



Aprendizagem Supervisionada





- ◊ “Antes imaginávamos que a automação afetaria apenas atividades operacionais, mas agora percebemos que ela atua também em atividades intelectuais”
- ◊ O objetivo de ML é extrair conhecimento e tomar decisões a partir dos dados
- ◊ Os sistemas de aprendizado de máquina são projetados para melhorar a medida que novos dados são coletados
- ◊ Basicamente podemos dividir os problemas de ML em duas categorias:
 - Supervisionados;
 - Não supervisionados;



“São apresentadas ao computador exemplos de entradas e saídas desejadas, fornecidas por um “professor”. O objetivo é aprender uma regra geral que mapeia as entradas para as saídas.”

Conceitos

Algoritmo treinado sobre dados rotulados:

- ◊ Algoritmos de Classificação (e.g. tumor maligno e/ou benigno)
 - Prediz Valores discretos
 - Algoritmos: Árvore de Decisão, Regressão Logística, KNN, etc.
- ◊ Algoritmos Regressão linear (e.g. prever o preço de um imóvel)
 - Prediz Valores contínuos
 - Algoritmos: Regressão Linear e Polinomial, SVM Regressor, Árvore de Decisão

+ Exemplos

- ◊ Identificação de fraudes
- ◊ Detecção de epidemias
- ◊ Precisão de tratamentos
- ◊ Análise de sentimento
- ◊ Filtros de spam
- ◊ Cálculo de empréstimo



Árvores de Decisão



Árvore de Decisão



Árvore de Decisão



Conceitos

- ◊ “Árvores de decisão são métodos de aprendizado de máquinas supervisionado não-paramétricos, muito utilizados em tarefas de classificação e regressão.”

- ◊ “A construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia, dificuldade de previsão, da variável alvo.”

Conceitos

- ◊ **Entropia:** A entropia da informação, no caso do aprendizado de máquina, mede a impureza de um determinado conjunto de dados. Em outras palavras mede **a dificuldade que se tem para saber qual a classificação** de cada amostra dentro do meu conjunto de dados.

Entropia

$$E(X) = - \sum_{i=1}^n p_i * \log_2 p_i$$

- ◊ **X:** é o atributo;
- ◊ **n:** quantidade de classes no conjunto de dados.
- ◊ **Pi:** probabilidade de cada uma delas acontecer para um dado atributo

“A entropia de um atributo é definida pela soma ponderada das entropias de suas partições.”

Conceitos

- ◊ **Ganho de Informação:** O ganho de informação ao contrário da entropia mede a pureza de um determinado conjunto de dados, essa definição nada mais é do que **a eficácia do atributo testado ao tentar classificar a base de dados.**

$$GI(x) = E(\text{Classe}) - E(x)$$

Conceitos

$$\diamond \quad GI(x) = E(Classe) - E(x)$$

- X: é o atributo
- E(Classe): é a entropia da classe no dataset
- E(x): é a entropia do atributo.

Definida pela soma ponderada das entropias de suas partições.

Information Gain

ESCOLA	IDADE	LABEL(Bolsa?)
Part_bolsa	>18	Não
Particular	<=18	Não
Part_bolsa	<=18	Sim
Particular	>18	Não
Pública	<=18	Sim
Pública	>18	Sim
Part_bolsa	>18	Não
Part_bolsa	<=18	Sim



Entropia da classe

◊ Entropia da **classe**:

- $E(\text{Bolsa}) = -\sum P_i \log_2 P_i$
 - ◊ $P_1(\text{bolsa=sim}) = 4/8 = 0,5$ ◊ 4 bolsistas
 - ◊ $P_2(\text{bolsa=não}) = 4/8 = 0,5$ ◊ 4 não bolsistas
- $E(\text{Bolsa}) = -P_1 \log_2 P_1 - P_2 \log_2 P_2$
- $E(\text{Bolsa}) = -0,5 \log_2 0,5 - 0,5 \log_2 0,5$
- $E(\text{Bolsa}) = 1$

Entropia do atributo

<u>ESCOLA</u>	IDADE	LABEL(Bolsa?)
Part_bolsa	>18	Não
Particular	<=18	Não
Part_bolsa	<=18	Sim
Particular	>18	Não
Pública	<=18	Sim
Pública	>18	Sim
Part_bolsa	>18	Não
Part_bolsa	<=18	Sim

Bolsa?	PU	PA	PB
Não	0	2	2
Sim	2	0	2

Information Gain

- ◊ Entropia do atributo (Escola):
 - $E_{esc}(Esc=PU) = -\sum P_i \cdot \log_2 P_i$
 - $P_1(\text{bolsa=sim}|Esc=PU) = 2/2 = 1$
 - $P_2(\text{bolsa=nao}|Esc=PU) = 0/2 = 0$
 - $E_{esc}(Esc=PU) = -P_1 \log_2 P_1 - P_2 \log_2 P_2$
 - $E_{esc}(Esc=PU) = -1 \log_2 1 - 0 \log_2 0$
 - **$E_{esc}(Esc=PU)=0$**

Bolsa?	PU	PB	PA
Não	0	2	2
Sim	2	2	0

Information Gain

◇ Entropia do atributo (Escola):

- $E_{esc}(Esc=PB) = -\sum P_i \cdot \log_2 P_i$
 - $P_1(\text{bolsa=sim}|Esc=PB) = 2/4 = 1/2$
 - $P_2(\text{bolsa=nao}|Esc=PB) = 2/4 = 1/2$
- $E_{esc}(Esc=PB) = -P_1 \log_2 P_1 - P_2 \log_2 P_2$
- $E_{esc}(Esc=PB) = -0,5 \log_2 0,5 - 0,5 \log_2 0,5$
- **$E_{esc}(Esc=PB)=1$**

Bolsa?	PU	PB	PA
Não	0	2	2
Sim	2	2	0

Information Gain

◇ Entropia do atributo (Escola):

- $E_{esc}(Esc=PA) = -\sum Pi * \log_2 Pi$
 - $P_1(bolsa=sim|Esc=PA) = 0/2 = 0$
 - $P_2(bolsa=nao|Esc=PA) = 2/2 = 1$
- $E_{esc}(Esc=PA) = -P_1 \log_2 P_1 - P_2 \log_2 P_2$
- $E_{esc}(Esc=PA) = -0 \log_2 0 - 1 \log_2 1$
- **$E_{esc}(Esc=PA)=0$**

Bolsa?	PU	PB	PA
Não	0	2	2
Sim	2	2	0



Information Gain

“A entropia de um atributo é definida pela soma ponderada das entropias de suas partições.”

- ◊ Entropia do atributo (Escola) $\Rightarrow E(\text{Escola}) = \sum P_i * E_{\text{esc}}(i)$
 - $E(\text{Esc=PU}) = 0$
 - $E(\text{Esc=PB}) = 1$
 - $E(\text{Esc=PA}) = 0$
- ◊ $E(\text{Esc}) = P(\text{Esc=PU}) * E(\text{Esc=PU}) + P(\text{Esc=PB}) * E(\text{Esc=PB}) + P(\text{Esc=PA}) * E(\text{Esc=PA})$
- ◊ $E(\text{Esc}) = (2/4) * 0 + (4/8) * 1 + (2/8) * 0$
- ◊ $E(\text{Esc}) = 0,5$

Fazendo todos esses cálculos para a idade temos:

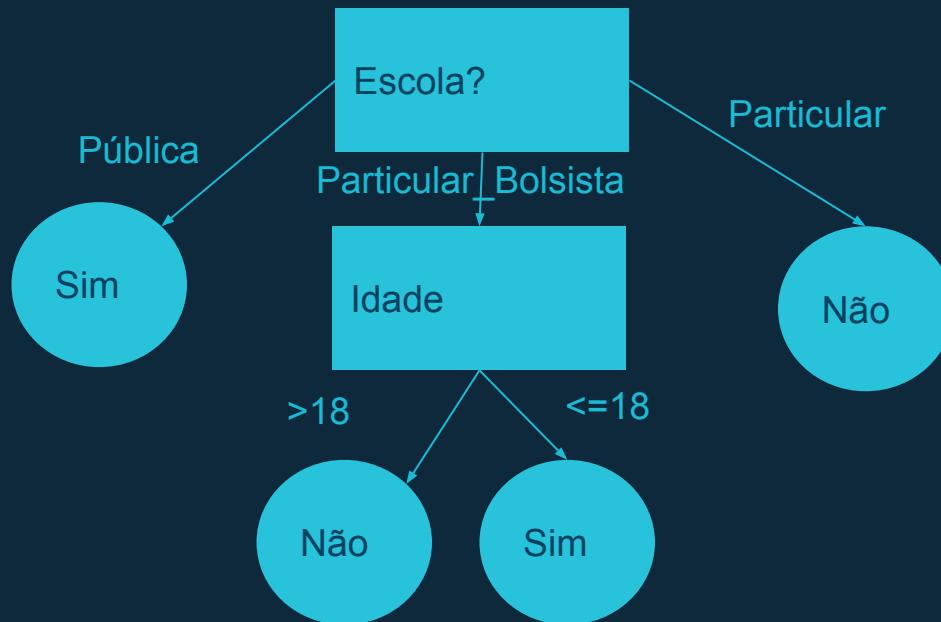
- ◊ $E(\text{Idade}) = 0,81$

Information Gain

- ◊ $E(\text{Bolsa})=1$
- ◊ $E(\text{Esc}) = 0,5$
- ◊ $E(\text{Idade}) = 0,81$
- ◊ $GI(\text{Esc}) = E(\text{Bolsa}) - E(\text{Esc}) = 1 - 0,5 = 0,5$
- ◊ $GI(\text{Idade}) = E(\text{Bolsa}) - E(\text{Idade}) = 1 - 0,811 = 0,189$

Qual o melhor atributo?

Indução da árvore

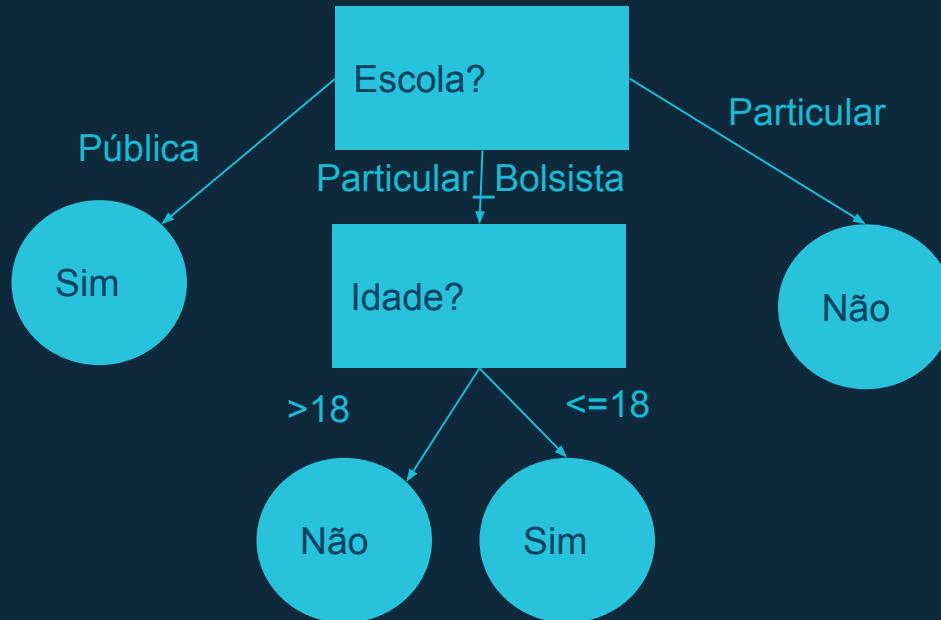


Algoritmo de Indução

Algoritmo de indução:

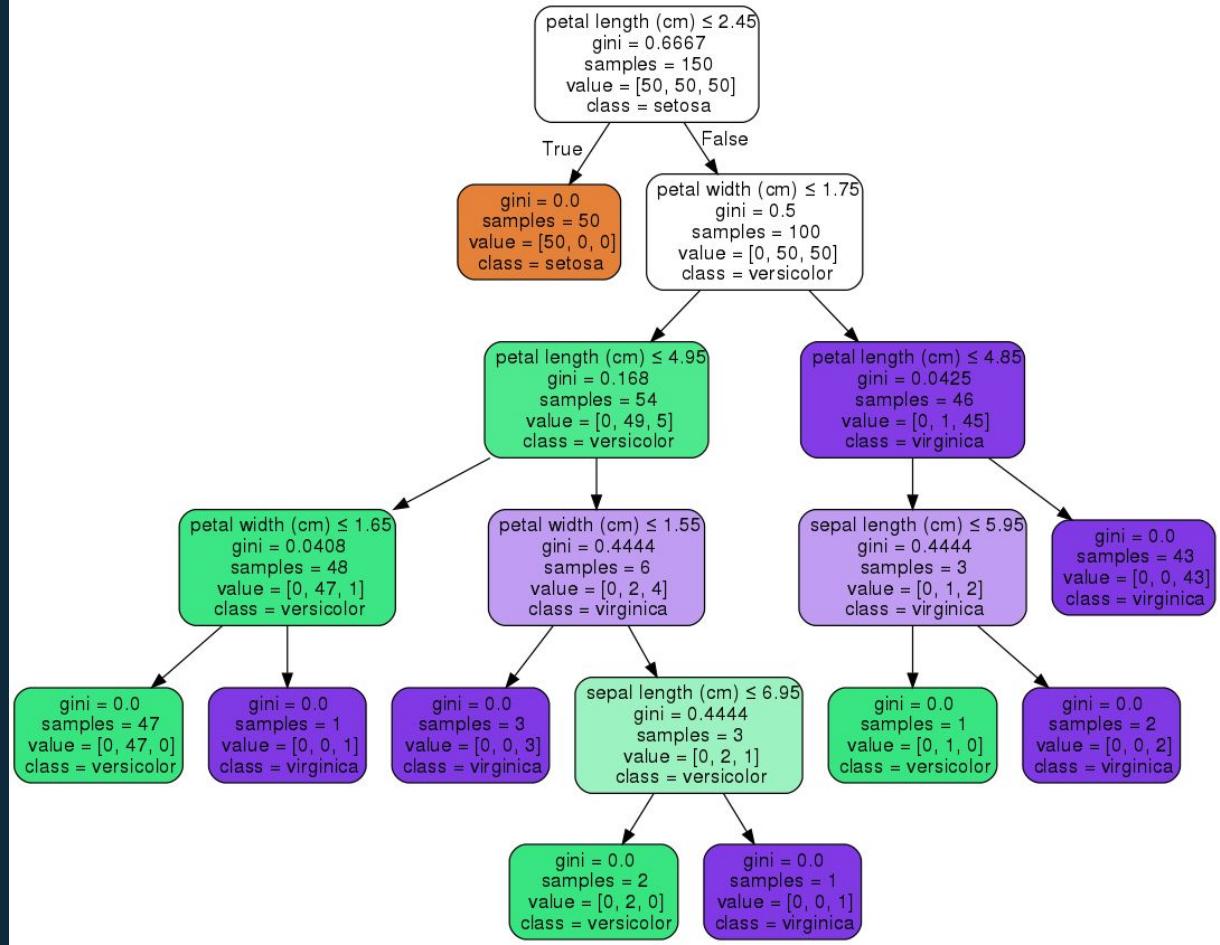
- 1) Escolher um atributo
- 2) Estender a árvore adicionando um ramo para cada valor do atributo
- 3) Filtrar as amostras de acordo com o valor do atributo e enviar as amostras para a folha
- 4) Para cada folha:
 - a) Se as amostras forem da mesma classe, associar essa classe à folha.
 - b) Senão, repetir os passos de 1 até 4

Indução da árvore



Árvore de Decisão

Exemplo de uma árvore
de decisão para o
problema de classificação
de flores



Considerações

- ◊ GI tem um bias que favorece a escolha de atributos com muitos valores;
- ◊ Para minimizar o *overfitting* deve-se:
 - Aplicar procedimentos de poda
 - Definir bem os hiperparâmetros
 - Selecionar atributos a priori, etc.



Random Forest

- ◆ “Floresta Aleatória (Random Forest) é um algoritmo de aprendizagem de máquina flexível e fácil de usar que produz excelentes resultados a maioria das vezes, mesmo sem ajuste de hiperparâmetros. É também um dos algoritmos mais utilizados, devido à sua simplicidade e o fato de que pode ser utilizado para tarefas de classificação e também de regressão.”¹

[1] - <https://medium.com/machina-sapiens/o-algoritmo-da-floresta-aleat%C3%B3ria-3545f6babdf8>



Random Forest

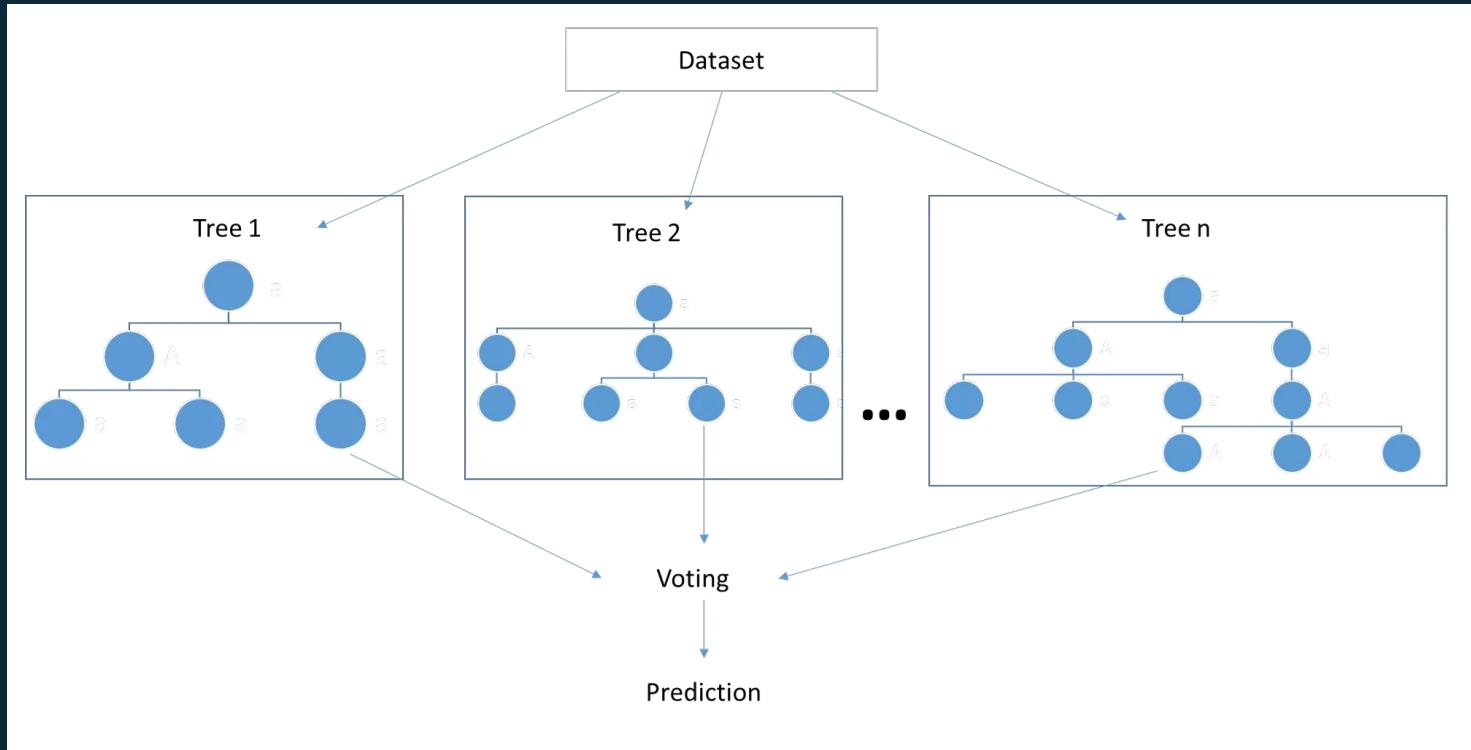
- ◊ A “floresta” criada é uma combinação (ensemble) de árvores de decisão
- ◊ Treinados com o método de bagging, (amostras diferentes da base de dados que são usadas para aprender hipóteses diferentes)
- ◊ Busca a melhor característica em um subconjunto aleatório de todas as características



Random Forest

- ◊ A previsão final para um exemplo de teste é a média da previsão de cada hipótese
- ◊ Cria diversidade, o que geralmente leva a geração de modelos melhores.
- ◊ Muito bom para se medir a importância relativa de cada característica (feature) para a predição

Random Forest



Considerações

◆ Vantagens:

- Poder ser utilizado tanto para regressão quanto para classificação
- É fácil visualizar a importância relativa que ele atribui para cada característica na suas entradas
- O número de hiperparâmetros não é tão grande e são fáceis de serem compreendidos.
- Diminui o overfitting se comparado a árvore de decisão

Considerações

◆ Desvantagens:

- Uma quantidade grande de árvores pode tornar o algoritmo lento e ineficiente para previsões em tempo real.
- Muito lento para fazer previsões depois de treinados (São rápidos para treinar)
- Uma previsão com mais acurácia requer mais árvores, o que faz o modelo ficar mais lento



Métricas de avaliação



Matriz de confusão

		Valor Observado (valor verdadeiro)	
		Label. Pos. ($Y=1$)	Label. Neg. ($Y=0$)
Valor Preditivo	Pred. Pos. ($Y=1$)	VP (verdadeiro positivo)	FP (falso positivo)
	Pred. Neg. ($Y=0$)	FN (falso negativo)	VN (verdadeiro negativo)

Exemplo

Após executar um classificador, que classifica os clientes da Cartoes&CIA entre bons e maus pagadores, sobre 100 amostras (55 como bons pagadores e 45 como maus pagadores), nós obtivemos o seguinte resultado:

Dos 55 **bons pagadores** apenas 40 foram preditos corretamente, e dos 45 **maus pagadores** apenas 35 foram preditos corretamente. Qual a matriz de confusão?

Exemplo

		Valor Real	
		$Y=BP=1$	$Y=MP=0$
Valor Preditivo	$Y'=BP=1$	40	10
	$Y'=MP=0$	15	35



Acurácia

- ◊ A proporção de previsões corretas, sem levar em consideração o que é positivo e o que é negativo. .
- ◊
$$\text{ACC} = \frac{\text{Total de Acertos}}{\text{Total de dados no conjunto}}$$
$$= \frac{(\text{VP} + \text{VN})}{(\text{VP} + \text{FP} + \text{VN} + \text{FN})}$$

	Label. Pos.	Label. Neg.
Pred. Pos.	VP	FP
Pred. Neg.	FN	VN

Precisão

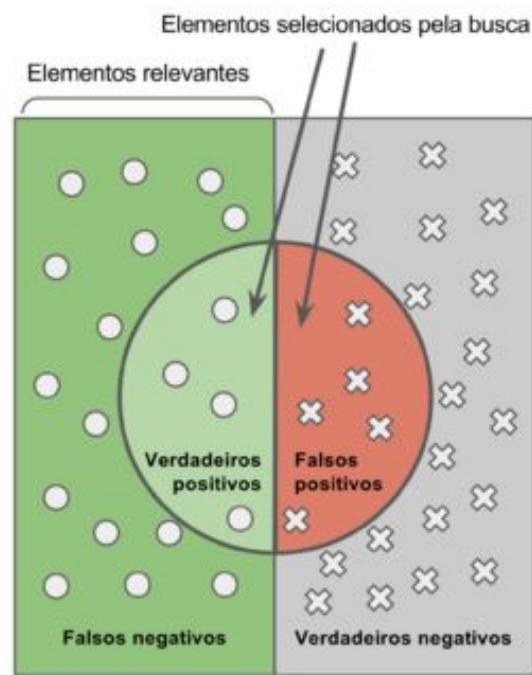
- ◊ É a fração de instâncias **recuperadas** que são relevantes.
- ◊ $\text{PREC} = \text{ACERTOS POSITIVOS} / \text{TOTAL DE ACERTOS PREDITOS COMO POSITIVOS}$
- ◊ $\text{PREC} = \text{VP} / (\text{VP} + \text{FP})$

	Label. Pos.	Label. Neg.
Pred. Pos.	VP	FP
Pred. Neg.	FN	VN



Precisão = 20/30

	Label. Pos. (Y=1)	Label. Neg. (Y=0)
Pred. Pos. (Y'=1)	20	10
Pred. Neg. (Y'=0)	40	30



$$\text{Precisão} = \frac{\text{Verdadeiros positivos}}{\text{Elementos selecionados}}$$

"Quantos elementos selecionados são relevantes?"

$$\text{Revocação} = \frac{\text{Verdadeiros positivos}}{\text{Elementos relevantes}}$$

"Quantos elementos relevantes foram selecionados?"

Recall (Sensibilidade)

- ◆ A capacidade do sistema em predizer corretamente a condição para casos que realmente a têm (é a proporção de verdadeiros positivos). É a fração de instâncias relevantes que são recuperadas
- ◆ Também conhecida como sensibilidade, revocação ou true positive rate (TPR)
- ◆ REC = ACERTOS POSITIVOS / TOTAL DE POSITIVOS

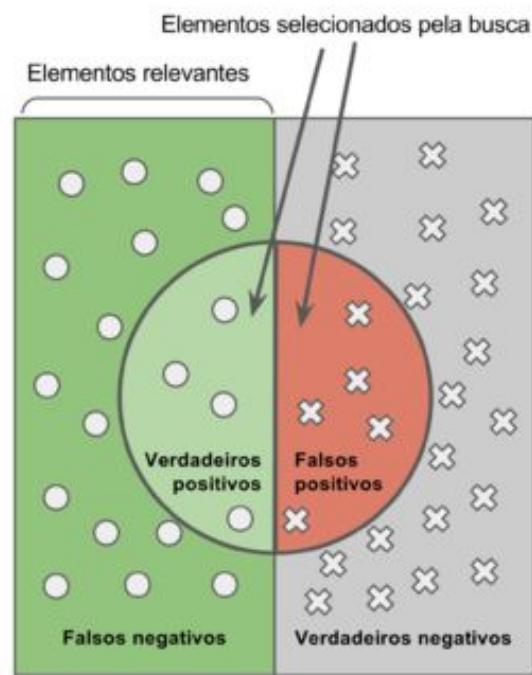
$$= VP / (VP + FN)$$

	Label. Pos.	Label. Neg.
Pred. Pos.	VP	FP
Pred. Neg.	FN	VN



$$\text{Recall} = 20/60$$

	Label. Pos. (Y=1)	Label. Neg. (Y=0)
Pred. Pos. (Y'=1)	20	10
Pred. Neg. (Y'=0)	40	30



F1-Score

- ◊ Essa métrica combina precisão e recall de modo a trazer um valor único que indique a qualidade geral do modelo
- ◊ Trabalha bem mesmo com conjuntos de dados que possuem classes desproporcionais.
- ◊
$$\text{F1-Score} = \frac{2 * \text{PRECISAO} * \text{RECALL}}{\text{PRECISAO} + \text{RECALL}}$$

Especificidade

- ◊ A proporção de verdadeiros negativos: a capacidade do sistema em predizer corretamente a ausência da condição para casos que realmente não a têm.
- ◊ Também conhecida como true negative rate (TNR)
- ◊ $SPEC = ACERTOS\ NEGATIVOS / TOTAL\ DE\ NEGATIVOS$

$$= VN / (VN + FP)$$

	Label. Pos.	Label. Neg.
Pred. Pos.	VP	FP
Pred. Neg.	FN	VN

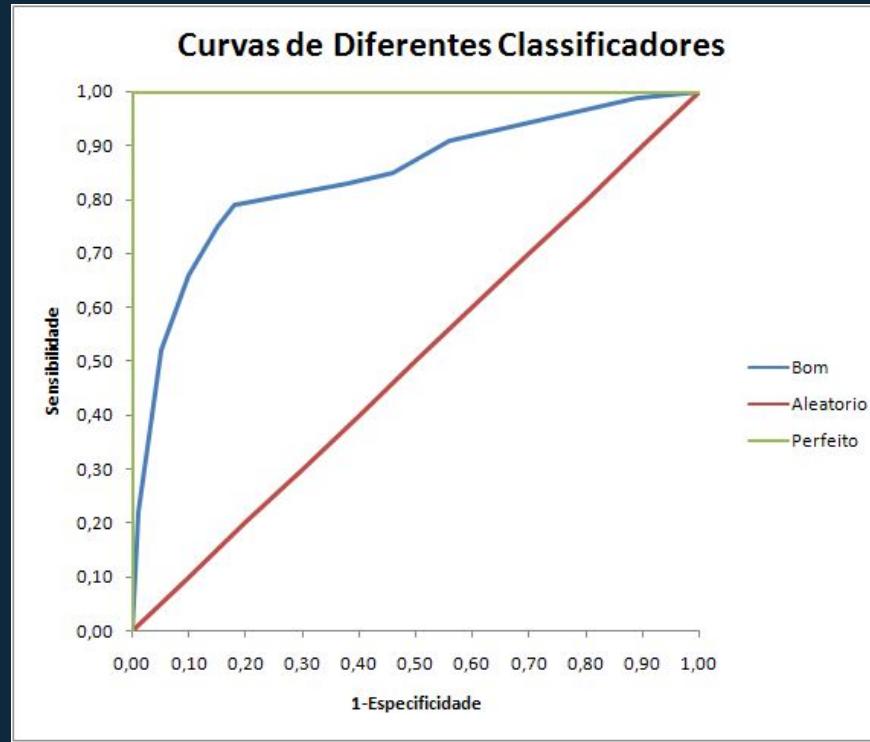
Curva ROC

- ◊ Criada por engenheiros elétricos e de sistemas de radar durante a Segunda Guerra Mundial para detectar objetos inimigos em campos de batalha
- ◊ Os algoritmos de classificação produzem um valor situado dentro de um determinado intervalo contínuo, como $[0;1]$, é necessário definir um ponto de corte, ou um limiar de decisão, para se classificar e contabilizar o número de previsões positivas e negativas.

Curva ROC

- ◊ Este limiar pode ser selecionado arbitrariamente, a melhor prática para se comparar o desempenho de diversos sistemas é estudar o efeito de seleção de diversos limiares sobre o resultado das previsões.

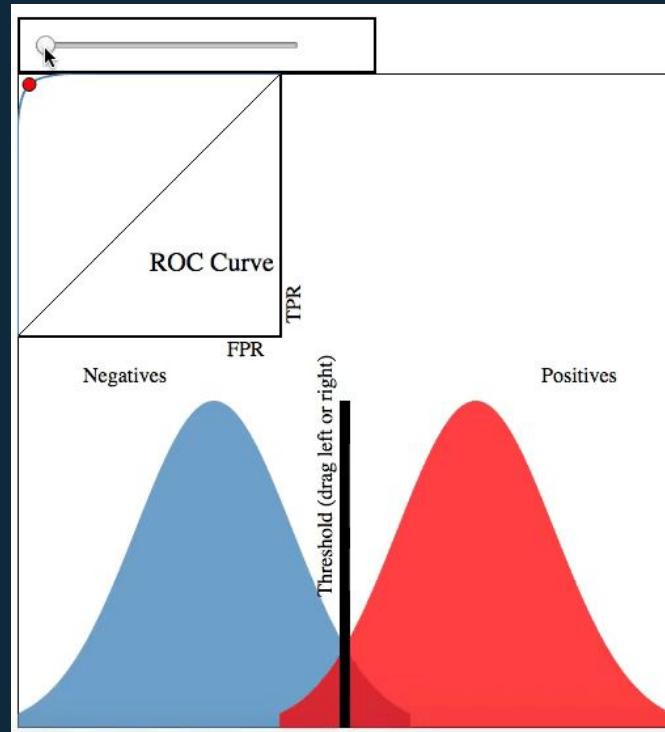
Área Sob Curva ROC





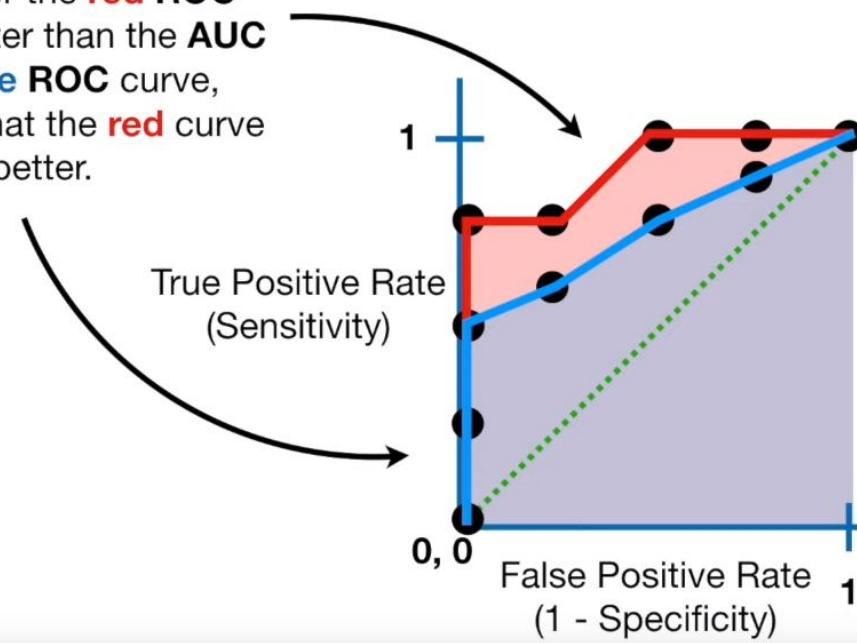
Área Sob Curva ROC

Área Sob Curva ROC

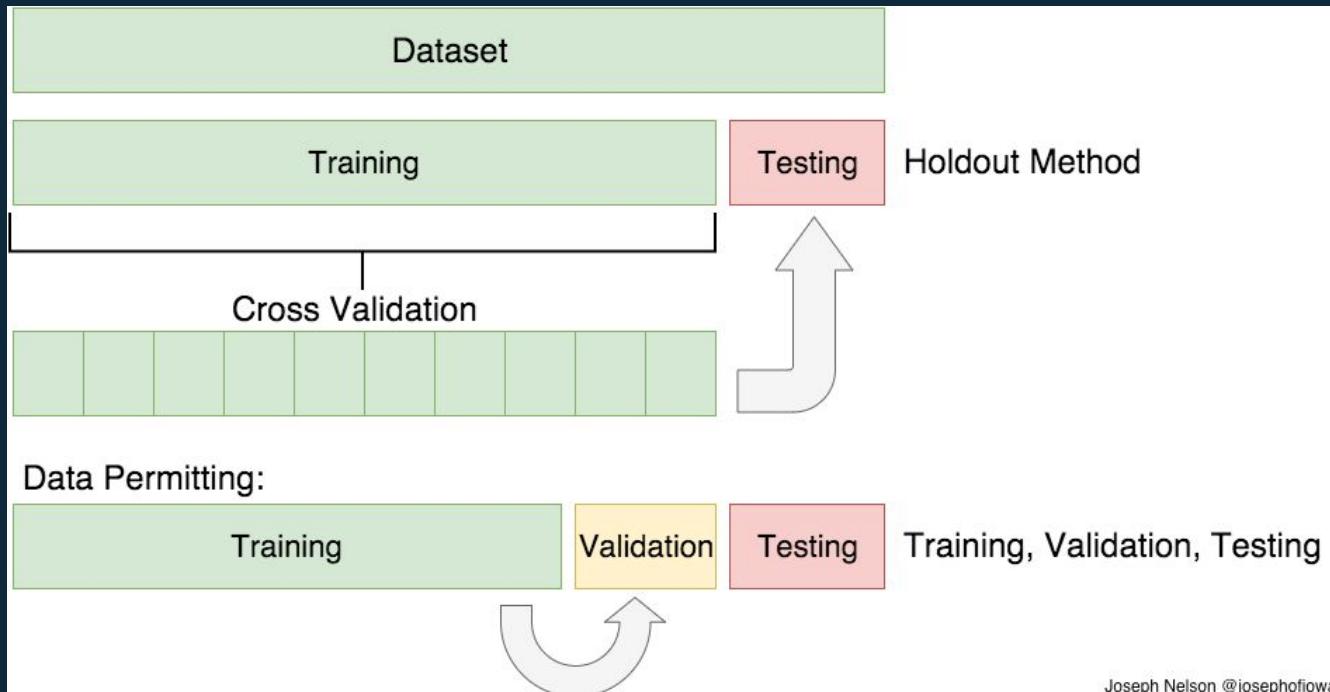


Área Sob Curva ROC

The **AUC** for the **red ROC** curve is greater than the **AUC** for the **blue ROC** curve, suggesting that the **red** curve is better.



Before HandsOn



http://dontpad.com/kdd_uni7

Hands-On

