Classificação

Luciano Barbosa

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Aprendizado de Máquina

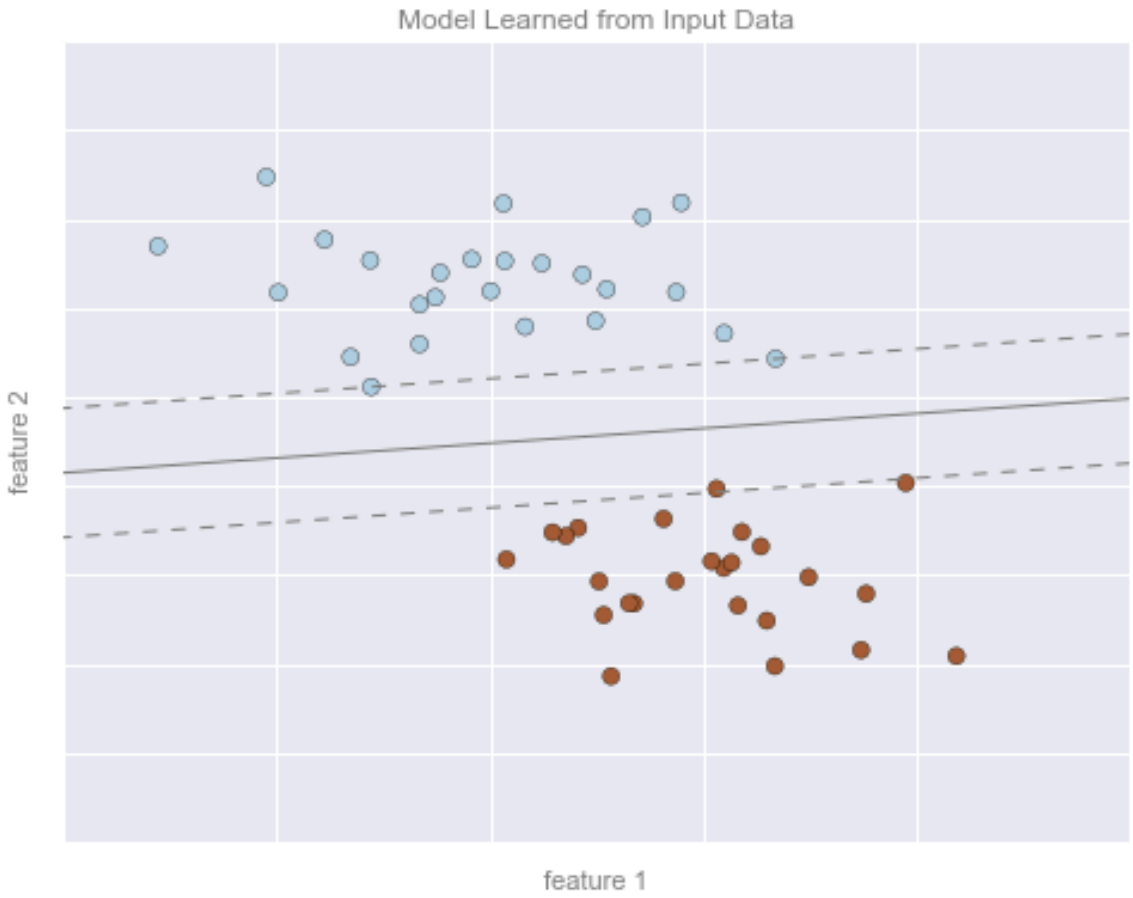
- Construção de modelos preditivos/descritivos a partir de dados
- Categorias:
  - Supervisionado: modela o relacionamento entre features e algum rótulo associado aos dados
  - Não-supervisionado: sem rótulos (“deixar os dados falarem por eles mesmos”)

Aprendizado Supervisionado

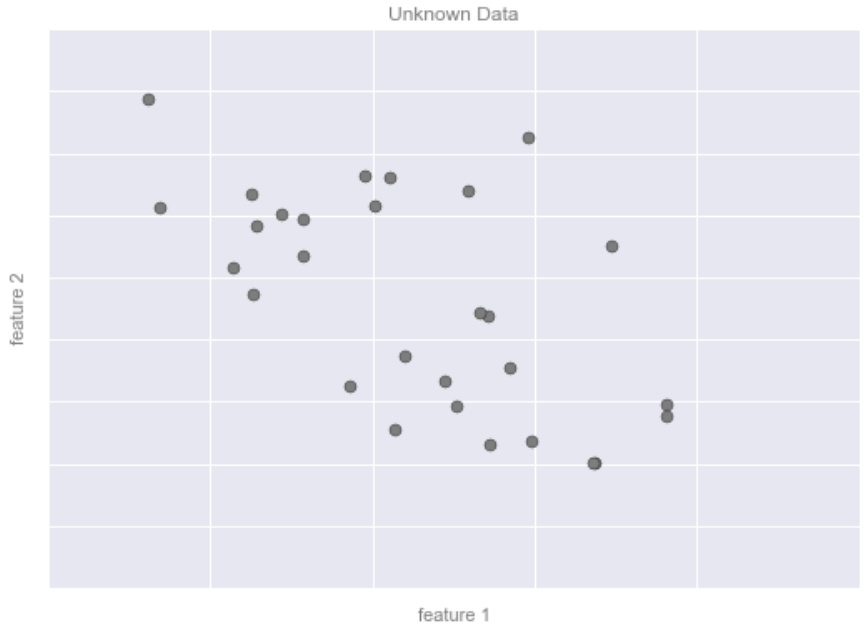
- Objetivo: inferir uma função a partir de exemplos dados para predizer classes de novos exemplos
- Fases:
  - Treinamento
  - Execução
- Tipos
  - Classificação: rótulos são categorias discretas
  - Regressão: rótulos são valores contínuos

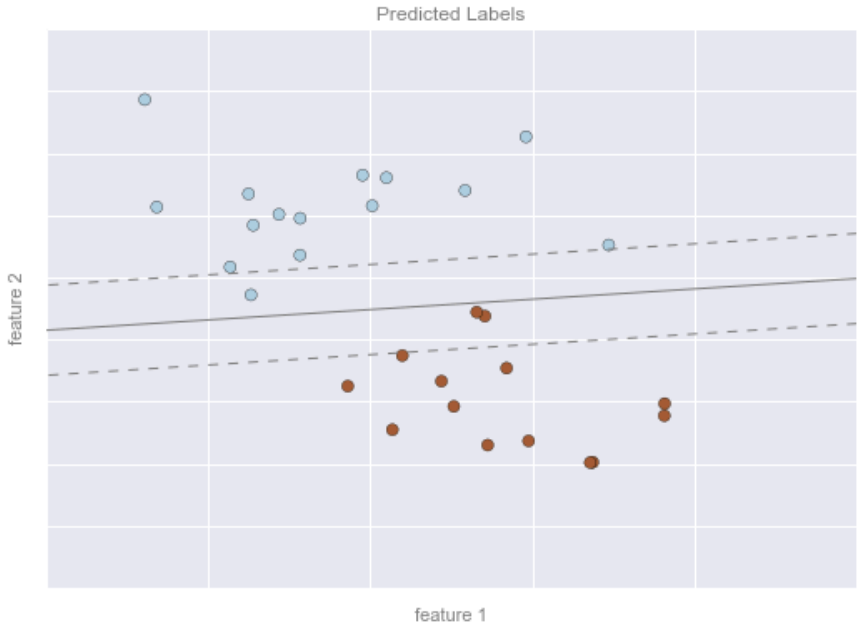
Exemplo de Classificação

Modelo de Classificação

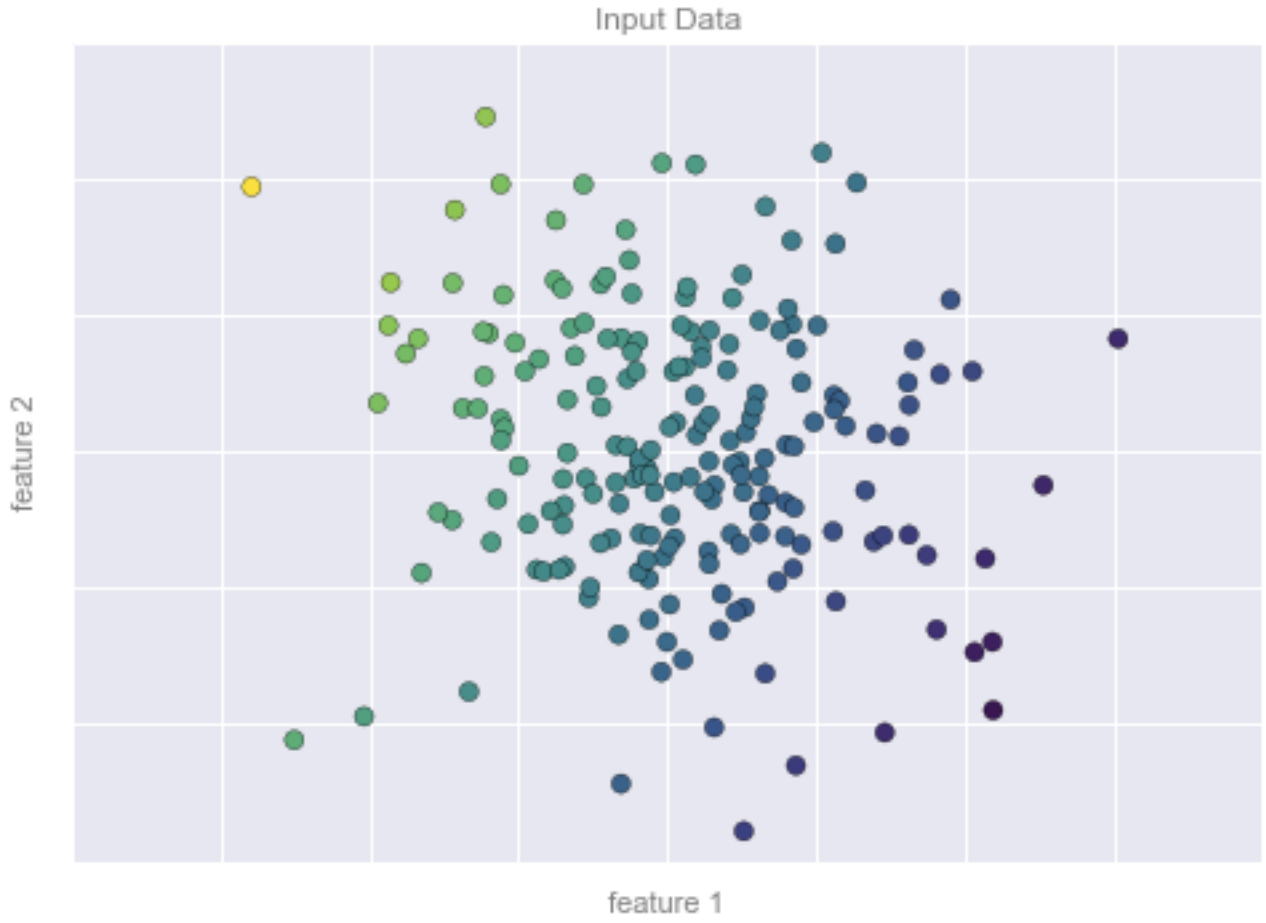


Aplicação do Modelo de Classificação

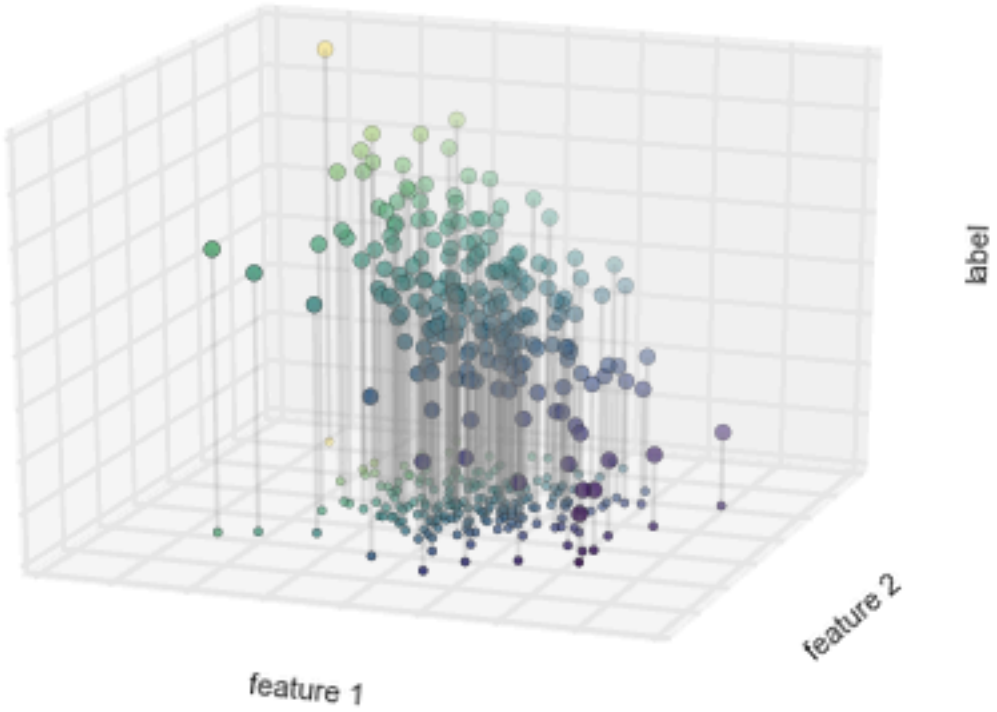




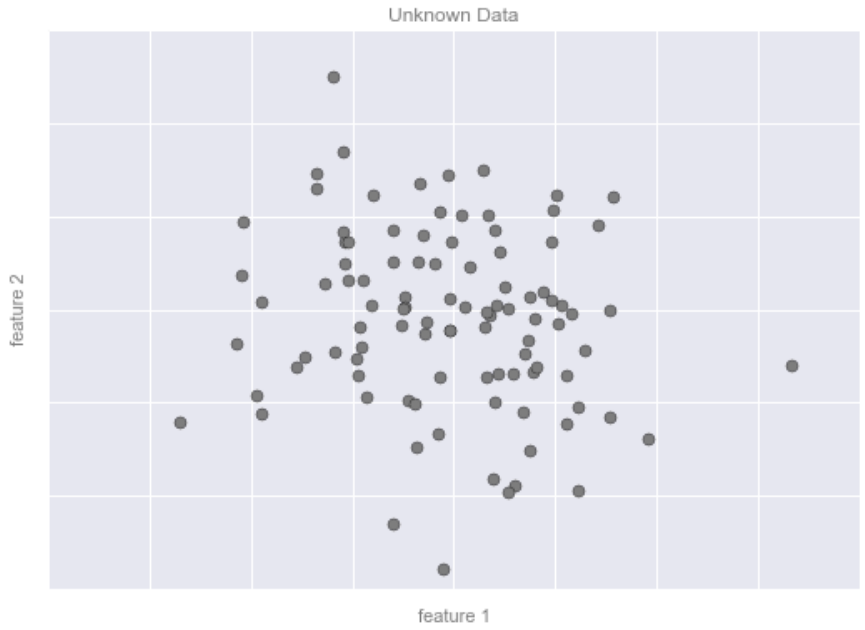
Regressão

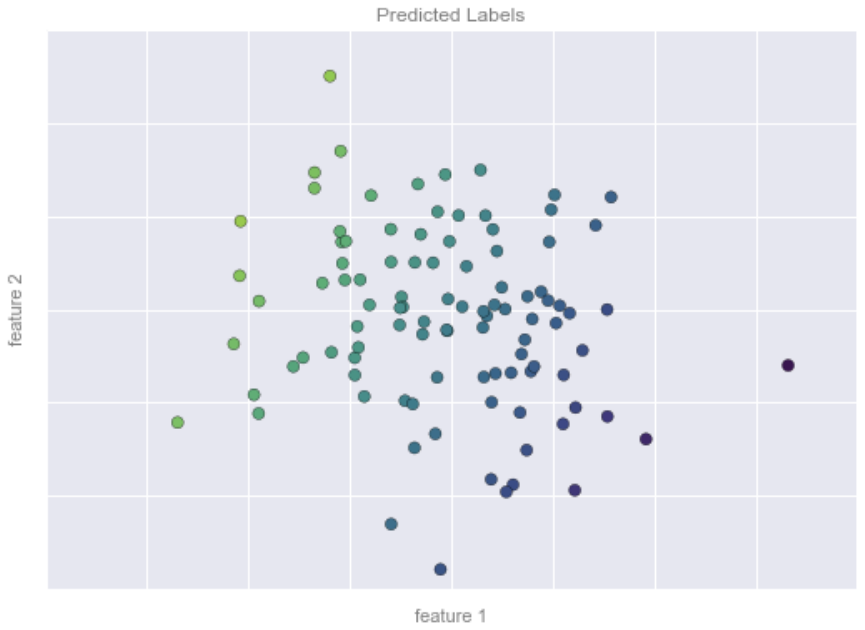


Modelo de Regressão



Aplicando Modelo de Regressão



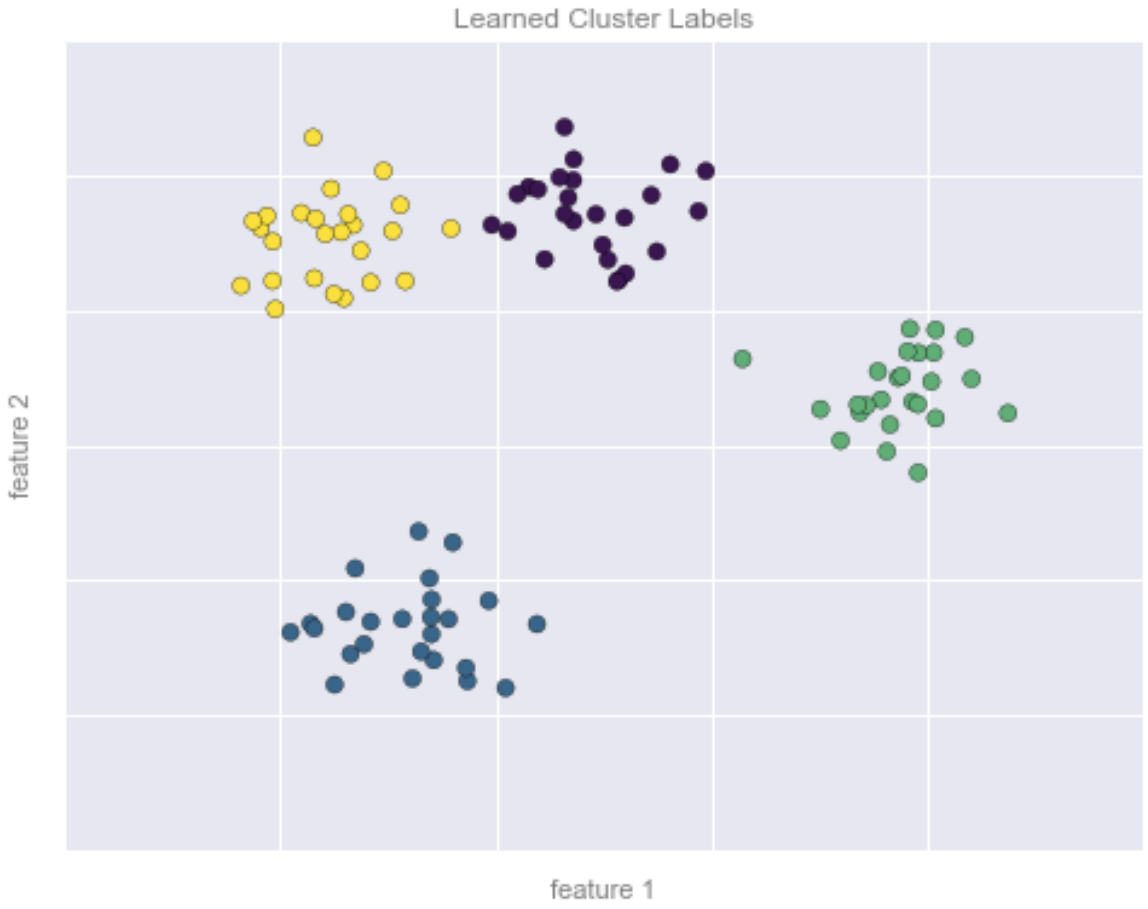


Aprendizado Não-Supervisionado

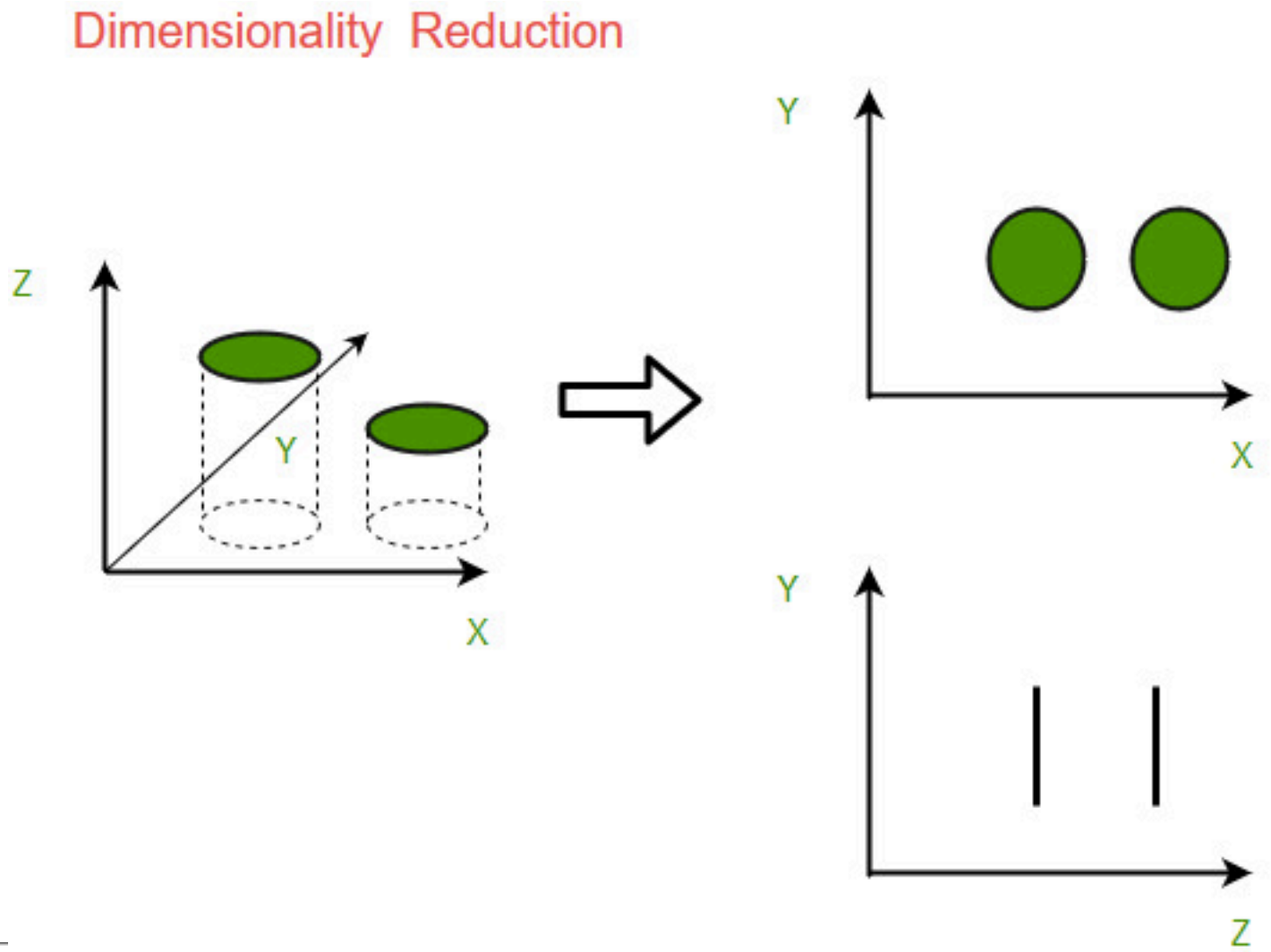
- Agrupamento: identifica grupos de dados
- Redução de dimensionalidade: cria representações mais suscintas dos dados

Agrupamento

Criando Agrupamentos



Redução de Dimensionalidade



Exemplo de Classificação: Filtragem de Spam

From: "" <takworlld@hotmail.com>
Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

==========

Click Below to order:
http://www.wholesaledaily.com/sales/nmd.htm

==========

Modelo Supervisionado

- Conjunto de treinamento: instâncias e rótulos
- Instância representada por seu vetor de características: x_i
- Aprender função f(x)=y que melhor prediz o valor de y dado x
- Para y categórico -> classificação
- Para y numérico -> regressão

Conjunto de Treinamento

	viagra	learning	the	dating	nigeria	*spam*?
$\vec{x}_1 = ($	1	0	1	0	0 $)$	$y_1 = 1$
$\vec{x}_2 = ($	0	1	1	0	0 $)$	$y_2 = -1$
$\vec{x}_3 = ($	0	0	0	0	1 $)$	$y_3 = 1$

Categórico

Modelo Supervisionado: Exemplo

- Filtragem de spam

	viagra	learning	the	dating	nigeria	*spam*?
$\vec{x}_1 = ($	1	0	1	0	0 $)$	$y_1 = 1$
$\vec{x}_2 = ($	0	1	1	0	0 $)$	$y_2 = -1$
$\vec{x}_3 = ($	0	0	0	0	1 $)$	$y_3 = 1$

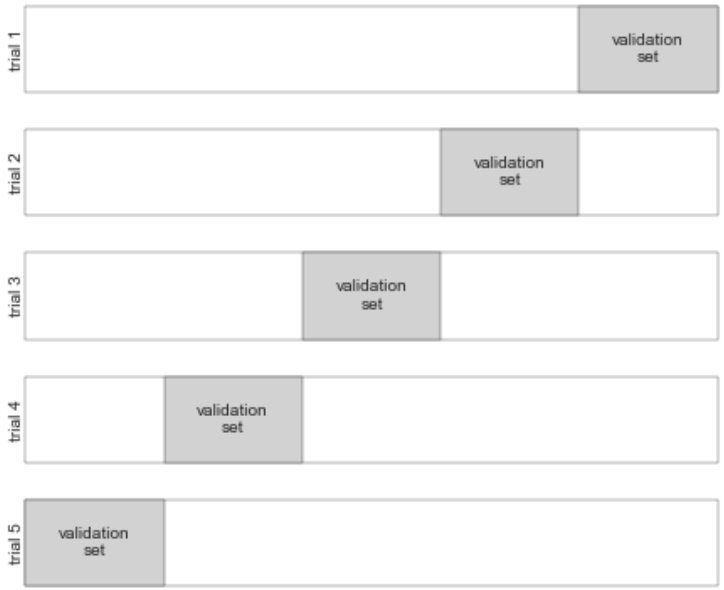
- Features:
  - Palavras: viagra, learning, the, dating nigeria
  - Presença ou ausência
- Classe y: spam (+1) ou não spam (-1)

Features

- Grande importância no resultado da classificação
- Importante: influência com a saída da classificação
  - Ex1: previsão de chuva: temperatura, humidade
  - Ex2: análise de sentimentos: palavras com polaridade (negativa/positiva)

Avaliação de Modelos

- Holdout set: não usado para treinamento
- Treinamento/validação/teste
- Cross-validation



Medidas de Avaliação

- Precision
- Recall
- F1 (F-measure)
- Accuracy

	in the class	not in the class
predicted to be in the class	true positives (TP)	false positives (FP)
predicted to not be in the class	false negatives (FN)	true negatives (TN)

$$\text{precision: } P = TP/(TP+FP)$$

$$\text{recall: } R = TP/(TP+FN)$$

$$F_1 = \frac{1}{\frac{1}{2}\frac{1}{P} + \frac{1}{2}\frac{1}{R}} = \frac{2PR}{P+R}$$

Accuracy = (TP+TN)/(TP+TN + FN + FP)

Selecionando o Melhor Modelo

- Usar conjunto de validação/cross-validation
- Buscar melhores valores dos hiper-parâmetros
  - Algoritmo de ML
  - Espaço de parâmetros
  - Método para buscar valores candidatos
  - Holdout set
  - Métrica de avaliação
- Estratégias:
  - Grid search
  - Automl: TPOT, SMAC, auto-sklearn

Introdução a Scikit-Learn

- Vários algoritmos implementados
- Uniformidade

Scikit-Learn's Estimator API

- Consistency: todos objetos compartilham uma interface única com poucos métodos
- Inspection: todos parâmetros específicos são atributos públicos
- Limited object hierarchy: somente algoritmos são representados como classes Python
- Composition: tarefas de ML podem ser compostas por tarefas mais simples
- Sensible defaults: parâmetros específicos têm valor default definidos

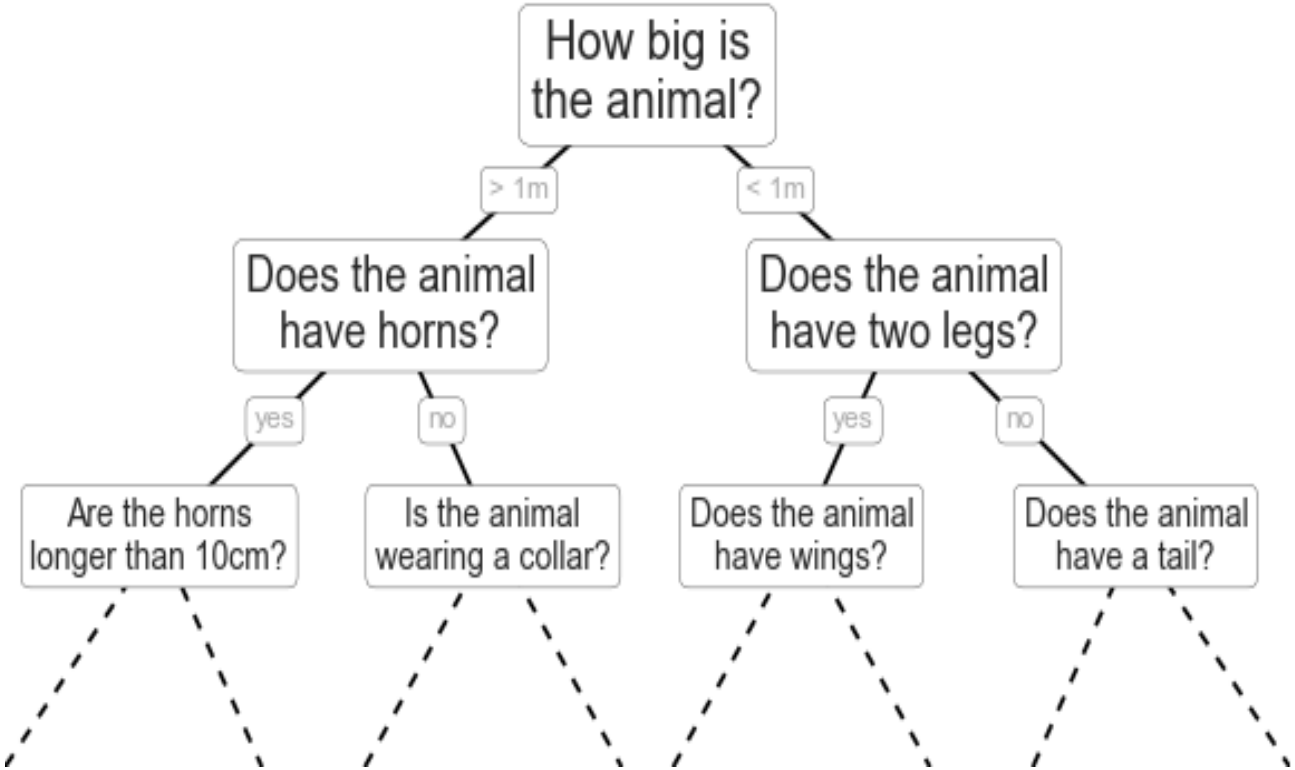
Passos

1. Importar modelo escolhido
2. Escolher os hiper-parâmetros do modelo
3. Criar modelo com função fit
4. Aplicar o modelo a novos dados
   - Supervisionado: função predict
   - Não-supervisionado: funções transform ou predict

Random Forest

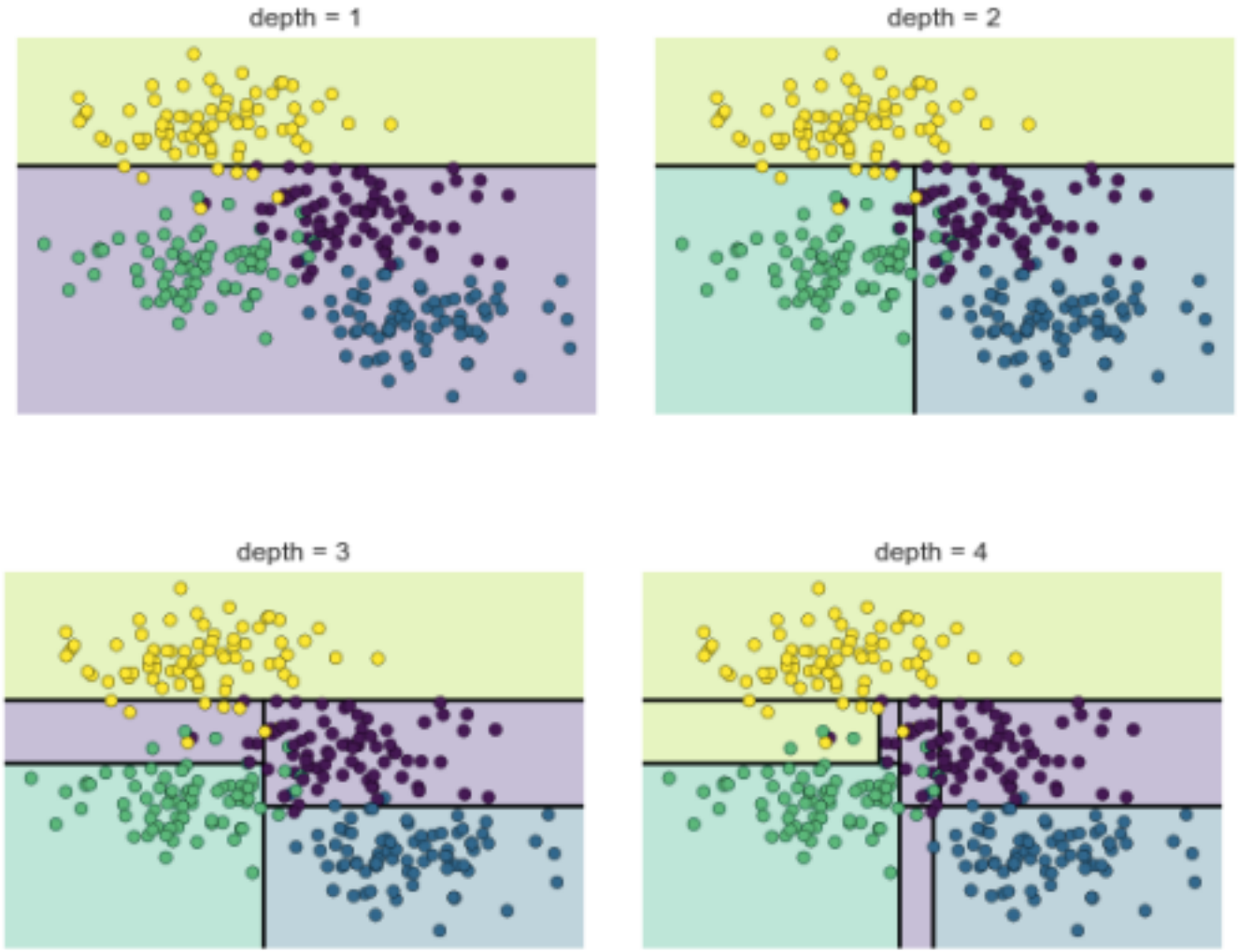
- Ensemble de árvores de decisão
- A soma é melhor que as partes: o voto da maioria é melhor do que modelos individualmente

Exemplo de Árvore de Decisão

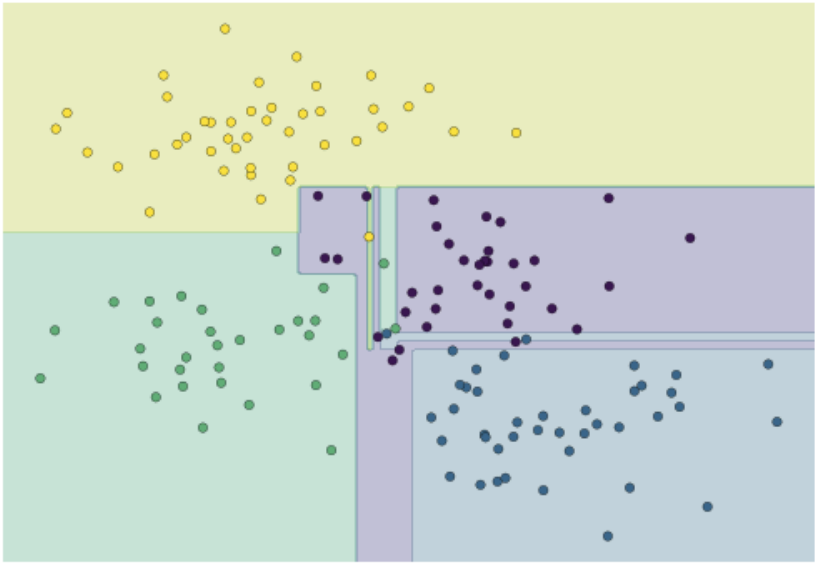


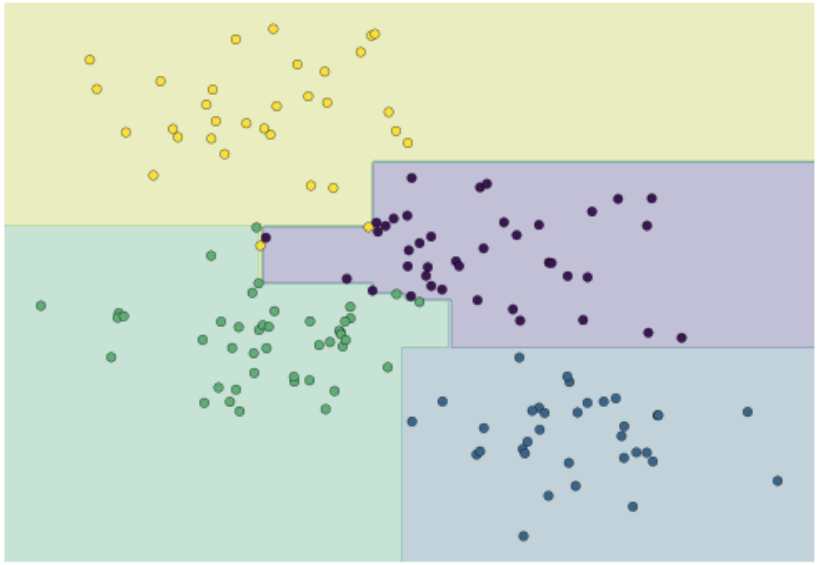
Outro Exemplo com Árvore de Decisão

Dividindo os Dados



Problema com Overfitting



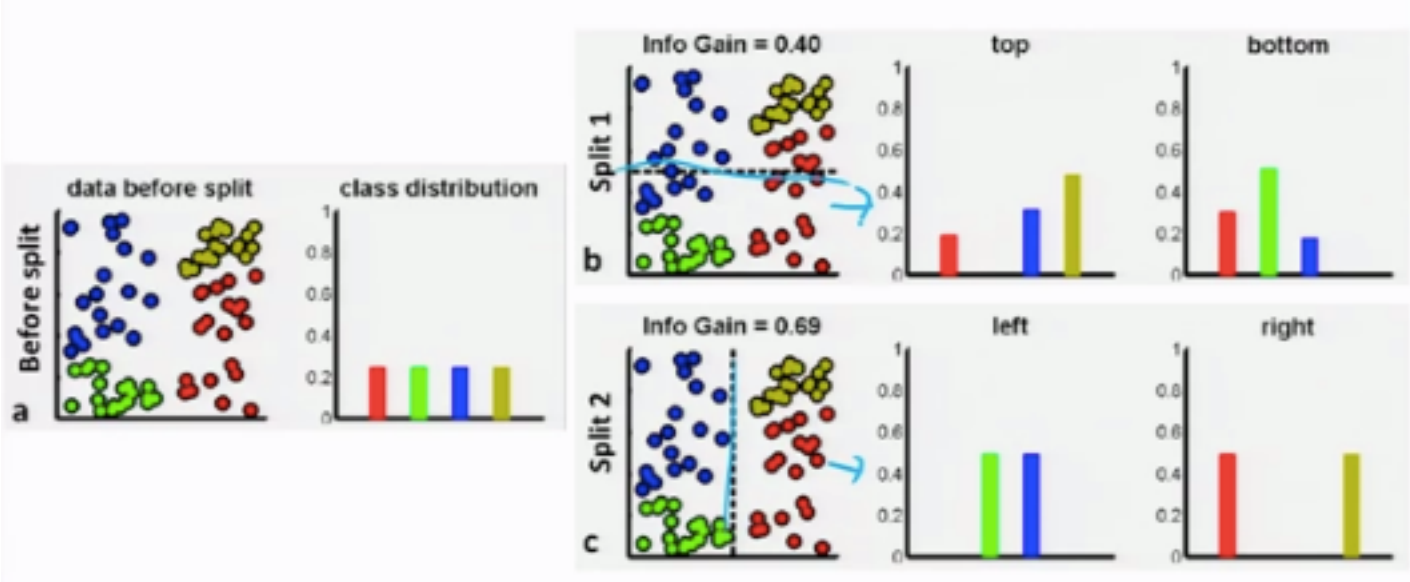


Random Forest

- Ensemble aleatório de árvore de decisão
- Vários estimadores combinados para evitar overfitting
- Bagging:
  - Ensemble de estimadores em paralelo que fazem overfitting
  - Calcula a média das predições (regressão) ou maioria dos votos (classificação)
  - Vários modelos que dão overfitting podem ser combinados para evitar overfitting
- Não assume distribuição gaussiana, relacionamento linear etc

Algoritmo de Construção

1. Seleciona uma amostra do treinamento
2. Para cada nó, seleciona uma amostra m das n features e dessas m é selecionada a com maior information gain ou gini
3. Até não haver mais dados para separar
4. Repete passos 1 a 3 para cada árvore



Information Gain

Vantagens

- Tem obtido excelentes resultados em diversos tipos de datasets
- Execução rápida
- Pode lidar com milhares de features
- Mostra a importância das features na classificação
- Lida bem com dados ausentes

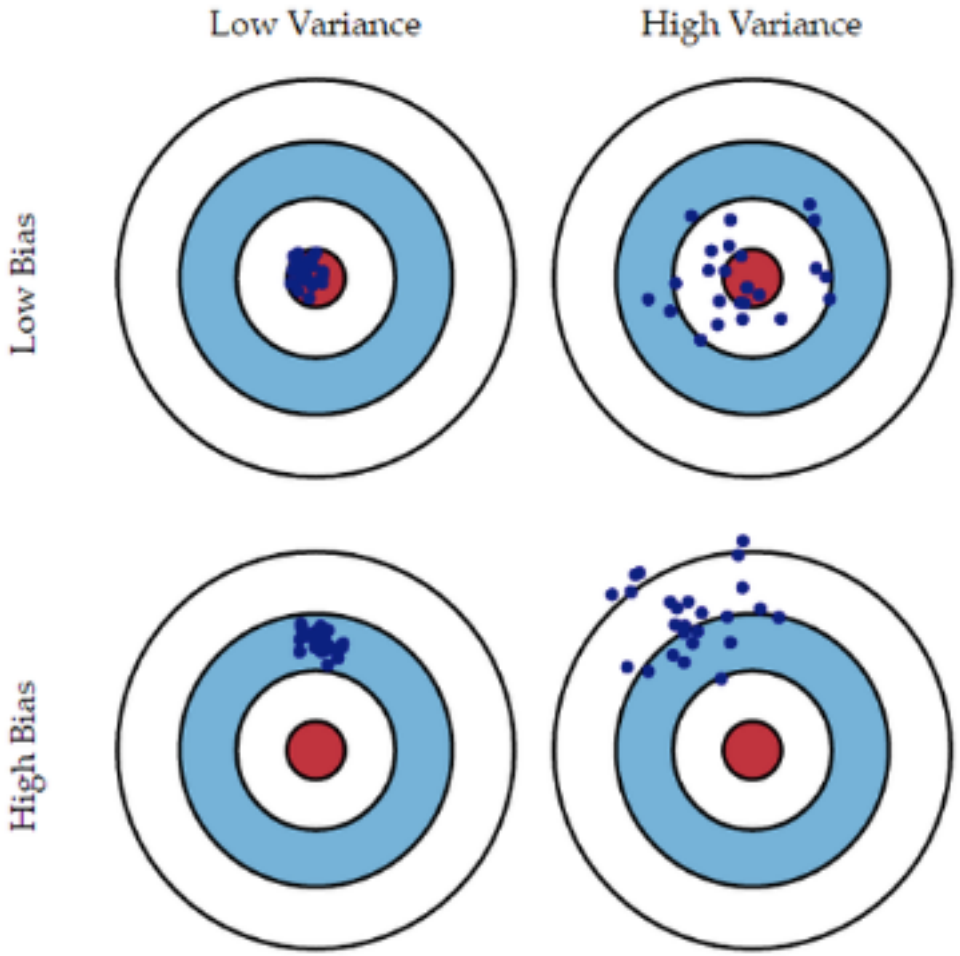
Diagnóstico de Modelos

- Como melhorar o classificador?
  - Mais/menos dados
  - Mais/menos features
  - Modelos mais/menos complexos
- Exemplo:
  - Alvo: 5% de erro
  - Erro no treinamento: 15% (**viés: 15-5= 10**)
  - Erro no teste: 16% (**variance: 16-15 = 1**)
  - Precisa melhor o desempenho no conjunto de treinamento

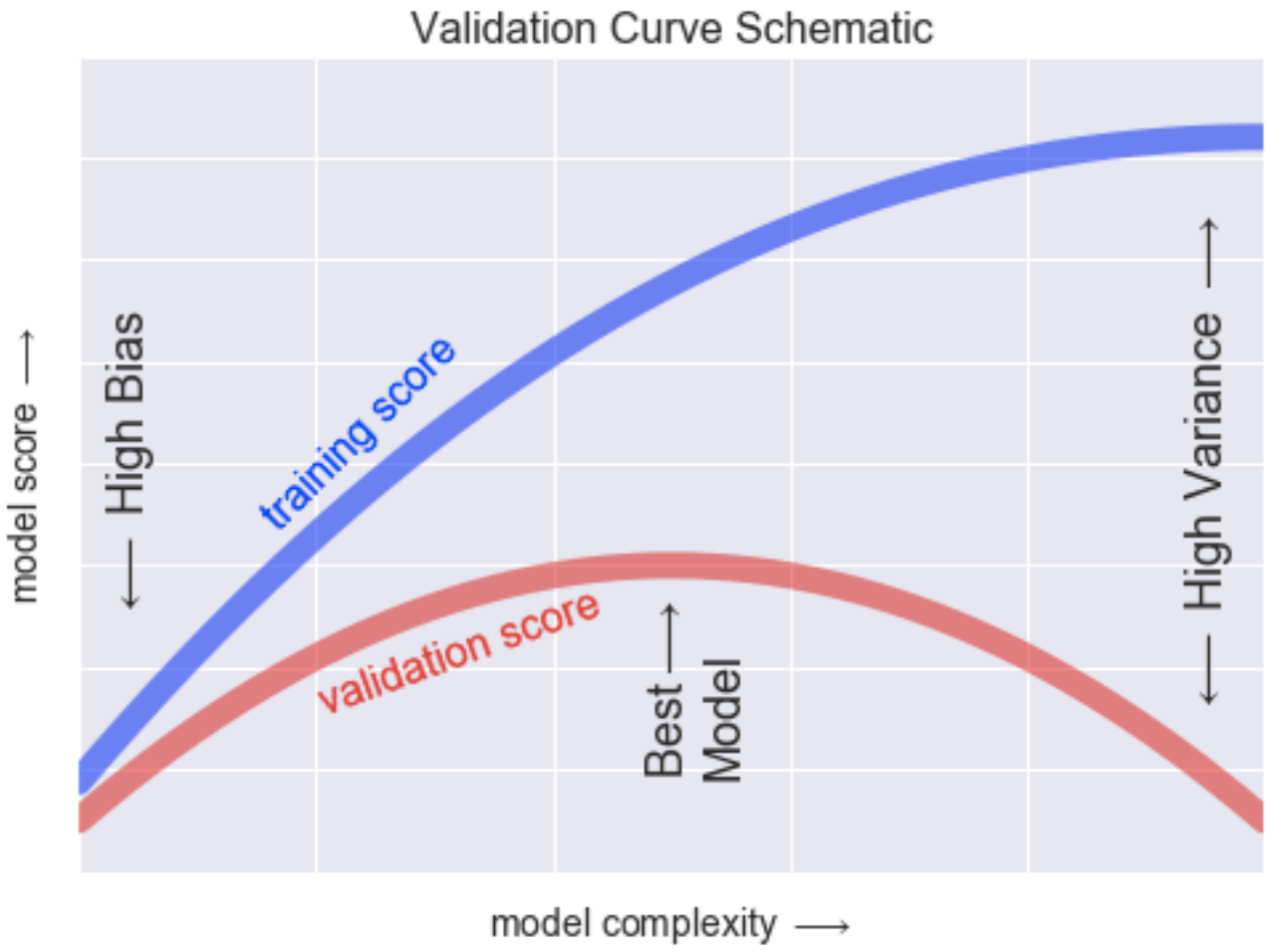
Diagnóstico de Modelos

- Viés:
  - Desempenho no conjunto de treinamento
  - Depende do alvo
- Variância: diferença de desempenho entre treinamento e teste

Viés e Variância

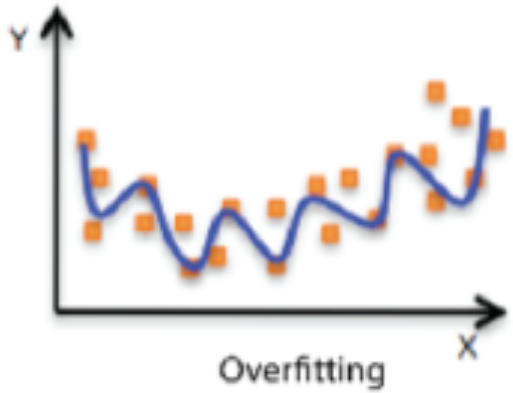


Viés e Variância



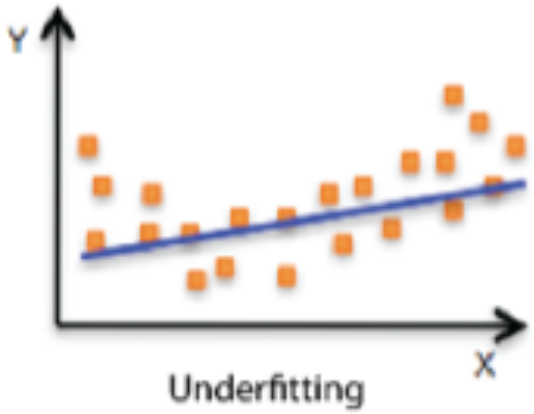
Overfitting

- Bom desempenho no conjunto de treinamento
- Problema em generalizar
- Baixo viés e alta variância
- Exemplo
  - Erro no treinamento: 1%
  - Erro no teste: 11%



Underfitting

- Modelo não modela bem o conjunto de treinamento
- Alto viés e baixa variância
- Exemplo
  - Erro no treinamento: 15%
  - Erro no teste: 16%

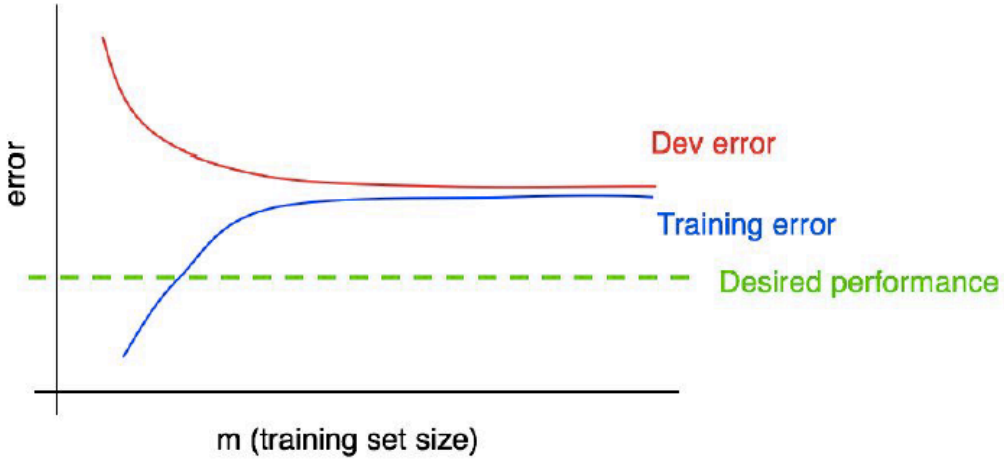


Outros Cenários

- Underfitting e overfitting
  - Exemplo
    - Erro no treinamento: 15%
    - Erro no teste: 30%
- Ideal
  - Exemplo
    - Erro no treinamento: 0.5%
    - Erro no teste: 1%

Lidando com Viés

- Alto viés (underfitting)



  - Aumentar a complexidade do modelo
  - Mais features
  - Não ajuda adicionar mais dados ao treinamento

Lidando com Variância

- Alta variância (overfitting)

  - Adicionar dados ao conjunto de treinamento
  - Feature selection
  - Modelos menos complexos