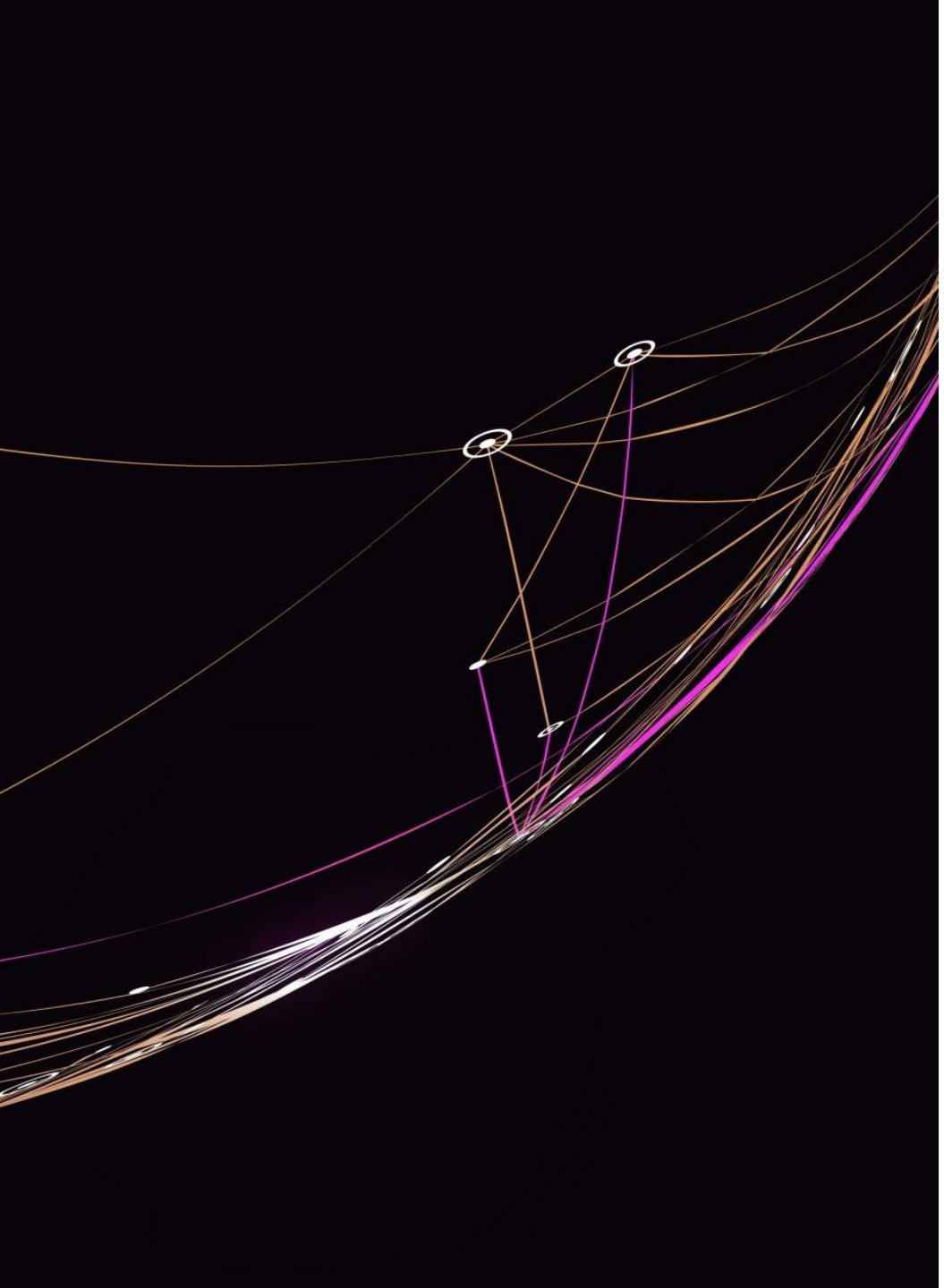


ADMINISTRAÇÃO

IBM0112 DATA MINING

Avaliação de modelos

Cassius Figueiredo

Princípios de avaliação de modelos

Dos objetivos às métricas

- A partir da(s) pergunta(s) que define(m) o problema como sabemos se alcançamos os objetivos estabelecidos?
- Tanto a inferência quanto a predição dependerão de métricas que farão a avaliação de quão bem o modelo responde ao objetivo.
- Vamos ver a partir de agora como avaliar se nosso modelo está cumprindo ou não seus objetivos.

Avaliação de modelos

- Duas questões sempre relevantes quando criamos modelos supervisionados são a variância e o overfitting, conceitos profundamente relacionados.
- Podemos entender a questão da variância a partir do entendimento do chamado trade-off viés-variância.
- Já o overfitting será discutido a partir de gráficos.

Trade-off viés-variância

- Quanto mais complexo o modelo menor o viés e maior a variância.
- Quanto menos complexo o modelo maior o viés e menor a variância.
- Quanto mais observações, menor a variância.
- Modelos mais complexos costumam apresentar maior variância para um mesmo número de observações.

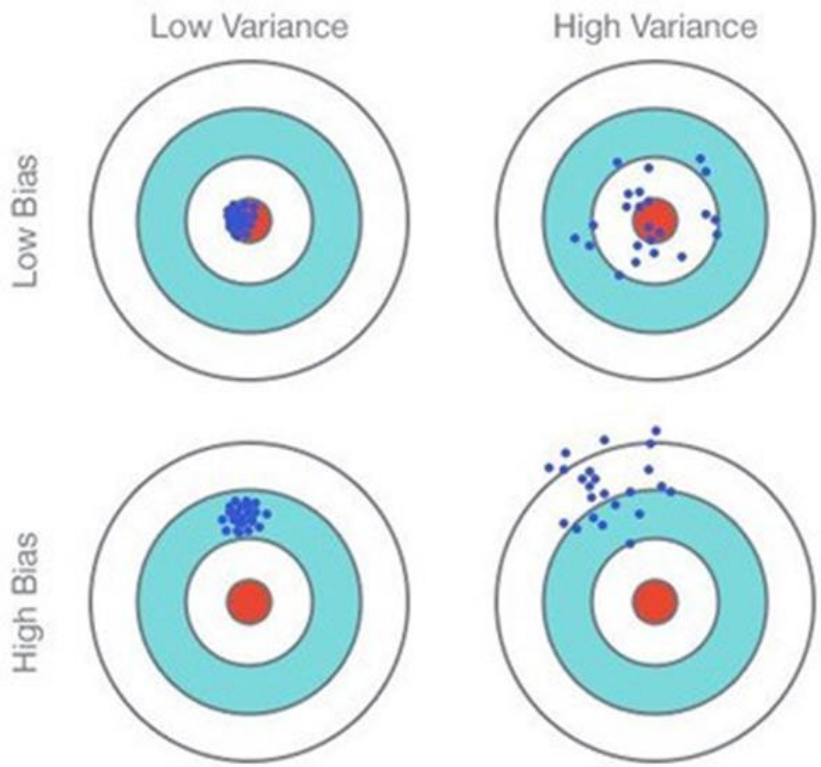


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

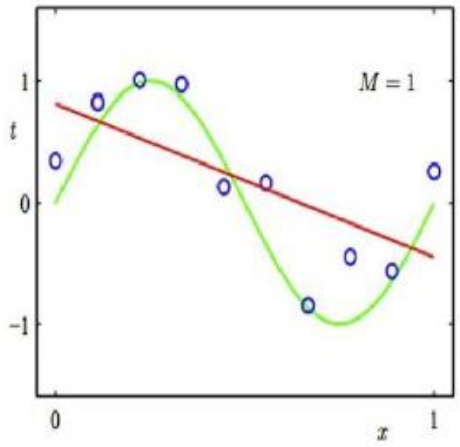
Imagem: Analytics Vidhya

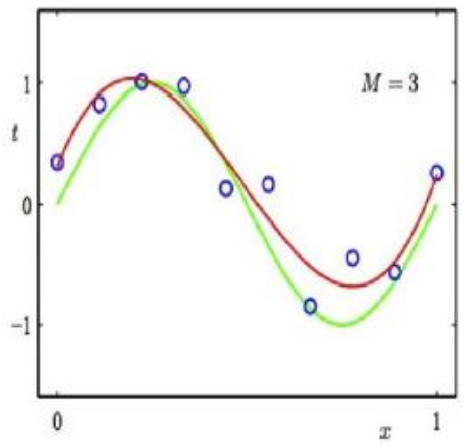
Overfitting

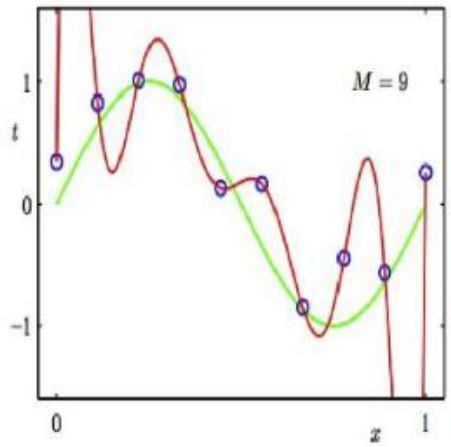
- Ocorre quando uma redução no erro do treinamento não corresponde a uma redução equivalente no erro fora da amostra.
- Normalmente causado pelo ruído, dependendo da complexidade do modelo e o número de observações disponíveis
  - Menos observações comportam modelos mais simples para um nível fixo de ruído.
- O overfitting pode ser combatido, por exemplo, através da aplicação de técnicas de sub-sampling no momento da modelagem e validação.
  - Treino – Validação – Teste
  - Validação Cruzada

Overfitting

Regression:



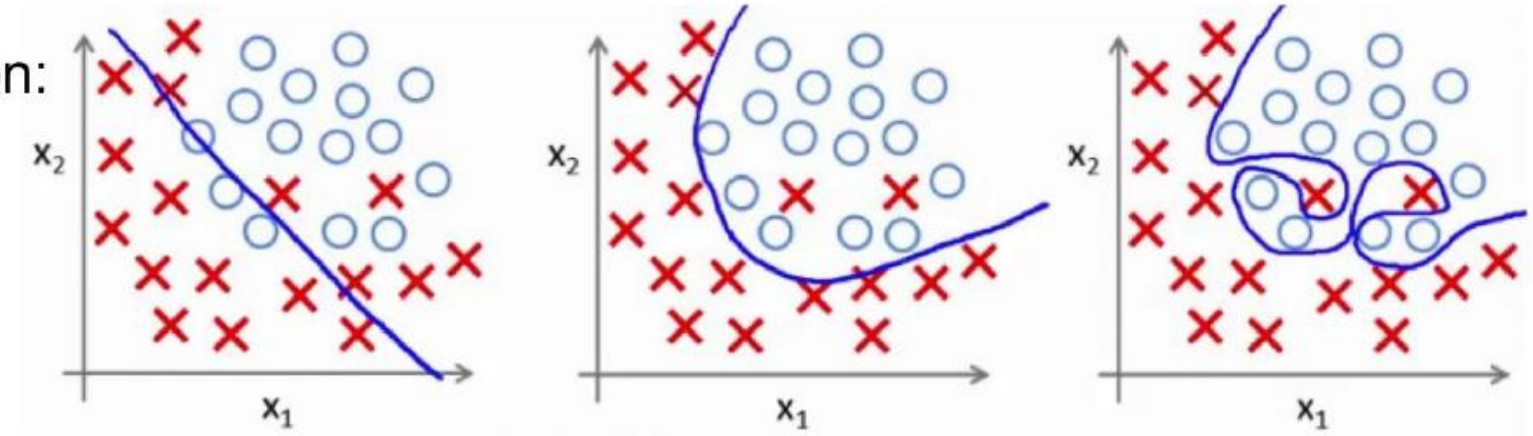




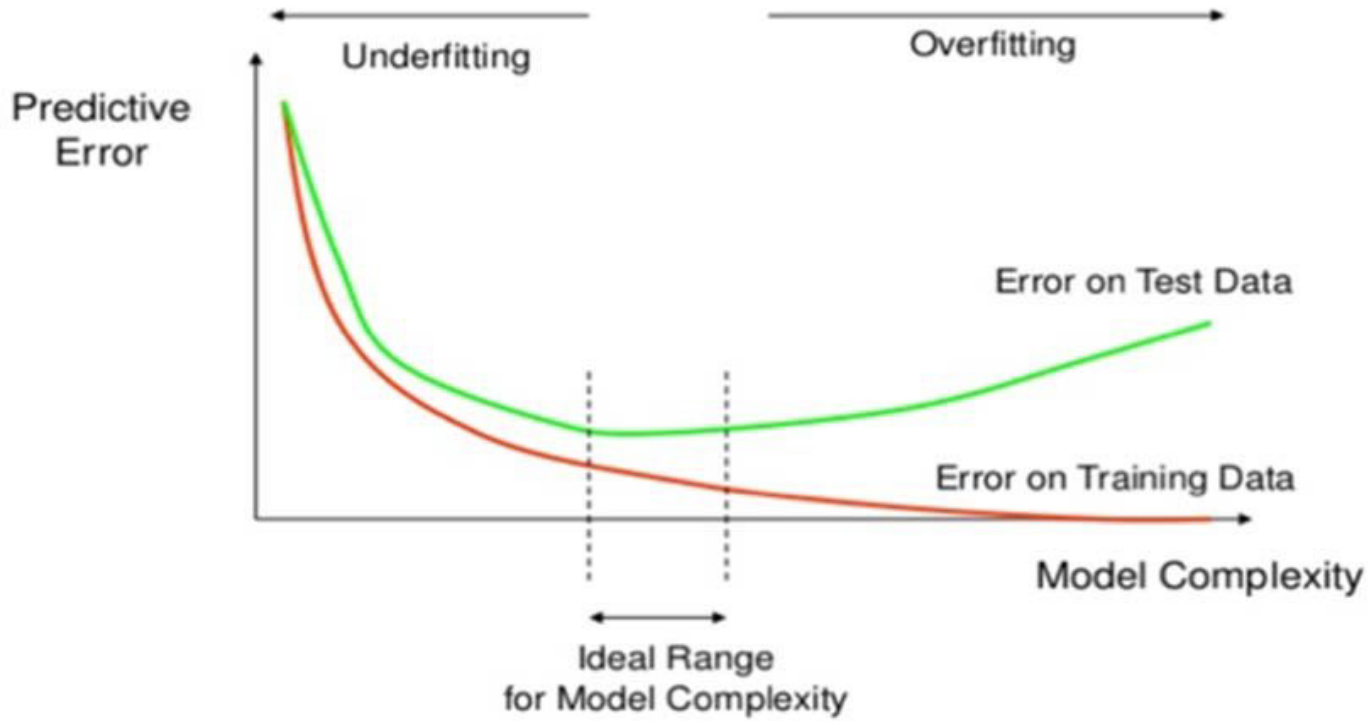
predictor too inflexible:
cannot capture pattern

predictor too flexible:
fits noise in the data

Classification:



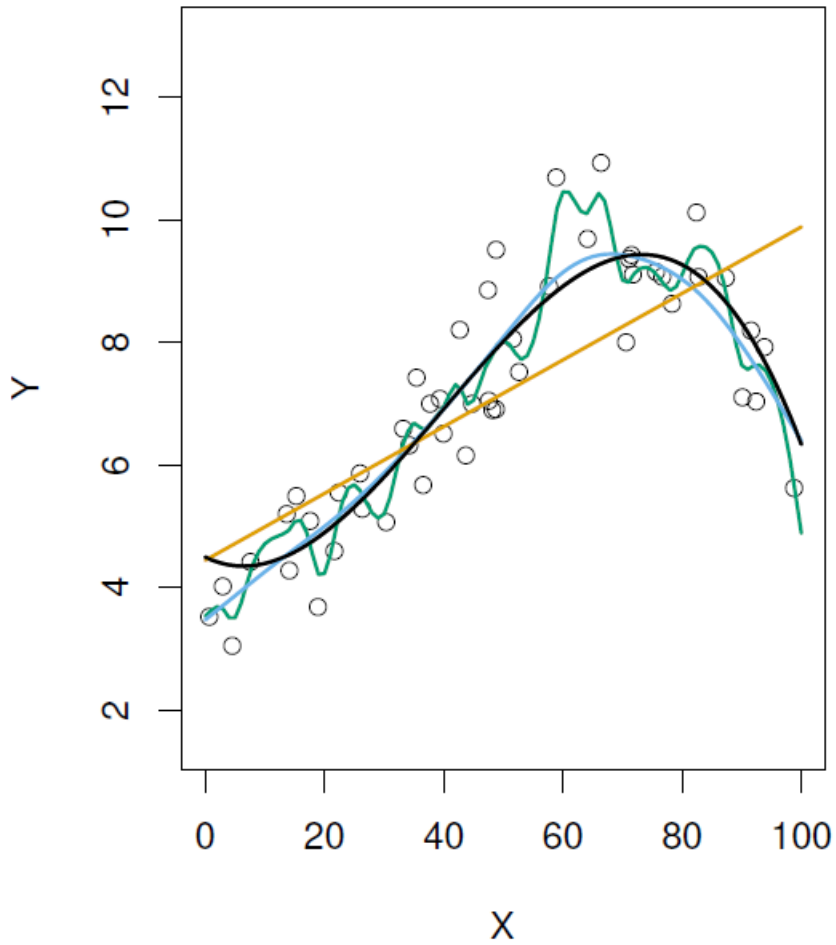
Overfitting

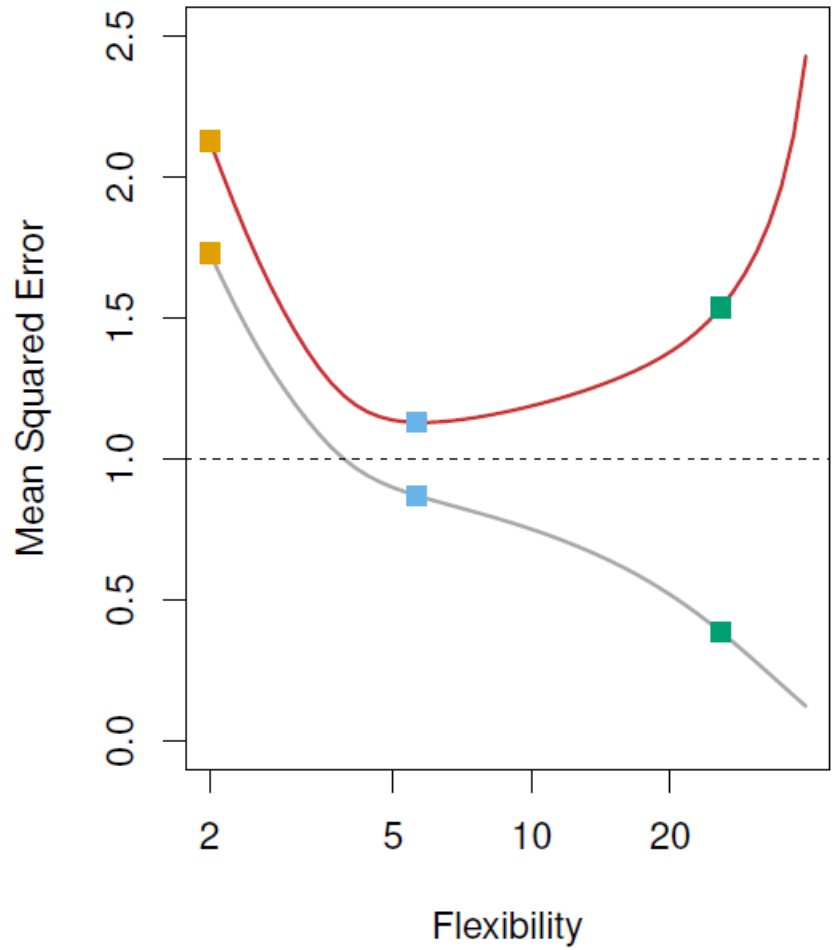


Overfitting

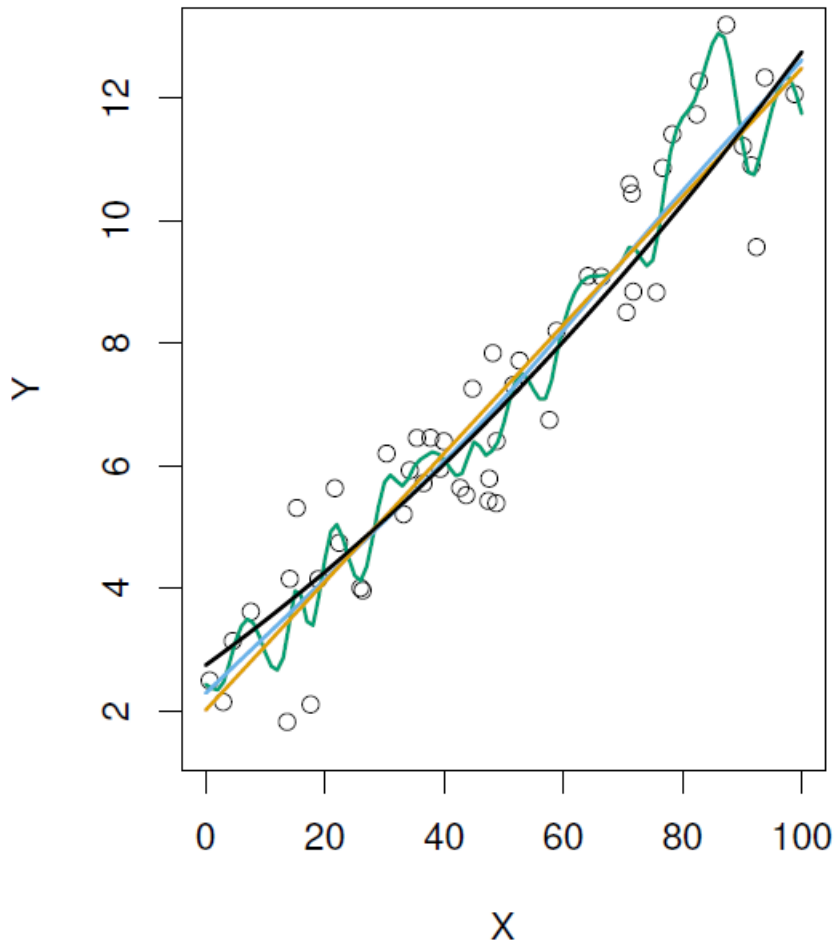
- Vamos discutir os gráficos levando em consideração o as informações ruído, erros dentro e fora da amostra, e complexidade do modelo.
  - Linha preta corresponde a função objetivo
  - Linha amarela corresponde a um modelo linear.
  - Linha azul e verde correspondem a modelos com complexidades diferentes.
- Os exemplos possuem funções objetivo de ordens (complexidades) diferentes (Flexibility).
- A linha vermelha é o erro fora da amostra e a linha cinza é o erro dentro da amostra.
- Estas curvas representam a aplicação em uma regressão.

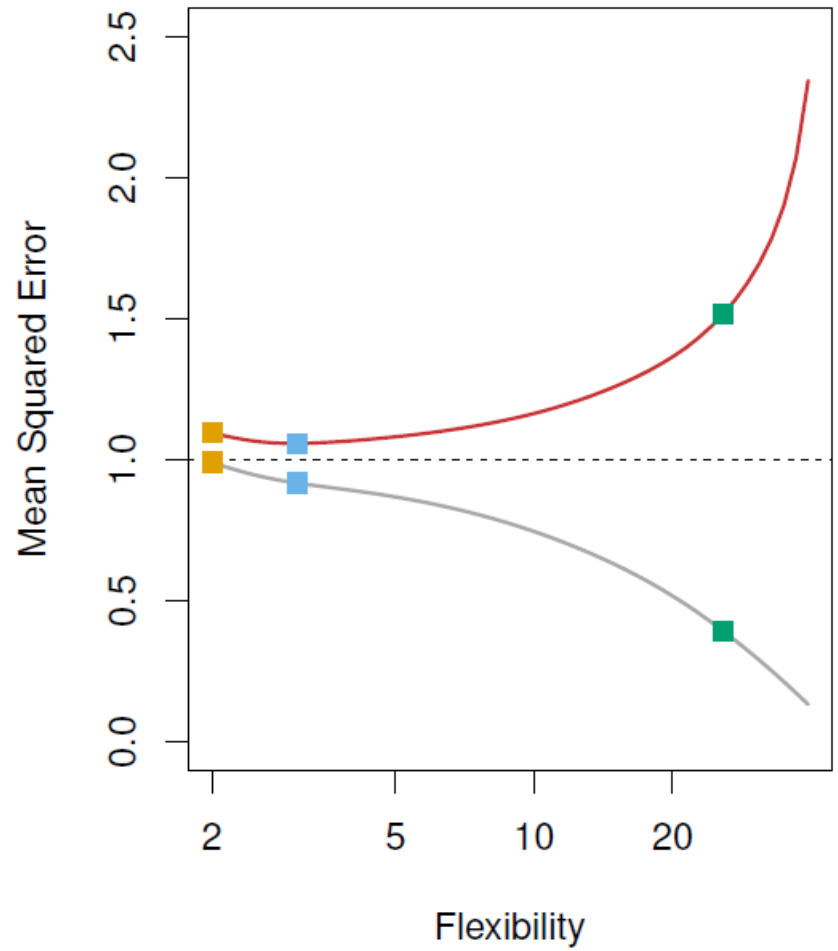
Cenário 1


Y
X
2
4
6
8
10
12
0
20
40
60
80
100

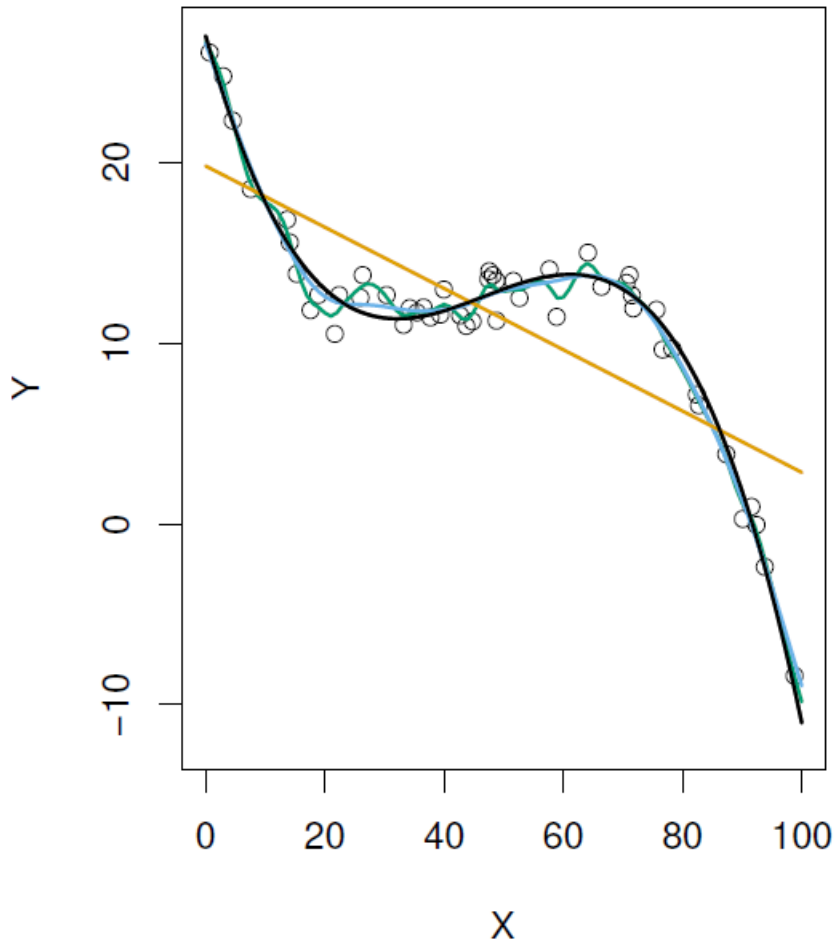

Mean Squared Error
Flexibility
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20

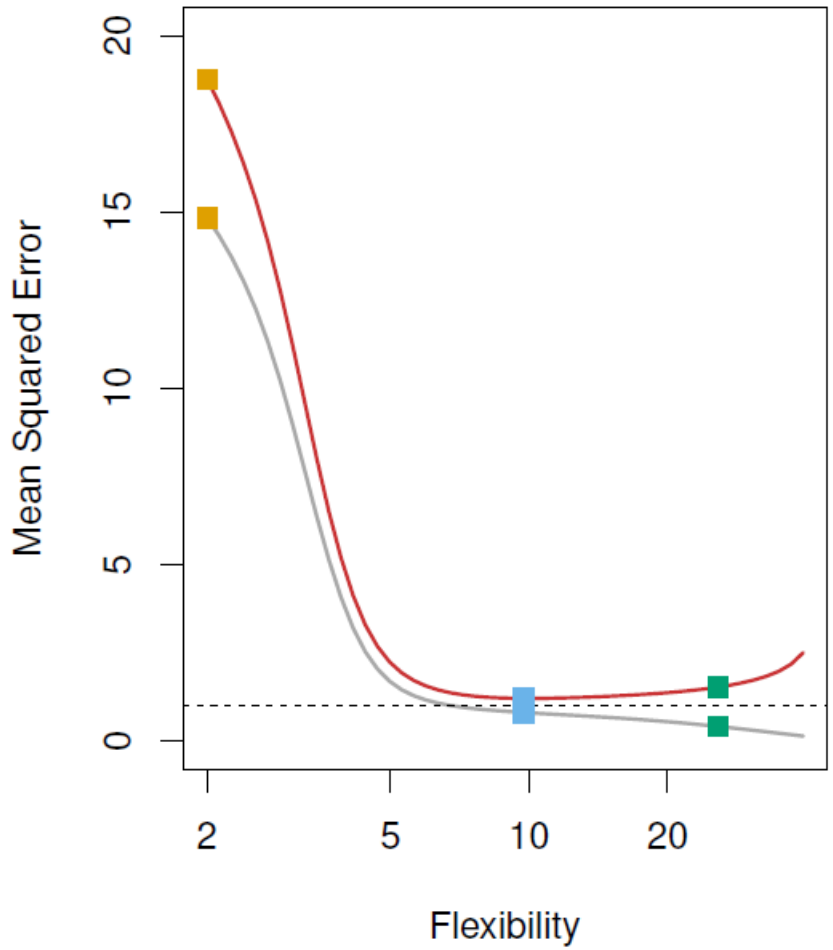
Cenário 2


Y
X
2
4
6
8
10
12
0
20
40
60
80
100


Mean Squared Error
Flexibility
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20

Cenário 3


Y
20
10
0
−10
0
20
40
60
80
100
X


Mean Squared Error
0
5
10
15
20
2
5
10
20
Flexibility

Para não esquecer...

Abordagem Treino-Validação-Teste

- Técnica de minimização de overfitting.
- Envolve a divisão do conjunto de dados que será utilizado para treinar o modelo em dois ou três grupos distintos:
  - Treino - Teste
  - Treino - Validação - Teste
- Utilizamos então o conjunto de Treino para treinar o modelo, o conjunto de Validação para validar o treino e o conjunto de Teste para a verificação final de performance.
- A partir destes conjuntos definimos erro dentro da amostra (Treino) e erro fora da amostra (Teste).

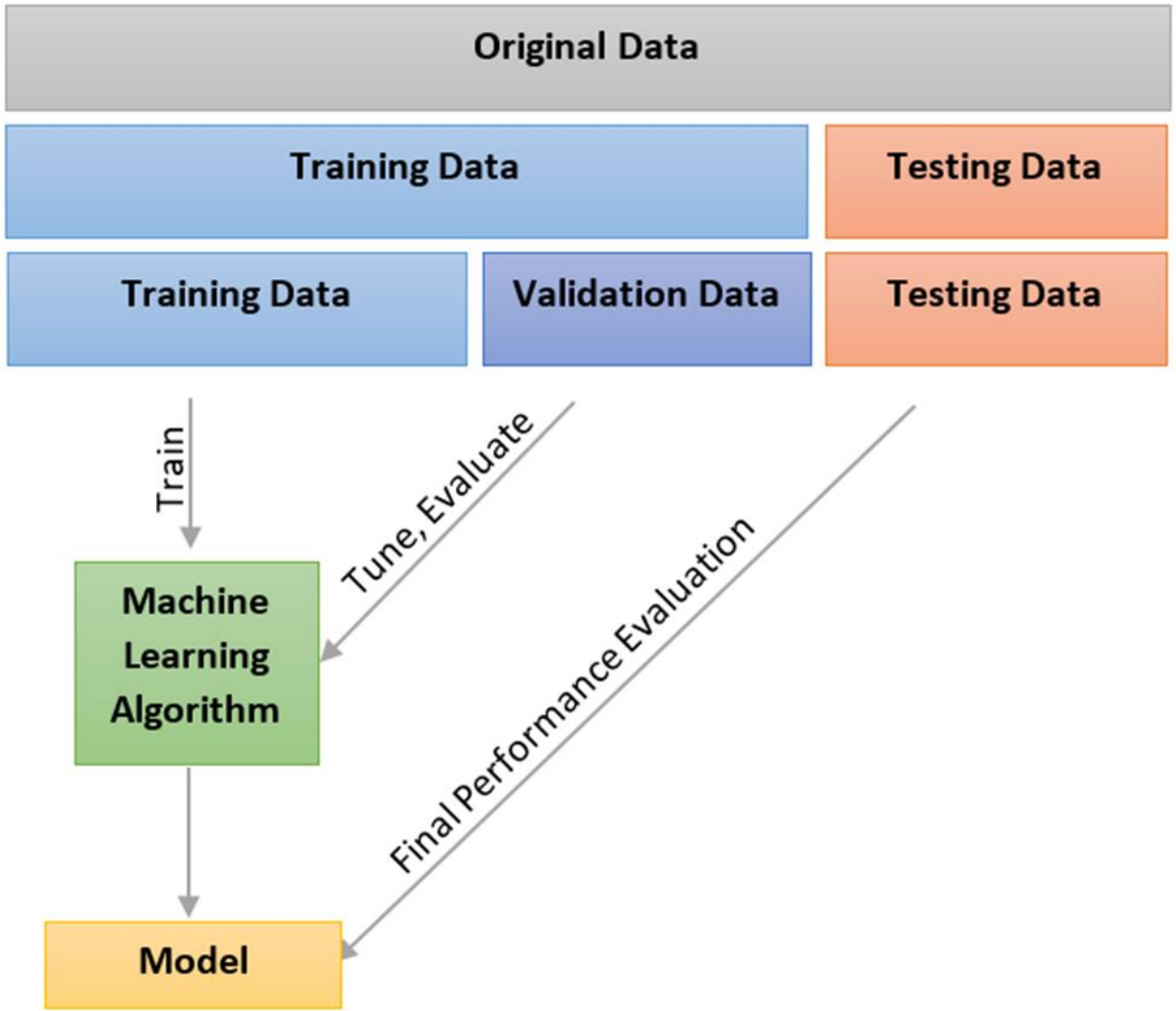
Abordagem Treino-Validação-Teste

- Divisões usuais são:
  - 80-20
  - 70-30
  - 60-40 (mais rara)
- Quando optamos pelo conjunto de Validação, este será escolhido da porção de Treino.

Abordagem Treino-Validação-Teste

- Pró
  - Fácil e rápido.
- Contras
  - Uma parte dos seus dados jamais será usado para treinar seu modelo.
  - Variedades pequenas da população podem ser sub-representadas.

Abordagem Treino-Validação-Teste

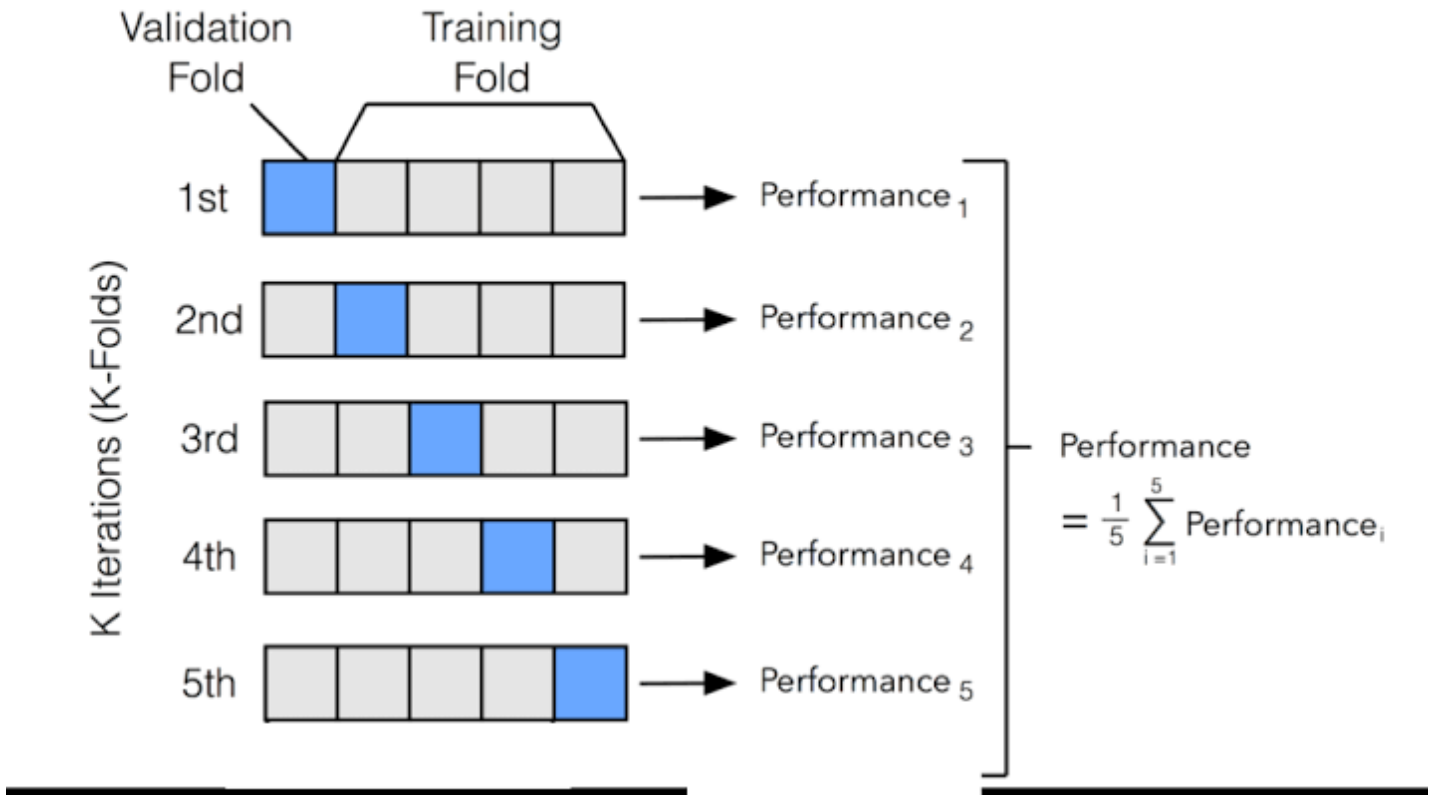


Problemas na geração dos conjuntos

- Escolha viesada
  - Geração aleatória, partindo do pressuposto que os dados são IID (Independently and Identically Distributed).
- Conjuntos desbalanceados
  - Undersampling
  - Oversampling
  - SMOTE, etc.
- Séries temporais
  - As divisões do conjunto de dados devem respeitar a ordem temporal para evitar o chamado lookahead bias.
- Contaminação Treino-Teste (Leakage)

Validação Cruzada (Cross-Validation)

- Minimiza a variância.
- Técnica que pode minimizar os problemas relacionados à subdivisão de bases pequenas.
- A principal abordagem chama-se "K-fold Cross Validation".



Validação Cruzada (Cross-Validation)

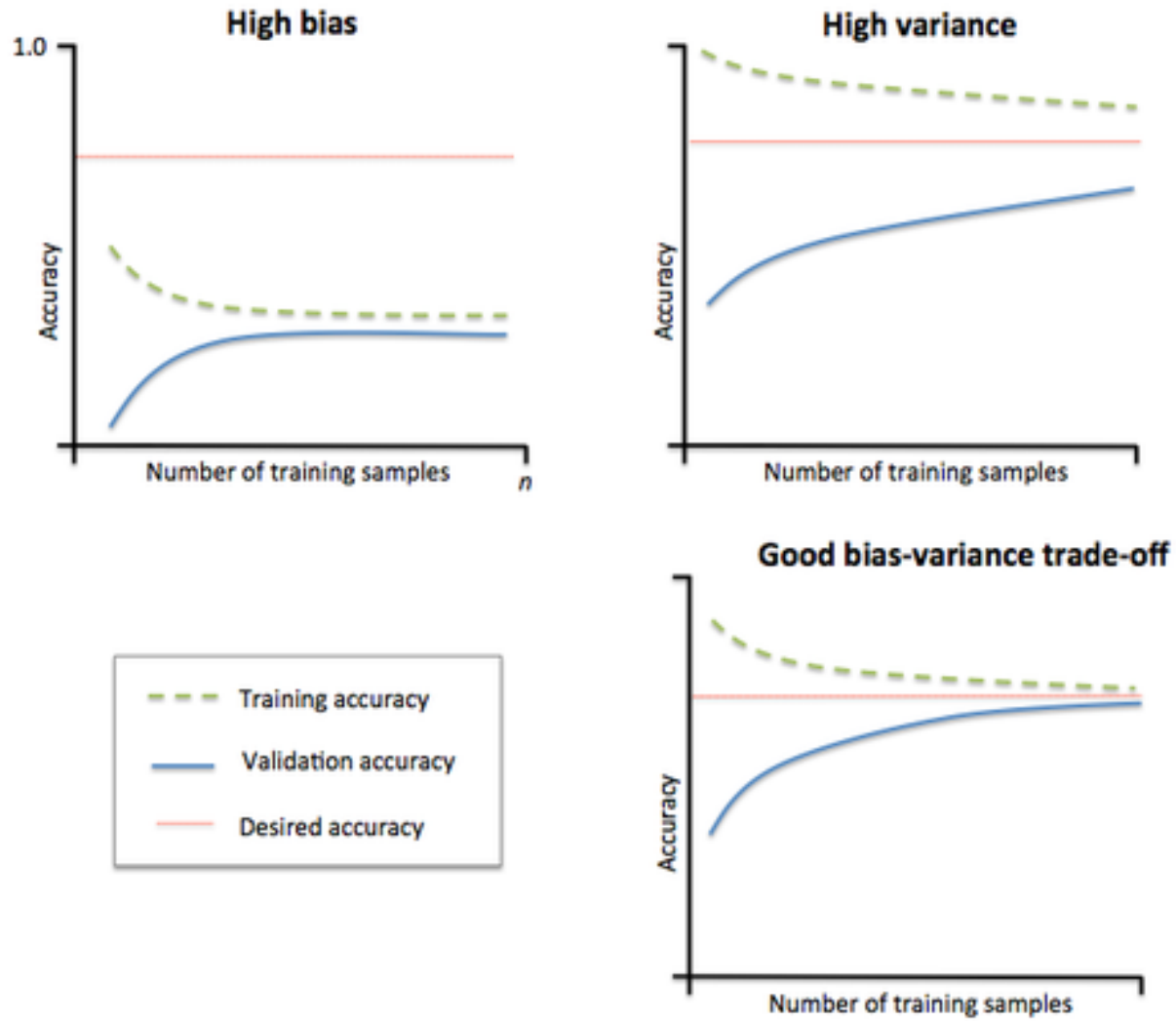
- Prós
  - Fácil de reproduzir.
  - Todos os dados participam tanto do treino quanto do teste, em momentos diferentes.
- Contra
  - Custo computacional: Envolve a criação de k (geralmente k=5 ou k=10) modelos.

Curvas de aprendizagem

- Apresenta a evolução dos erros nos conjuntos de Treino e Teste contra o tamanho do conjunto de dados.
- Muito usadas em problemas de classificação.
- Ajuda a identificar o trade-off viés-variância para determinado modelo e também ajuda a responder se o aumento de observações pode ajudar a melhorar a performance do modelo.

- Normalmente utilizamos a medida de erro no eixo Y (em alguns casos podemos usar métricas como acurácia, precision ou recall).
- Já no eixo X, depende do cenário em estudo:
  - Regressores – Complexidade do modelo (grau)
  - Classificadores tradicionais – Tamanho da amostra
  - Deep Learning – Número de epochs

Curvas de aprendizagem – Bias vs. Variance



Métricas de classificação

Classificação

- O objetivo da classificação é prever uma variável discreta, que representa um grupo, categoria ou classe.
- A avaliação de performance de classificadores envolve conceitos mais abrangentes, que veremos a partir de agora.

- A maioria dos classificadores entregará, na verdade, um valor contínuo que será classificado de acordo com um critério de separação de classes.
- No caso de um classificador binário (duas classes), por exemplo, o score gerado pelo classificador será um valor entre zero e um. A classificação final será definida pelo critério de corte (ex.: $\geq 0{,}5 \rightarrow 1; < 0{,}5 \rightarrow 0$).

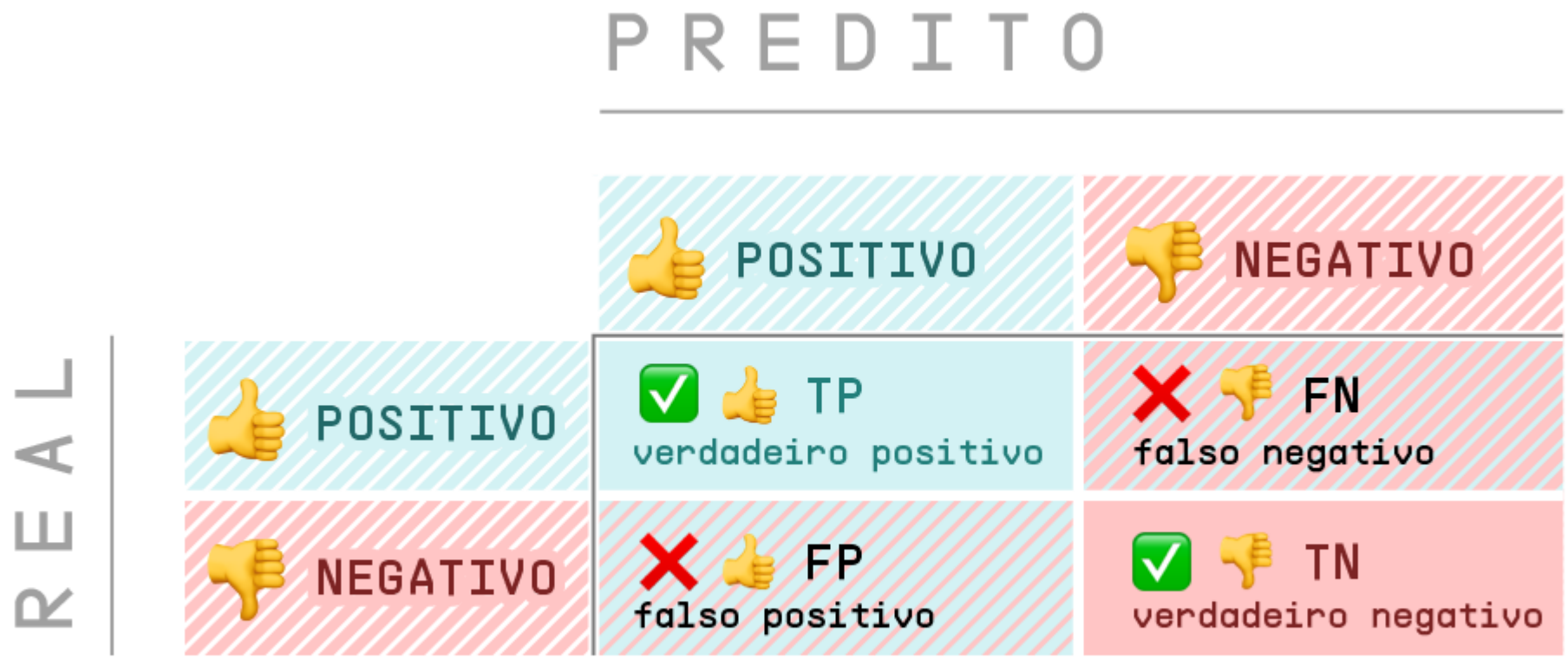
Métricas

- Ao compararmos o resultado de uma classificação com a variável-alvo pré-existente, podemos ter 4 resultados possíveis:
  - **True Positives (TP)**: real e predito são verdadeiros (boa acurácia).
  - **True Negatives (TN)**: real e predito são falsos (boa acurácia).
  - **False Positives (FP)**: predito verdadeiro e real falso, conhecido por Erro do Tipo 1.
  - **False Negatives (FN)**: predito falso e real verdadeiro, conhecido por Erro do Tipo 2.

Métricas

- Todas as métricas de avaliação de classificadores são calculadas a partir da quebra das predições entre os tipos possíveis.
- O resultado da aplicação da técnica é uma matriz que quantifica cada caso, chamada conhecida como **Matriz de Confusão** (Confusion Matrix).
- Esta matriz faz a associação entre as classes reais e preditas.

Matriz de Confusão



Fonte: Métricas de Avaliação em Machine Learning: Classificação (Kunumi Blog)

Matriz de Confusão

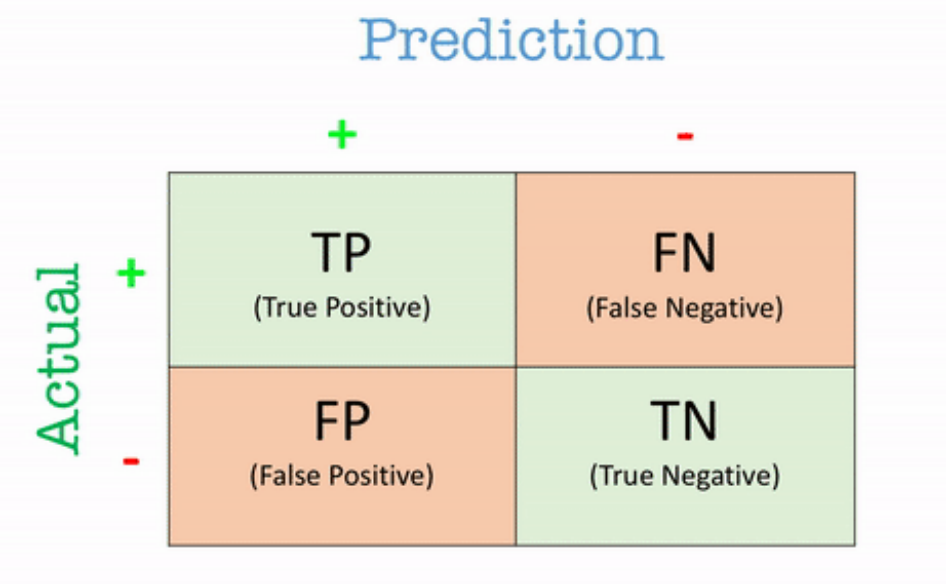
		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	∑
Actual	Iris-setosa	100.0 %	0.0 %	0.0 %	50
	Iris-versicolor	0.0 %	88.7 %	6.4 %	50
	Iris-virginica	0.0 %	11.3 %	93.6 %	50
	∑	50	53	47	150

Acurácia (ACC – Accuracy)

- É a medida de quão bom é o modelo em sua capacidade total de predição.
- Quanto mais próximo de um, melhor a performance do modelo.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

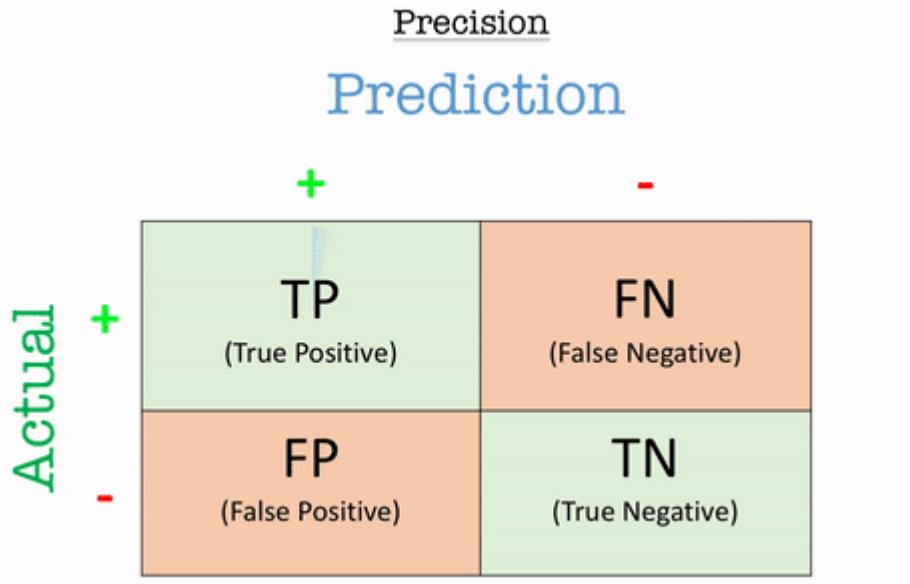


Animações: Amey Naik & Pooja Mehta

Precisão (PRE – Precision)

- Indica quantos dos preditos foram corretos.
- Quanto mais próxima de um, mais acurada é a previsão.
- Responde a pergunta: dos exemplos classificados como positivos, quantos realmente são positivos?

$$Precision = \frac{TP}{TP + FP}$$

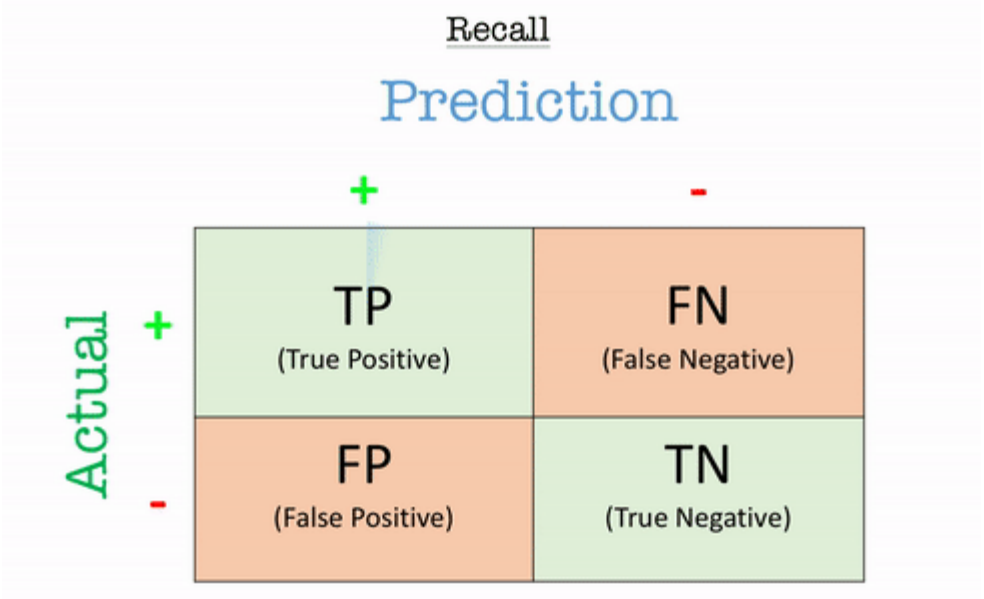


Animações: Amey Naik & Pooja Mehta

Revocação (REC – Recall)

- Também conhecido por TPR (True Positive Rate), Sensibilidade (Sensitivity) ou Hit-Rate.
- Indica o grau de acerto da classe positiva perante todos os positivos.
- Quanto mais próximo de um, melhor.
- Responde a pergunta: de todos os exemplos que são positivos, quantos foram classificados corretamente como positivos?

$$Recall = \frac{TP}{TP + FN}$$



Animações: Amey Naik & Pooja Mehta

F1-Score

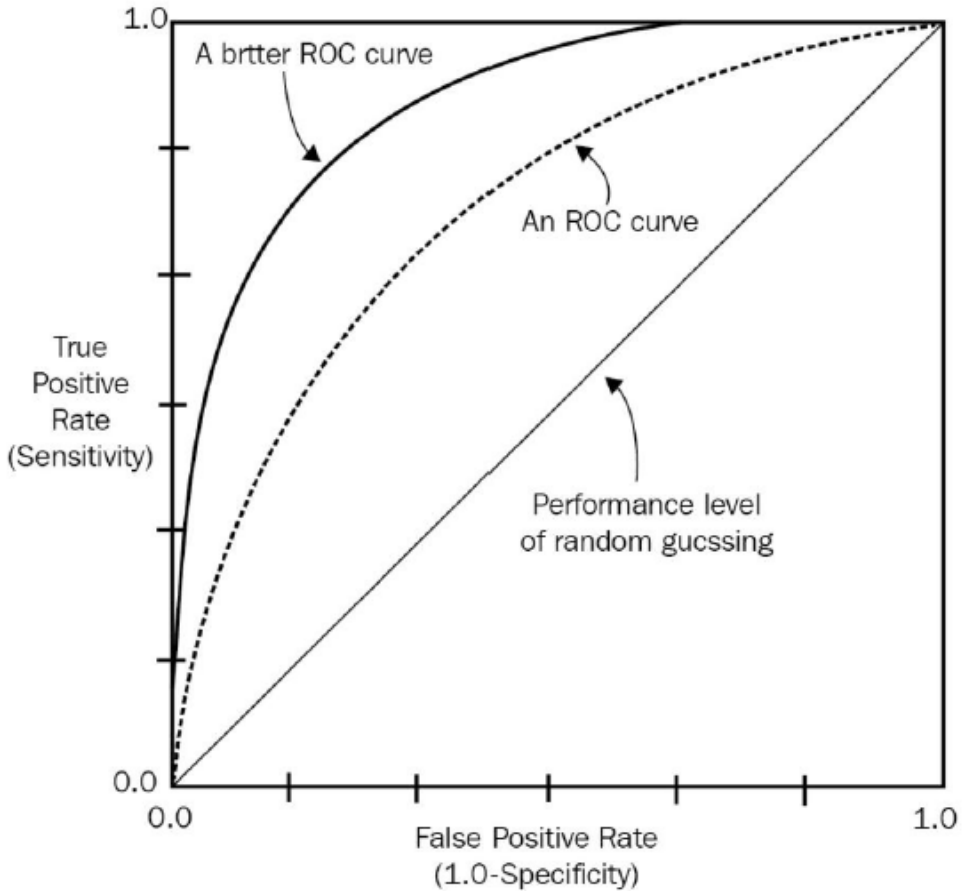
- É uma relação entre Precision e Recall.
- Varia entre zero e um e quanto mais próximo de um, melhor.
- Indicado para problemas com desbalanceamento entre classes.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

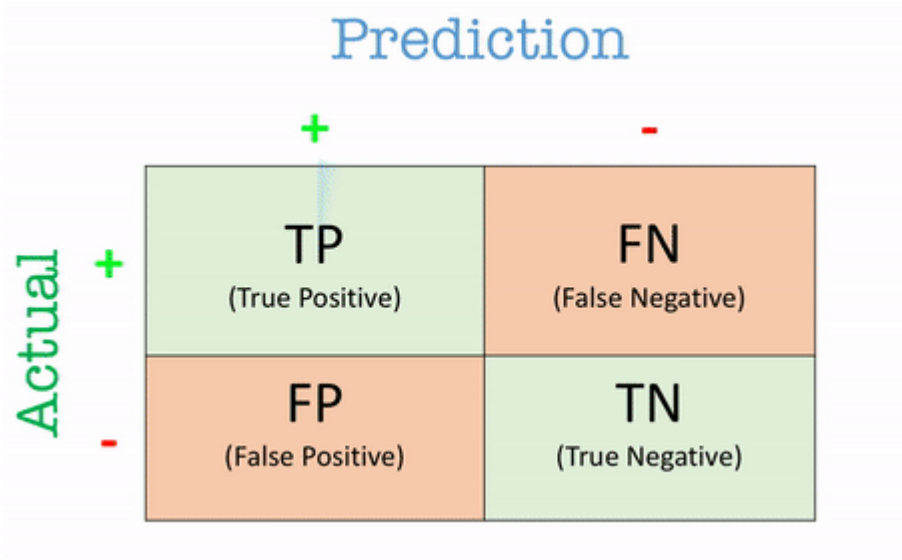
Curva ROC

- Receiver Operating Characteristics (ROC).
- Usada para visualizar, comparar e selecionar classificadores.
- Calcula pares de TPRs e FPRs utilizando todos os scores preditos como pontos de corte.
- Representada em um quadrado de lado unitário.
- A métrica associada é chamada de AUC (Area Under Curve), varia entre zero e um e quanto maior, melhor.

Curva ROC



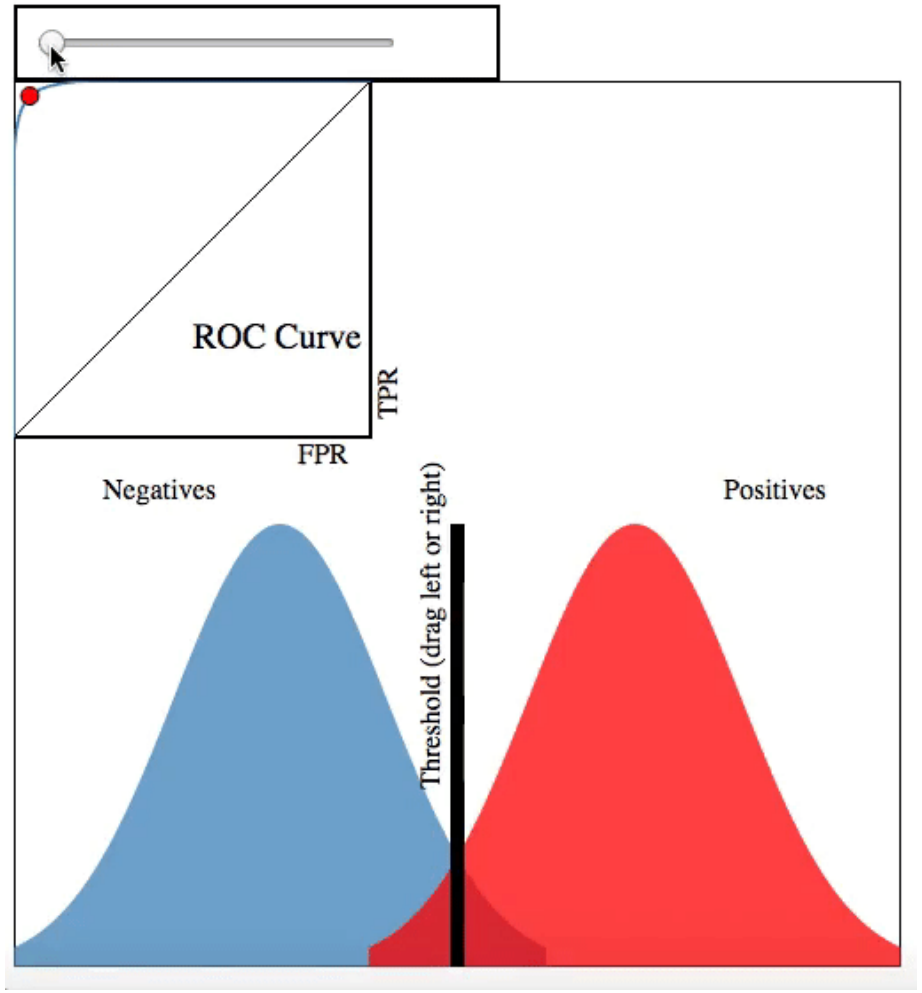
$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$



Animações: Amey Naik & Pooja Mehta

Curva ROC

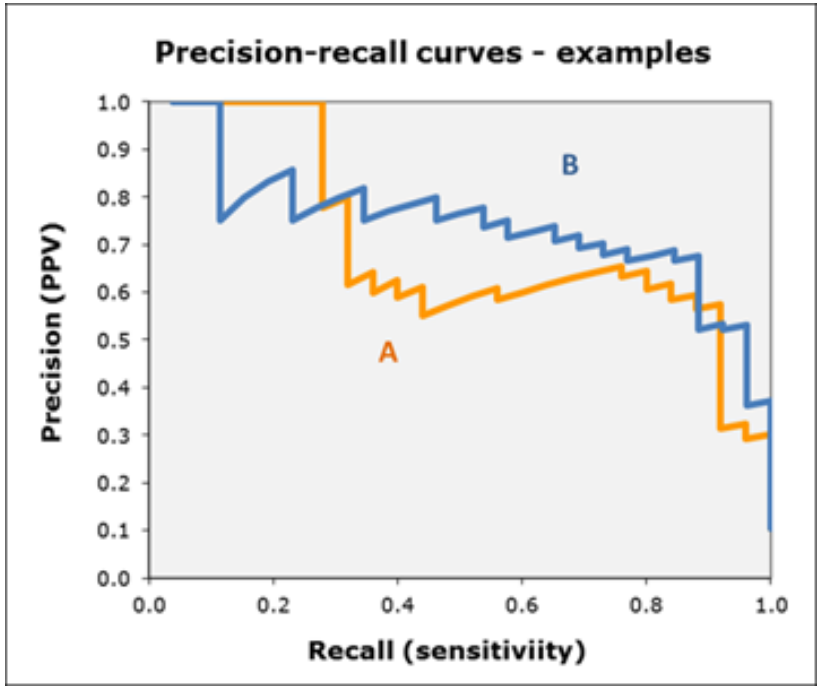
- A AUC é maior quando as curvas de distribuição das classes apresenta pouca interseção.



Animações: Amey Naik & Pooja Mehta

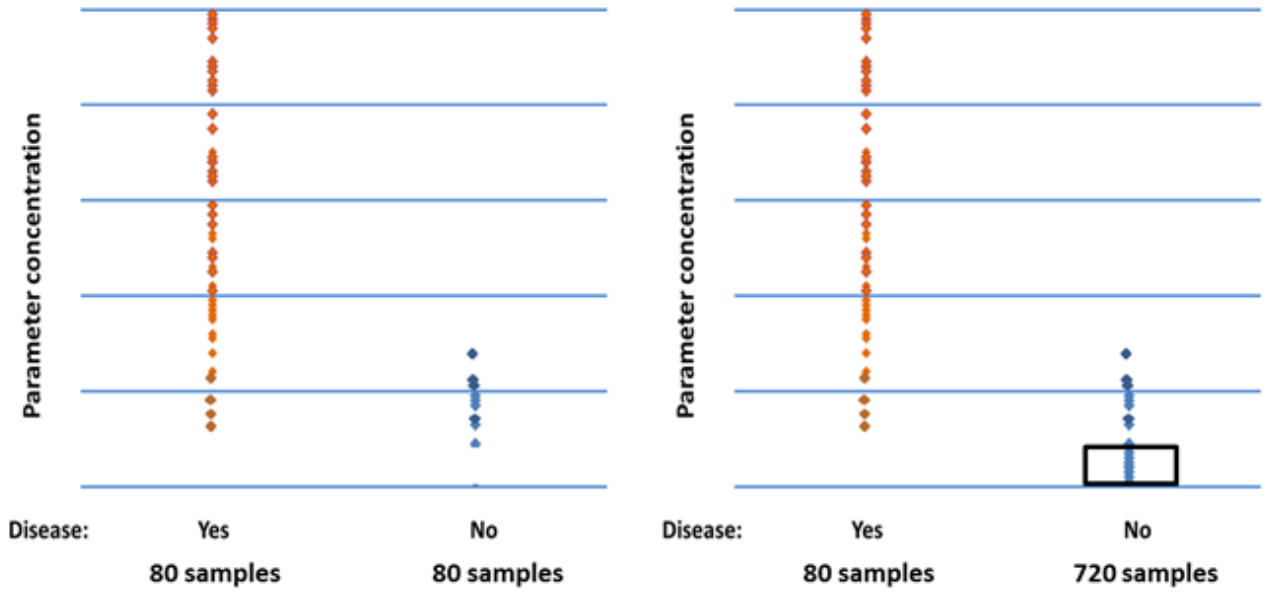
Curva Precision-Recall

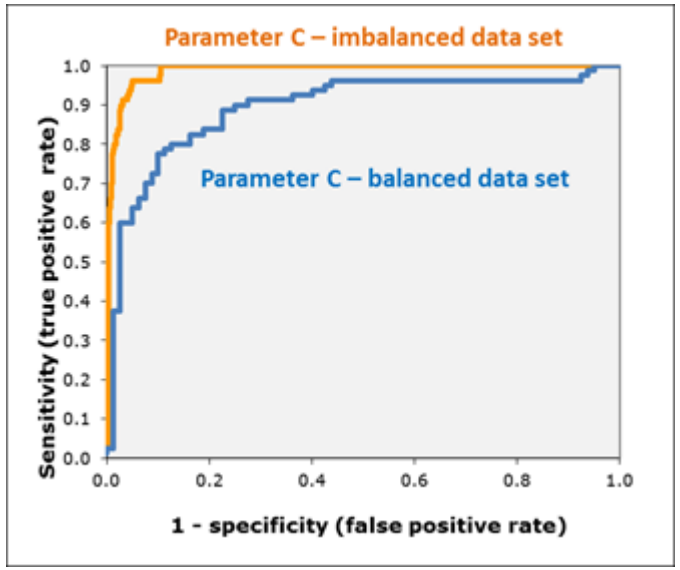
- Mais indicada que a Curva ROC para o caso de problemas com desbalanceamento entre classes de moderado a alto.
- De forma similar à curva ROC, apresenta um gráfico de Precision e Recall, construído a partir de vários pontos de corte.

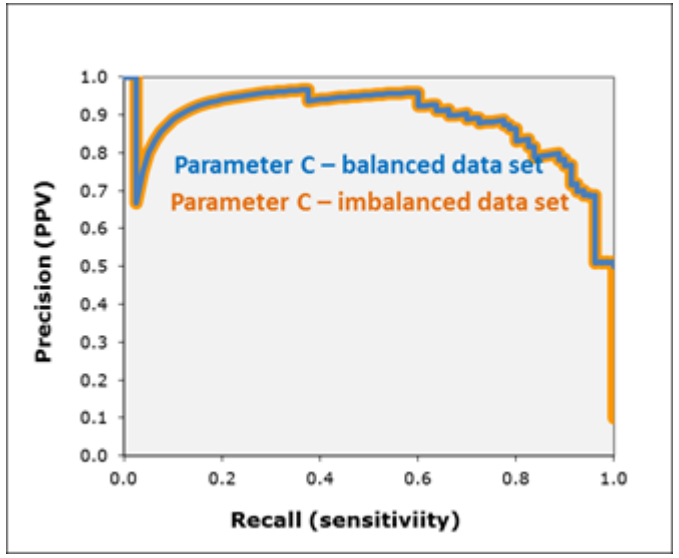


Visualizações: Suzanne Ekelund

Curva Precision-Recall vs. Curva ROC

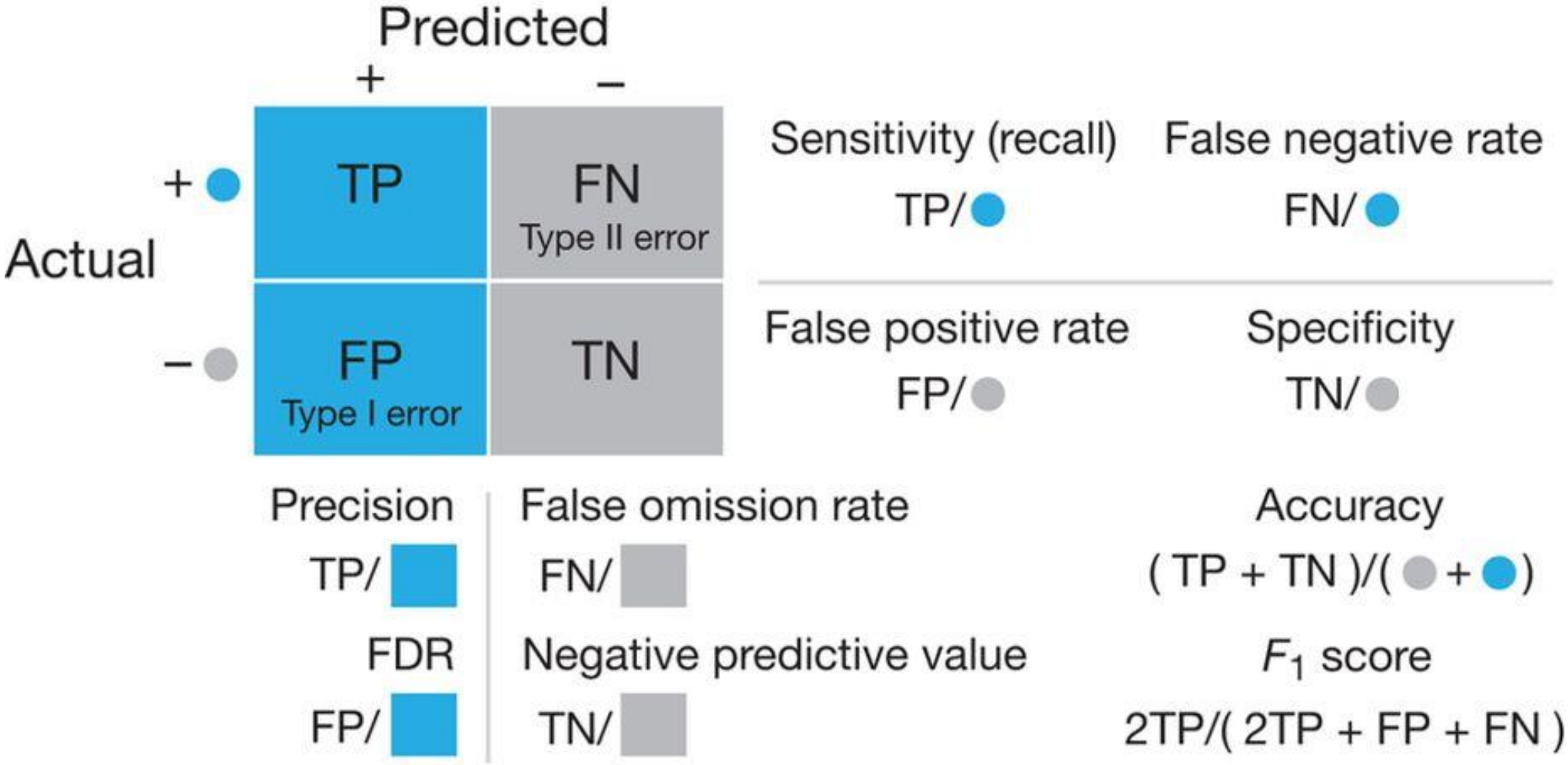






Visualizações: Suzanne Ekelund

Resumo de métricas para classificadores



Fonte: https://www.nature.com/articles/nmeth.3945