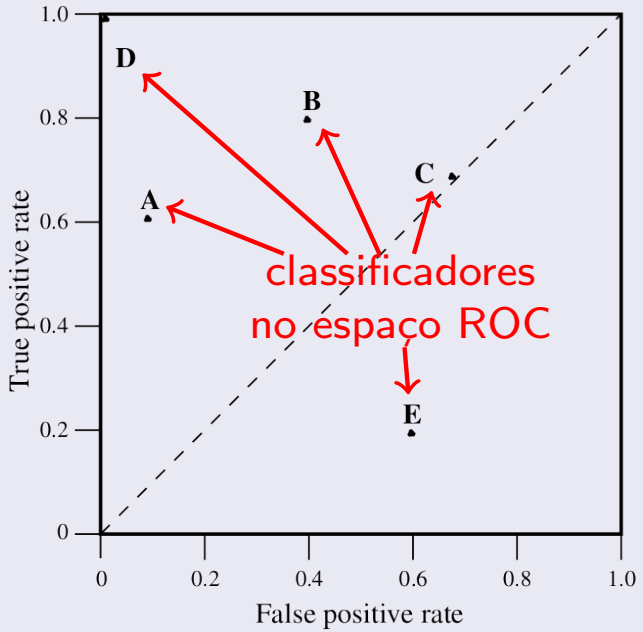


Fonte: [6]

Medidas alternativas

Gráficos ROC

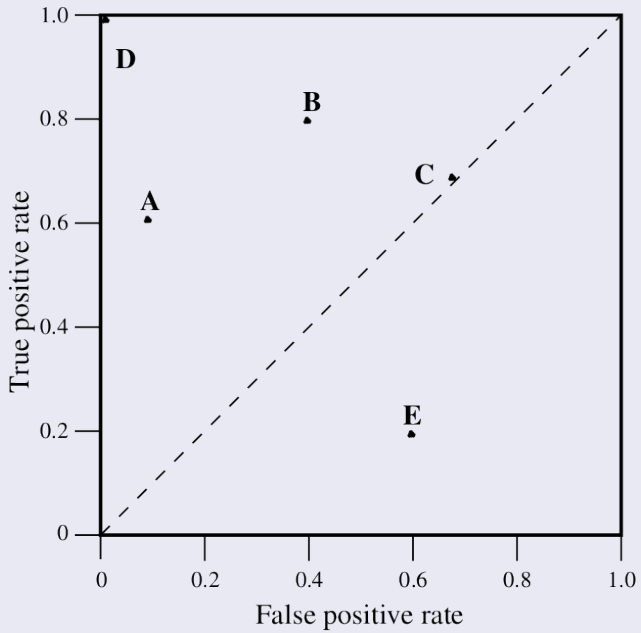
- E como criamos um gráfico ROC?
  - Depende do tipo de classificador
- Classificadores discretos:
  - São os que produzem um rótulo de classe apenas
    - Ex: árvores de decisão
  - Representados por um único ponto no gráfico
    - Pois produzem uma única matriz de confusão para o conjunto de teste



Fonte: [6]

Gráficos ROC – Classificadores discretos

- Mesmo parecendo simples, esse gráfico já nos mostra muita coisa

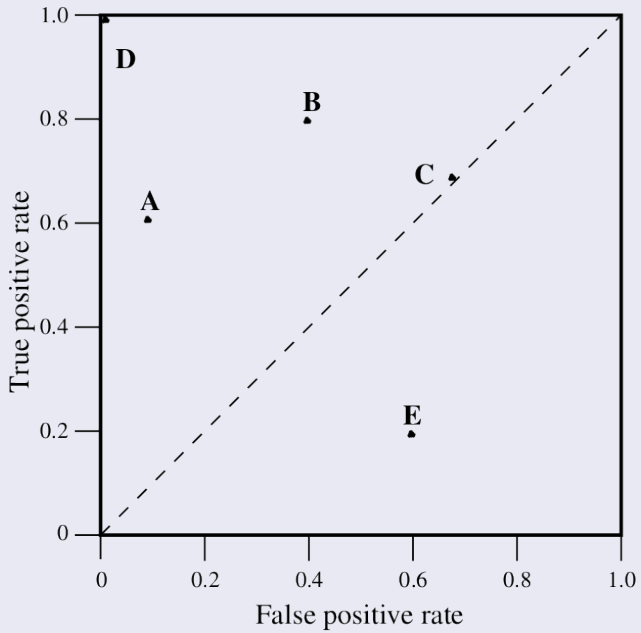


Fonte: [6]

Medidas alternativas

Gráficos ROC – Classificadores discretos

- Mesmo parecendo simples, esse gráfico já nos mostra muita coisa
  - Se queremos aumentar os positivos verdadeiros, mesmo que isso gere falsos positivos, então B é melhor que A

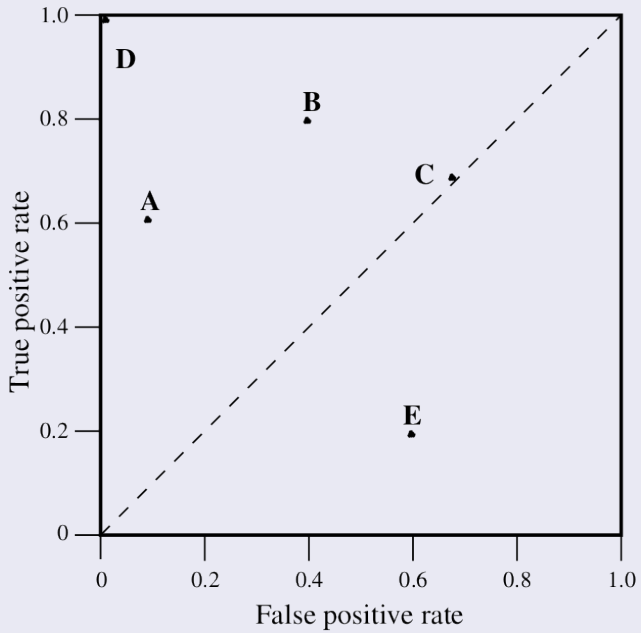


Fonte: [6]

Medidas alternativas

Gráficos ROC – Classificadores discretos

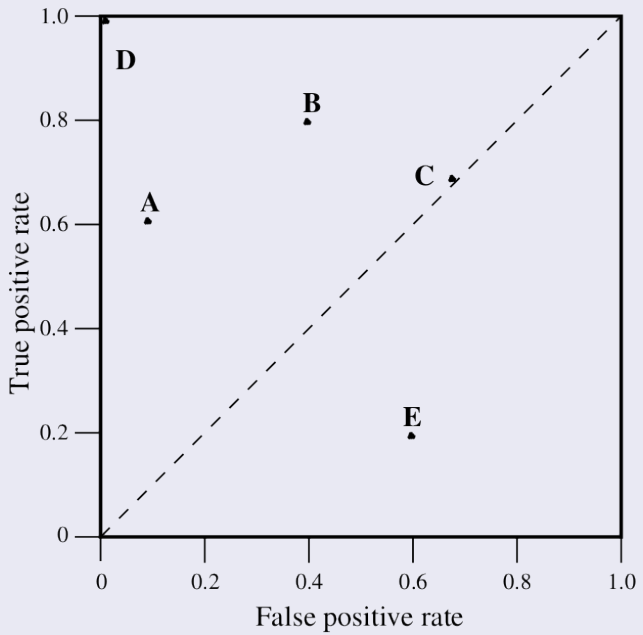
- Mesmo parecendo simples, esse gráfico já nos mostra muita coisa
  - Se queremos aumentar os positivos verdadeiros, mesmo que isso gere falsos positivos, então B é melhor que A
  - D é o melhor de todos: 100% de acertos nos positivos, sem falsos positivos



Fonte: [6]

Gráficos ROC – Classificadores discretos

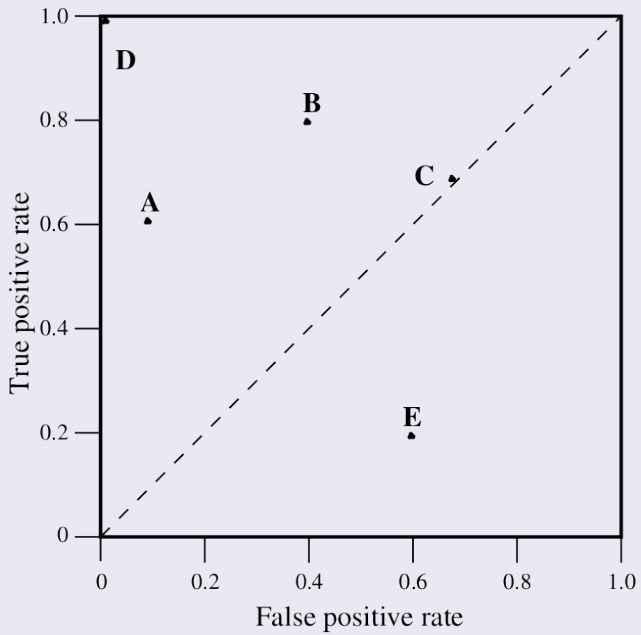
- Note também que
  - O ponto $(0, 0)$ representa a estratégia de nunca classificar algo como positivo
    - Não acerta um positivo, mas também não tem falsos positivos



Fonte: [6]

Gráficos ROC – Classificadores discretos

- Note também que
  - O ponto $(0, 0)$ representa a estratégia de nunca classificar algo como positivo
    - Não acerta um positivo, mas também não tem falsos positivos
  - A estratégia oposta, de classificar tudo como positivo, corresponde ao ponto $(1, 1)$

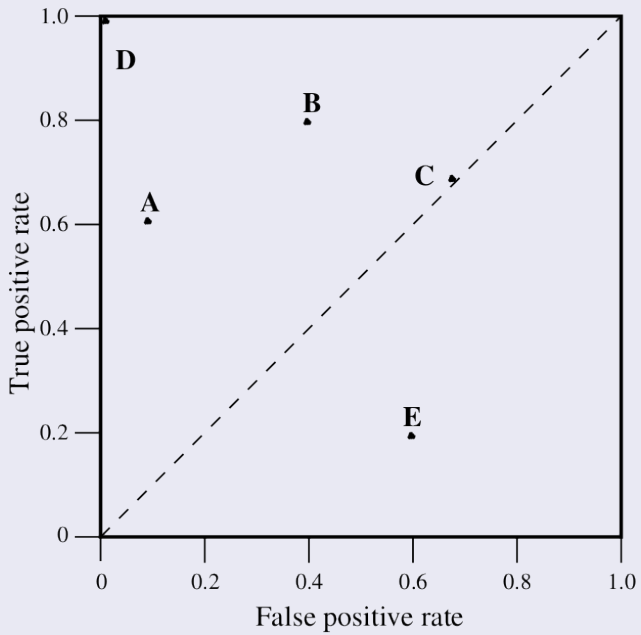


Fonte: [6]

Medidas alternativas

Gráficos ROC – Classificadores discretos

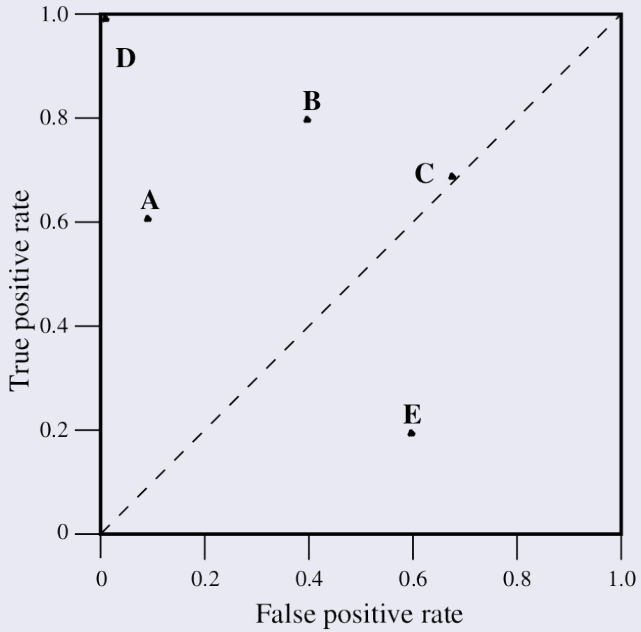
- Note também que
  - O ponto $(0, 0)$ representa a estratégia de nunca classificar algo como positivo
    - Não acerta um positivo, mas também não tem falsos positivos
  - A estratégia oposta, de classificar tudo como positivo, corresponde ao ponto $(1, 1)$
  - E $(0, 1)$ representa a classificação perfeita
    - É nesse ponto que desejamos estar



Fonte: [6]

Gráficos ROC – Classificadores discretos

- A linha $y = x$ representa a decisão aleatória
  - Se um classificador aleatoriamente escolher a classe positiva 50% das vezes, espera-se que classifique como positivos 50% dos positivos, mas também 50% dos negativos
    - O que o coloca no ponto $(0.5, 0.5)$

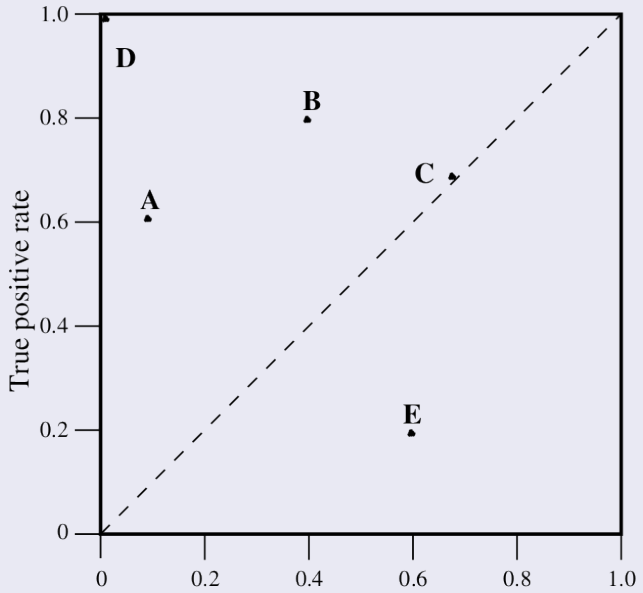


Fonte: [6]

Medidas alternativas

Gráficos ROC – Classificadores discretos

- A linha $y = x$ representa a decisão aleatória
  - Se um classificador aleatoriamente escolher a classe positiva 50% das vezes, espera-se que classifique como positivos 50% dos positivos, mas também 50% dos negativos
    - O que o coloca no ponto $(0.5, 0.5)$
  - Se escolher a positiva 90% das vezes, espera-se que acerte 90% dos positivos, e erre 90% dos negativos
    - Levando ao ponto $(0.9, 0.9)$

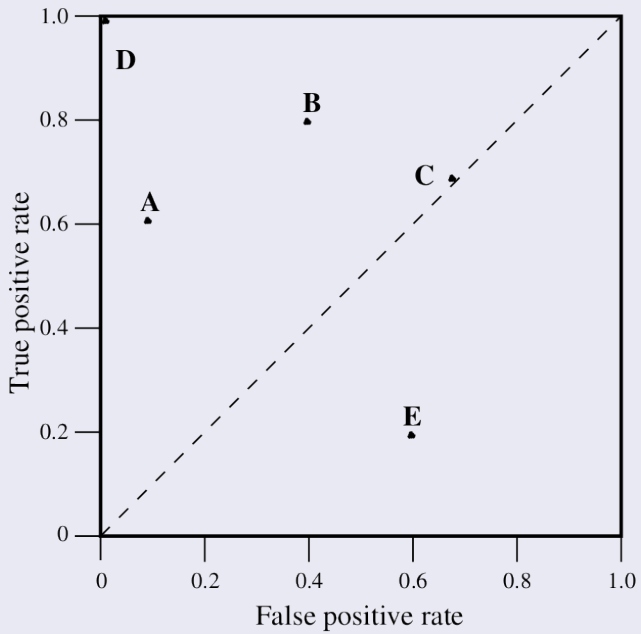


Fonte: [6]

Medidas alternativas

Gráficos ROC – Classificadores discretos

- Note que C não difere de um classificador que escolha positivo 70% das vezes
  - Dizemos que classificadores na diagonal não têm qualquer informação sobre a classe

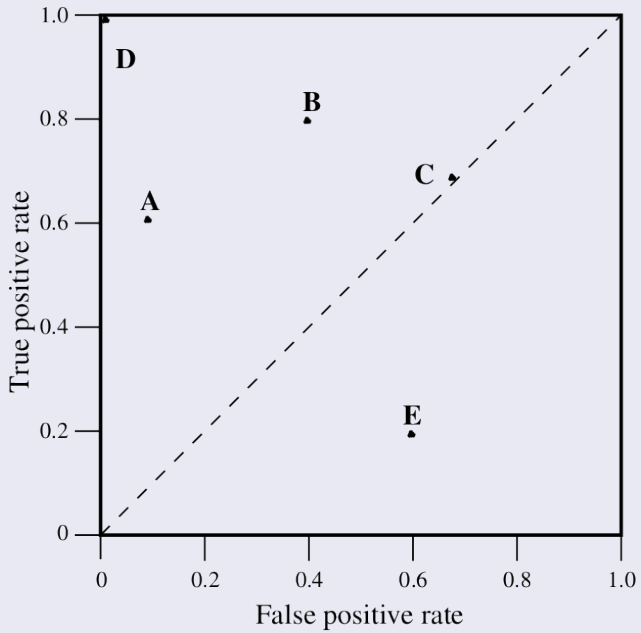


Fonte: [6]

Medidas alternativas

Gráficos ROC – Classificadores discretos

- Note que C não difere de um classificador que escolha positivo 70% das vezes
  - Dizemos que classificadores na diagonal não têm qualquer informação sobre a classe
- Mas pior é E
  - Se negarmos suas decisões obteremos B
  - Classificadores assim têm informação útil, mas a aplicam de forma errada



Fonte: [6]

Medidas alternativas

Gráficos ROC

- Classificadores contínuos
  - São os que produzem um valor (ex: uma probabilidade) representando o quanto uma instância pertence a uma determinada classe
    - Ex: Naïve Bayes e redes neurais

Medidas alternativas

Gráficos ROC

- Classificadores contínuos
  - São os que produzem um valor (ex: uma probabilidade) representando o quanto uma instância pertence a uma determinada classe
    - Ex: Naïve Bayes e redes neurais
  - Aplica-se então um limiar (*threshold*), para predizer a classe dessa instância
    - Se a saída do classificador for maior ou igual ao limiar, ele produz um **Y**, do contrário produzirá um **N**

Medidas alternativas

Gráficos ROC

- Classificadores contínuos
  - São os que produzem um valor (ex: uma probabilidade) representando o quanto uma instância pertence a uma determinada classe
    - Ex: Naïve Bayes e redes neurais
  - Aplica-se então um limiar (*threshold*), para predizer a classe dessa instância
    - Se a saída do classificador for maior ou igual ao limiar, ele produz um **Y**, do contrário produzirá um **N**
  - Transformamos assim o classificador em um classificador discreto

Gráficos ROC

Classificadores contínuos

- E como os representamos no gráfico ROC?
  - Inserimos um ponto diferente para cada limiar

Gráficos ROC

Classificadores contínuos

- E como os representamos no gráfico ROC?
  - Inserimos um ponto diferente para cada limiar
- Considere o conjunto de teste

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

Fonte: [6]

Gráficos ROC

Classificadores contínuos

- E como os representamos no gráfico ROC?
  - Inserimos um ponto diferente para cada limiar
- Considere o conjunto de teste
  - Temos 10 exemplos **p** e 10 **n**
  - Ordenados pelo valor dado pelo classificador
  - Um limiar de 0,7 classificaria como **P** apenas os 3 exemplos do topo ($score \geq 0,7$)

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

Fonte: [6]

Gráficos ROC

Classificadores contínuos

- Variamos então esse limiar
  - 0,9: apenas o exemplo 1 recebe o rótulo **P**, os demais **N**
  - 0,8: apenas 2 e 3 são **P**, os demais **N**
  - ...
  - 0,3: do 1 ao 19 são **P**, os demais **N**
  - 0,1: todos são **P**

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

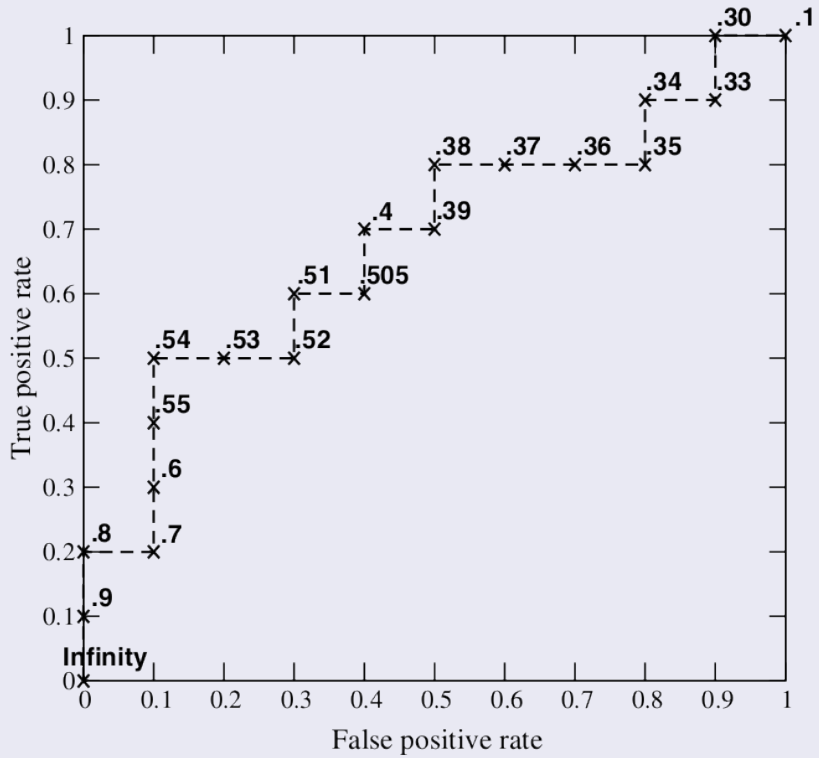
Fonte: [6]

Classificadores contínuos

- Levando ao gráfico

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

Fonte: [6]



Fonte: [6]

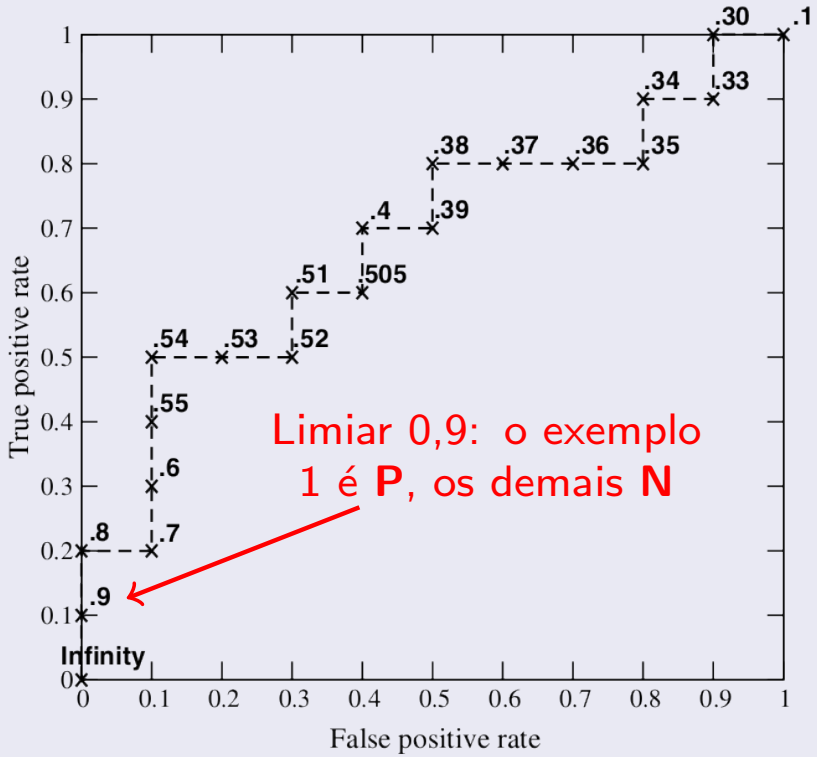
Gráficos ROC

Classificadores contínuos

- Levando ao gráfico

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

Fonte: [6]



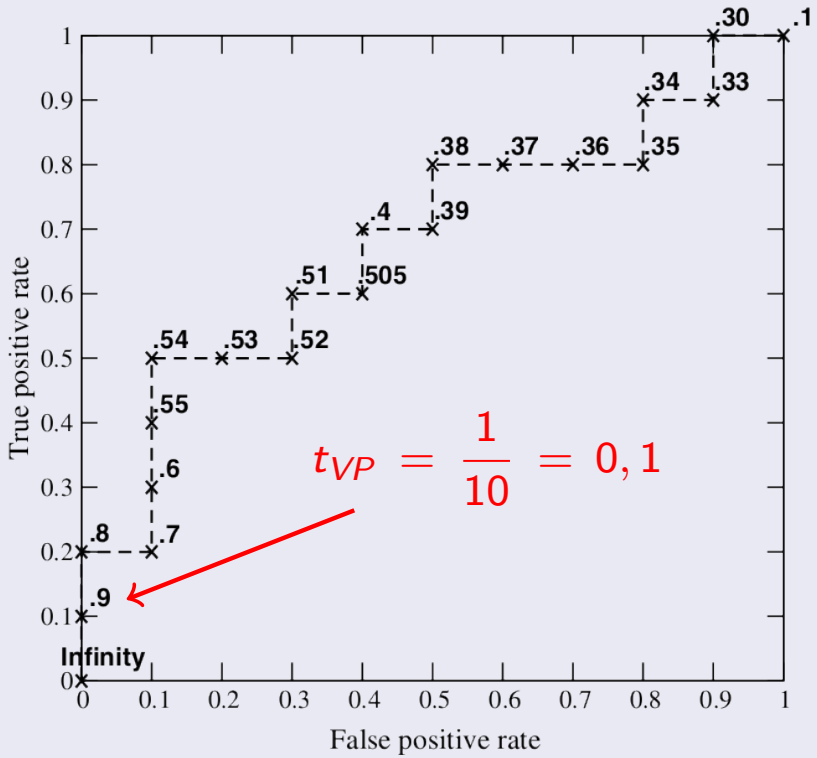
Fonte: [6]

Classificadores contínuos

- Levando ao gráfico

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

Fonte: [6]



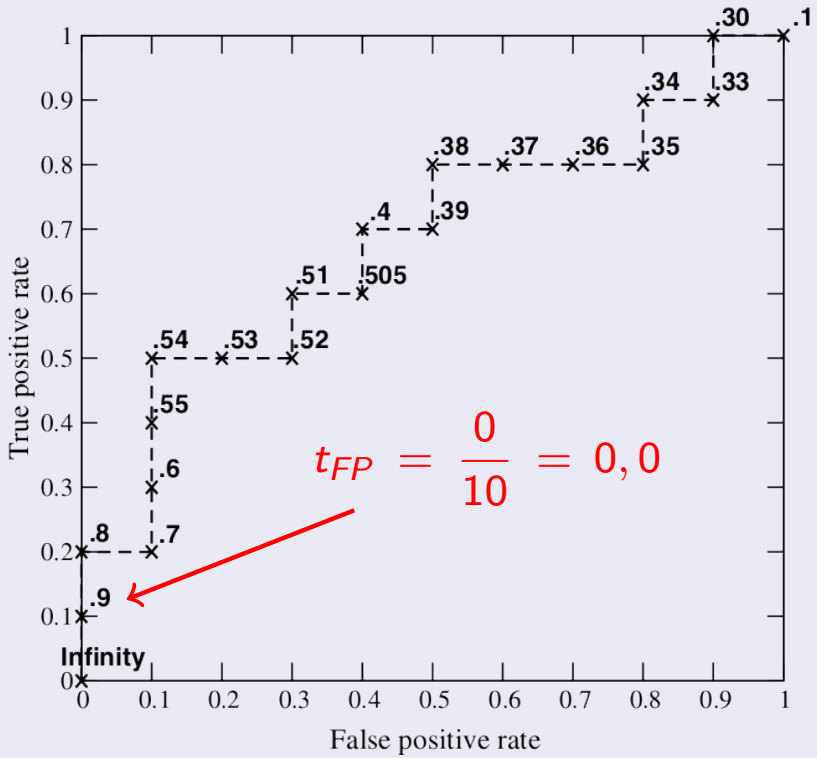
Fonte: [6]

Classificadores contínuos

- Levando ao gráfico

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

Fonte: [6]



Fonte: [6]

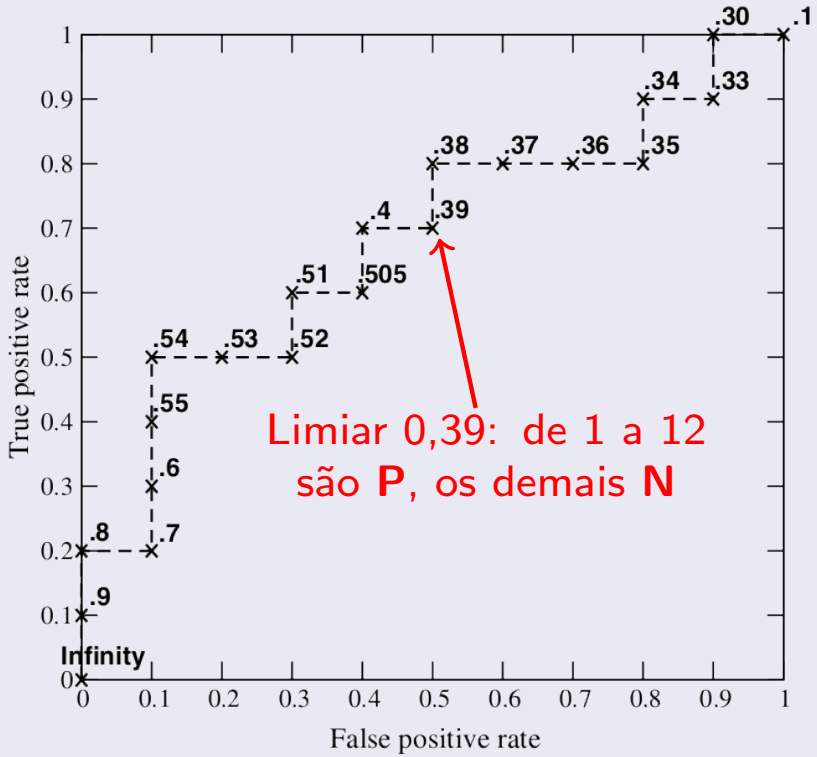
Gráficos ROC

Classificadores contínuos

- Levando ao gráfico

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

Fonte: [6]



Fonte: [6]

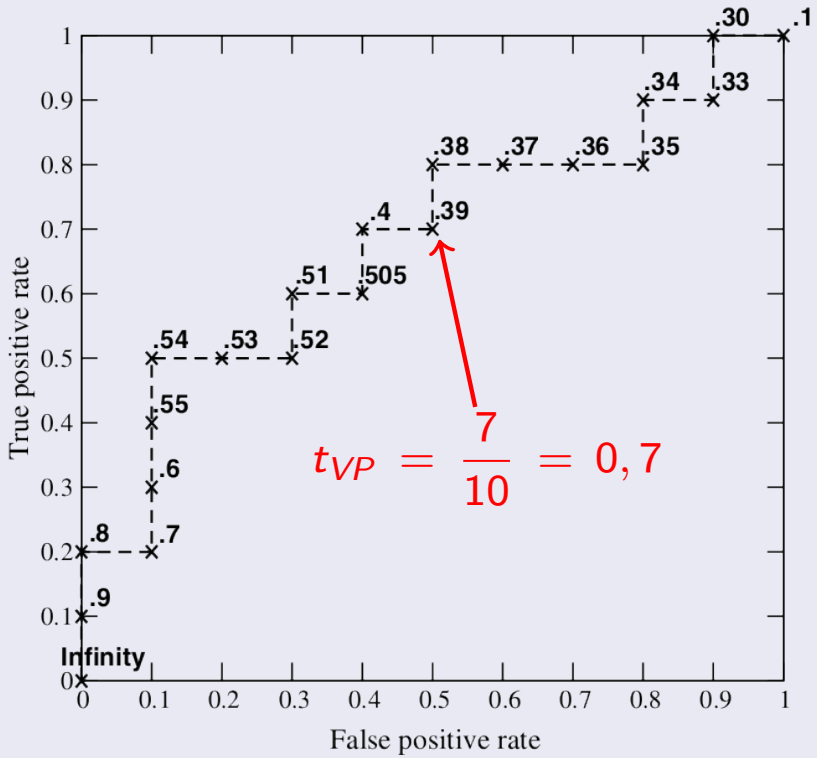
Gráficos ROC

Classificadores contínuos

- Levando ao gráfico

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

Fonte: [6]



Fonte: [6]

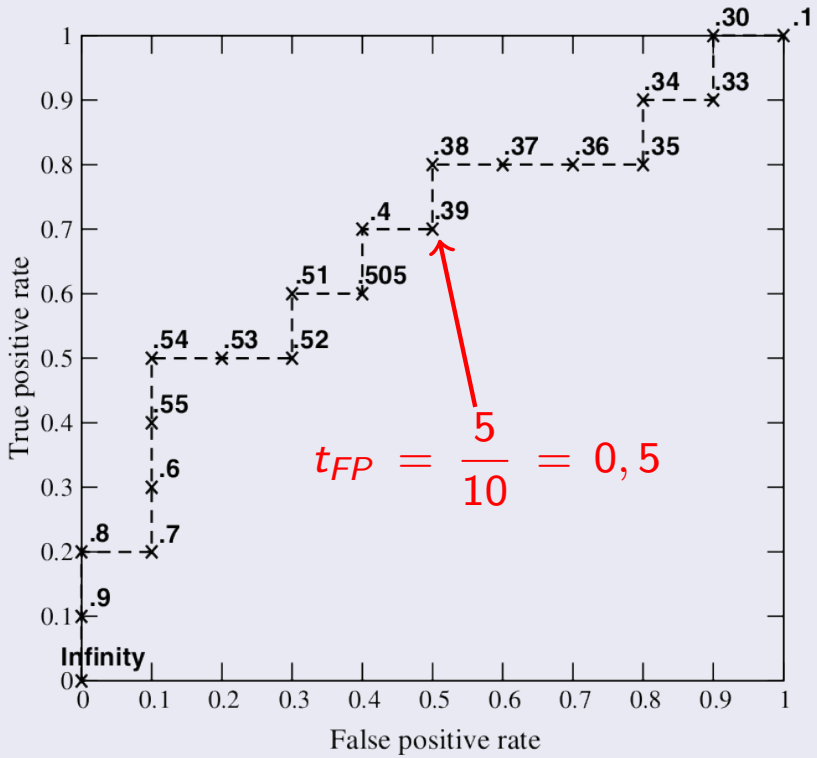
Gráficos ROC

Classificadores contínuos

- Levando ao gráfico

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

Fonte: [6]



Fonte: [6]

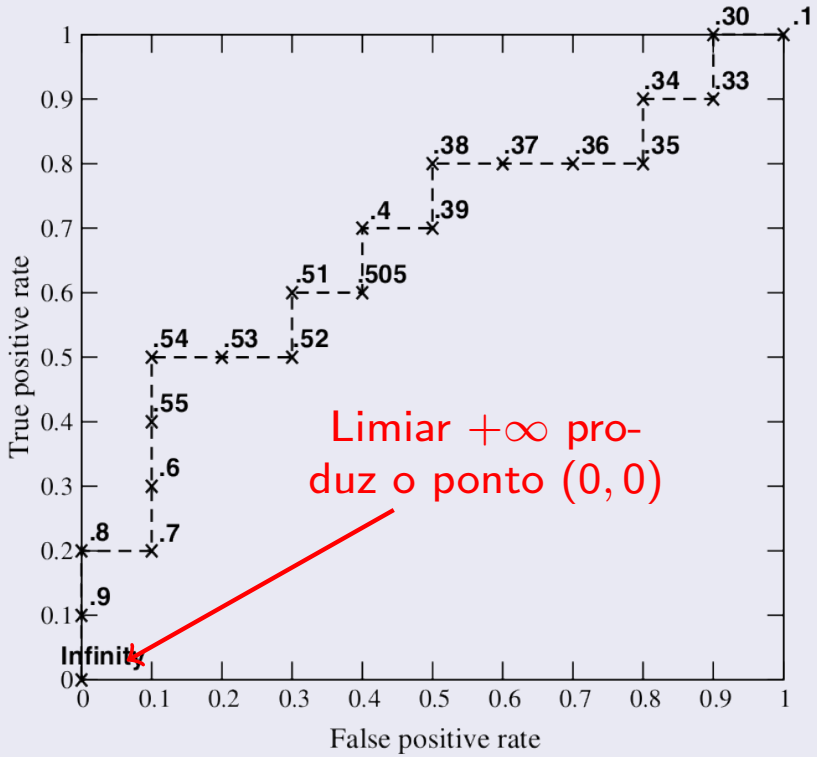
Gráficos ROC

Classificadores contínuos

- Levando ao gráfico

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

Fonte: [6]



Fonte: [6]

Gráficos ROC

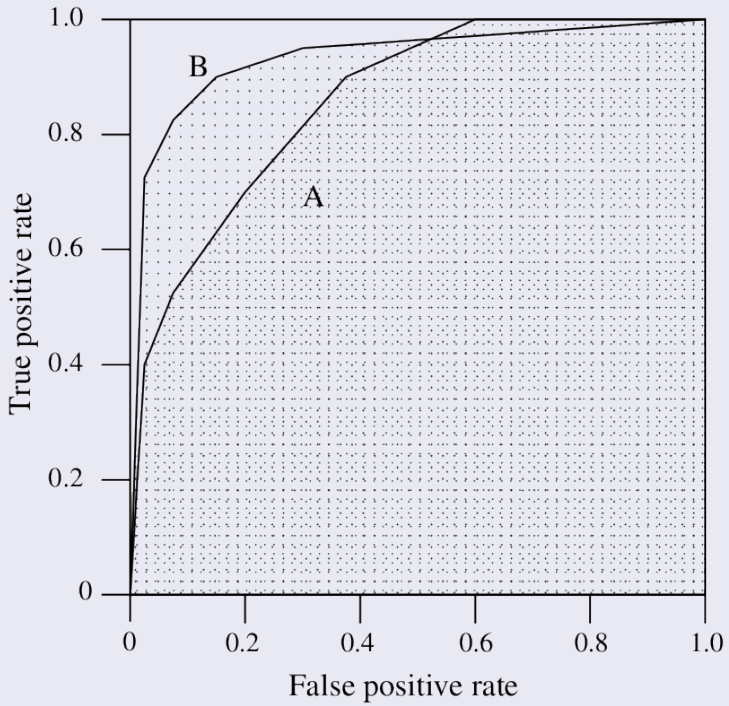
Área sob a Curva

- Para comparar classificadores pode ser útil reduzir a curva ROC a um único valor
  - Representando o desempenho esperado (ou seja, médio) de cada classificador

Gráficos ROC

Área sob a Curva

- Para comparar classificadores pode ser útil reduzir a curva ROC a um único valor
  - Representando o desempenho esperado (ou seja, médio) de cada classificador
- Uma forma de fazer isso é calcular a área sob a curva
  - $0 \leq$ área ≤ 1

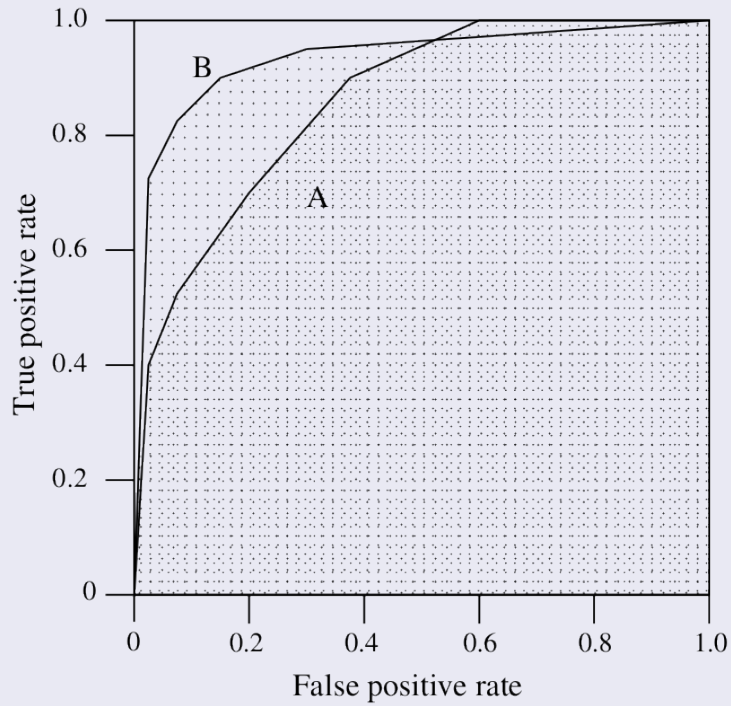


Fonte: [6]

Gráficos ROC

Área sob a Curva

- O classificador com maior área será aquele que, em média, possui o melhor desempenho
  - No caso, B

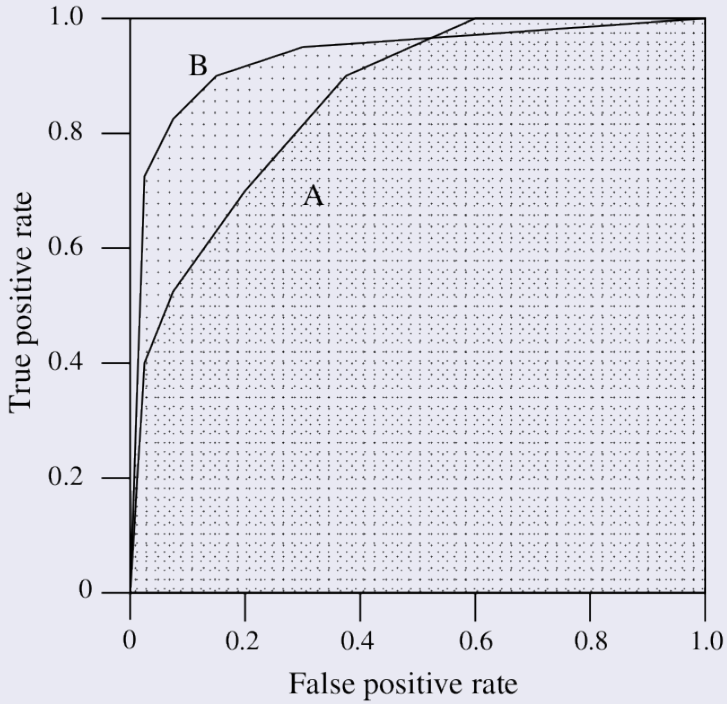


Fonte: [6]

Gráficos ROC

Área sob a Curva

- O classificador com maior área será aquele que, em média, possui o melhor desempenho
  - No caso, B
- Lembre que a escolha aleatória produz a linha $y = x$
  - Nenhum classificar real pode ter uma área menor que 0,5



Fonte: [6]

Gráficos ROC

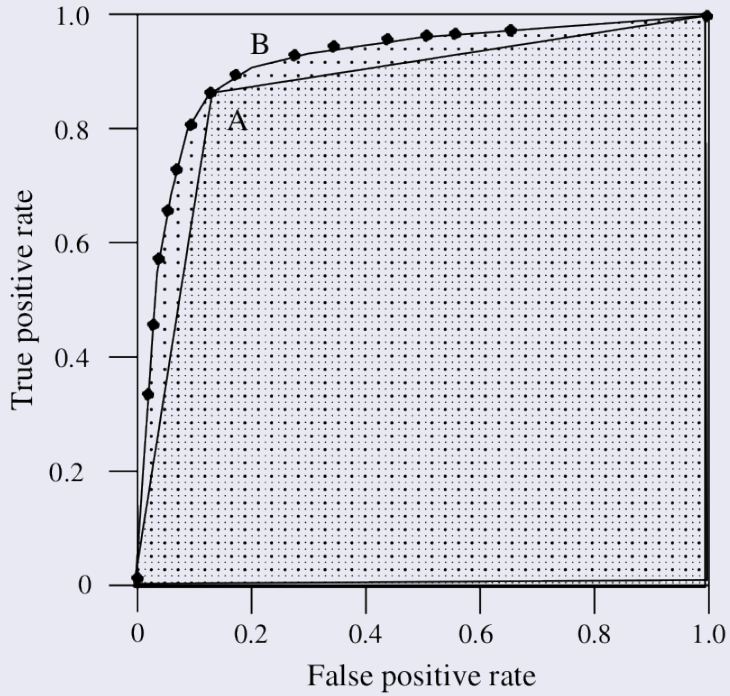
Área sob a Curva

- Podemos inclusive comparar classificadores discretos e contínuos

Gráficos ROC

Área sob a Curva

- Podemos inclusive comparar classificadores discretos e contínuos
- No exemplo, A é discreto e B contínuo
  - Vemos que A representa o desempenho de B quando este é usado com um único limiar fixo
  - Embora nesse ponto o desempenho de ambos seja igual, o desempenho de A é inferior nos demais pontos

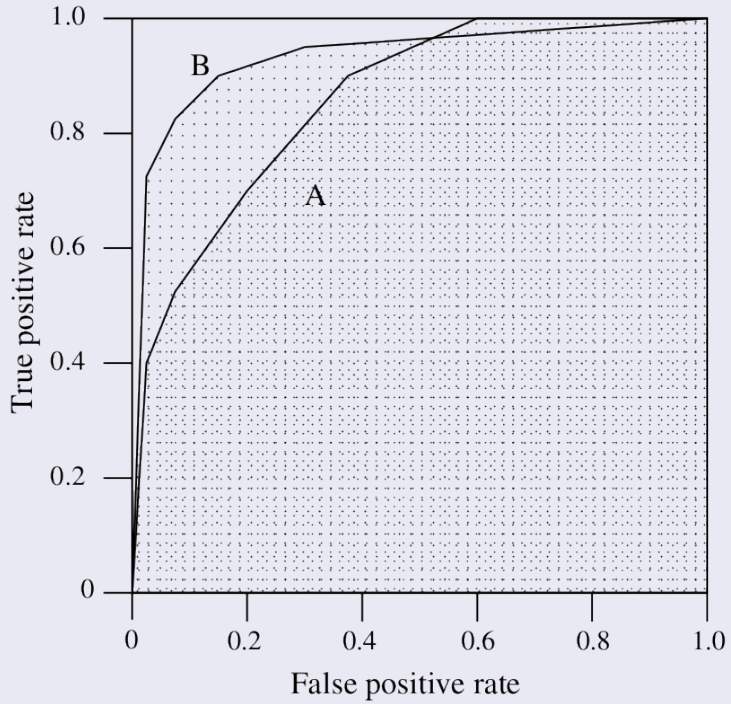


Fonte: [6]

Gráficos ROC

Área sob a Curva

- A área possui uma propriedade interessante
  - Equivale à probabilidade do classificador elencar um exemplo positivo escolhido aleatoriamente mais alto que um exemplo negativo escolhido aleatoriamente
    - Se $p = 1$, então há um limiar dividindo perfeitamente positivos de negativos
  - Corresponde ao teste de Wilcoxon (*Wilcoxon rank-test*)



Fonte: [6]

Referências

1. Witten, I.H.; Frank, E. (2005): Data Mining: Practical Machine Learning Tools and Techniques. Elsevier. 2a ed.
2. Goodfellow, I.; Bengio, Y.; Courville, A. (2016): Deep Learning. MIT Press.
3. Mitchell, T.M.: Machine Learning. McGraw-Hill. 1997.
4. Hastie, T.; Tibshirani, R.; Friedman, J. (2009): The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. 2a ed.
5. Alpaydın, E. (2010): Introduction to Machine Learning. MIT Press. 2 ed.
6. Fawcett, T. (2006): An Introduction to ROC Analysis. Pattern Recognition Letters, 27, pp. 861-874.
7. `https://www.dataquest.io/blog/learning-curves-machine-learning/`

Referências

8. http://scott.fortmann-roe.com/docs/BiasVariance.html
9. https://en.wikipedia.org/wiki/Precision_and_recall

Apêndice I

Variância

$$
\begin{aligned}
Var(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2) \\
&= \mathbb{E}(\hat{\theta}^2 - 2\hat{\theta}\mathbb{E}(\hat{\theta}) + \mathbb{E}^2(\hat{\theta})) \\
&= \mathbb{E}(\hat{\theta}^2) - 2\mathbb{E}(\hat{\theta})\mathbb{E}(\hat{\theta}) + \mathbb{E}^2(\hat{\theta}) \\
&= \mathbb{E}(\hat{\theta}^2) - 2\mathbb{E}^2(\hat{\theta}) + \mathbb{E}^2(\hat{\theta}) \\
&= \mathbb{E}(\hat{\theta}^2) - \mathbb{E}^2(\hat{\theta})
\end{aligned}
$$

Apêndice II

Erro quadrático médio

$$
\begin{aligned}
EQM(\hat{\theta}) &= \mathbb{E}((\hat{\theta} - \theta)^2) \\
&= \mathbb{E}(\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2) \\
&= \mathbb{E}(\hat{\theta}^2) - 2\theta\mathbb{E}(\hat{\theta}) + \theta^2 \\
&= \mathbb{E}(\hat{\theta}^2) - 2\theta\mathbb{E}(\hat{\theta}) + \theta^2 + \mathbb{E}^2(\hat{\theta}) - \mathbb{E}^2(\hat{\theta}) \\
&= (\mathbb{E}^2(\hat{\theta}) - 2\theta\mathbb{E}(\hat{\theta}) + \theta^2) + (\mathbb{E}(\hat{\theta}^2) - \mathbb{E}^2(\hat{\theta})) \\
&= (\mathbb{E}(\hat{\theta}) - \theta)^2 + (\mathbb{E}(\hat{\theta}^2) - \mathbb{E}^2(\hat{\theta})) \\
&= Viés^2(\hat{\theta}) + Var(\hat{\theta})
\end{aligned}
$$