Técnicas essenciais e qualidade dos modelos

Agente Educacional
Sérgio M. Dias

Agenda

- Classificando as principais técnicas aplicadas em ciência de dados
- Criando bons modelos de dados
- Dados para treinamento, teste e validação (*Holdout* e Validação cruzada)
- Medidas de avaliação (Matriz de confusão, Acurácia e Curva ROC)

Técnicas para Ciência de Dados

As principais técnicas para Ciência de Dados podem ser analisadas a partir da sua capacidade de realizar um conjunto de tarefas

Descrição

Os dados utilizados em uma análise podem descrever um comportamento ou tendência

Técnicas essenciais e qualidade dos modelos

Técnicas para Ciência de Dados

Classificação

A tarefa de classificação consiste em determinar a classe de um registro. Nessa tarefa, os algoritmos utilizados produzem modelos que descrevem as características de cada classe

Regressão

De forma similar ao processo de classificação, a regressão procura predizer o valor de um registro a partir de um modelo gerado através de dados conhecidos

Predição

Similar ao processo de classificação e regressão, a tarefa de predição visa estimar o valor futuro de uma variável

Agrupamento

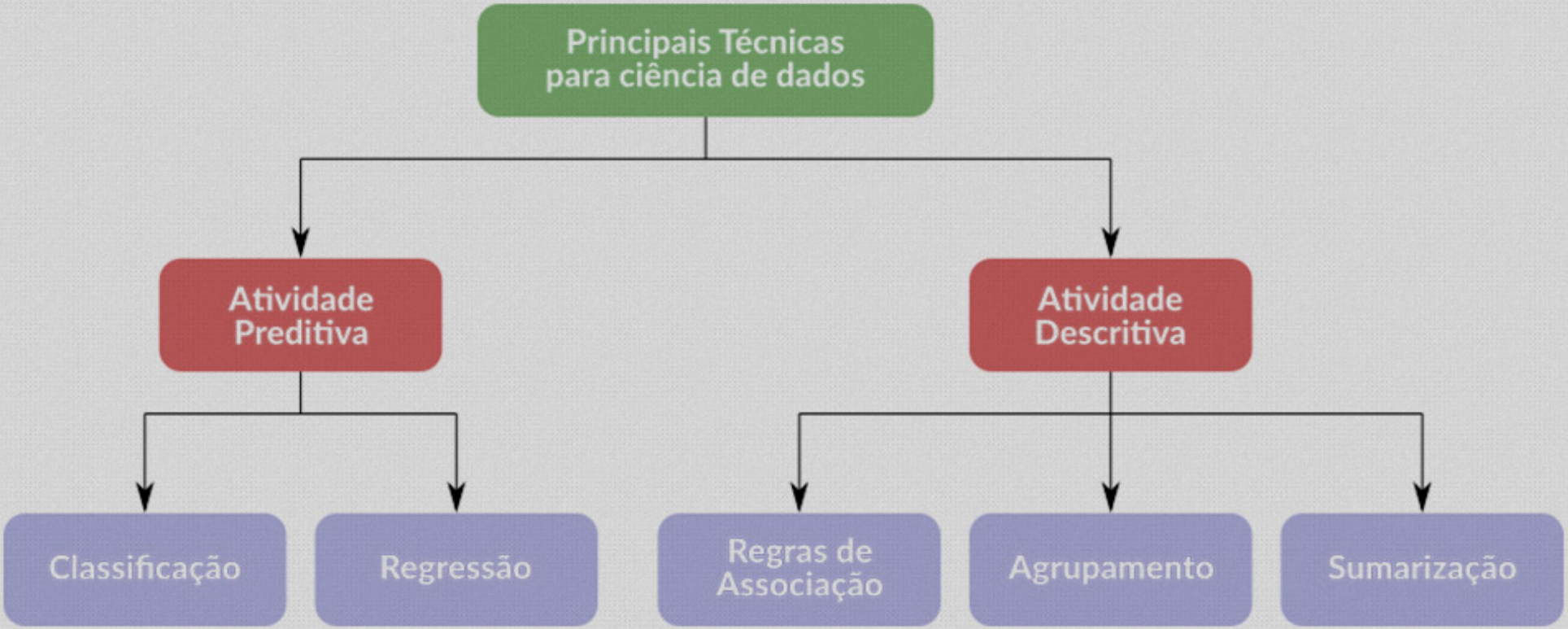
Na tarefa de agrupamento, registros similares são identificados. Cada grupo (*cluster*) é formado por um conjunto de registros similares entre si; entretanto, diferentes dos registros pertencentes aos demais grupos.

Associação

A tarefa de associação consiste em identificar atributos relacionados. Em geral, a associação é expressa através de regras do tipo Se X então Y; em que X e Y são conjuntos de atributos categóricos.

Técnicas essenciais e qualidade dos modelos

Técnicas para Ciência de Dados



Técnicas essenciais e qualidade dos modelos

Técnicas para Ciência de Dados

As principais técnicas para Ciência de Dados também podem ser classificadas a partir da perspectiva de aprendizagem de máquina

Aprendizado supervisionado

- Neste tipo de aprendizagem existe um "professor" que avalia a resposta
- Algoritmos para classificação, regressão

Aprendizado não supervisionado

- Nesta forma de aprendizagem não existe "professor"
- Algoritmos para agrupamento

Aprendizagem por reforço

- Aprendizagem dando recompensas ocasionais

Metodologia de construção e avaliação de modelos

- Não é suficiente se preocupar apenas com a execução do algoritmo que implementa técnicas e gera um *(qualquer)* modelo

- É necessário se preocupar também com o processo de construção dos modelos, o qual deve ser baseado em estratégias que possibilitem a geração de bons modelos

Técnicas essenciais e qualidade dos modelos

Metodologia de construção e avaliação de modelos

- Independentemente da medida de avaliação a ser usada para atestar a qualidade de um modelo, não é adequado avaliá-lo por seu desempenho em relação aos exemplares apresentados no processo de treinamento *(indução)*

- É sempre necessário saber como o modelo se comporta quando aplicado a exemplares que ainda não conhece

- O motivo para essa ressalva é que modelos preditivos, a depender de como são gerados, podem levar à manifestação de um fenômeno bastante conhecido, o sobreajuste *(do inglês overfitting)*.

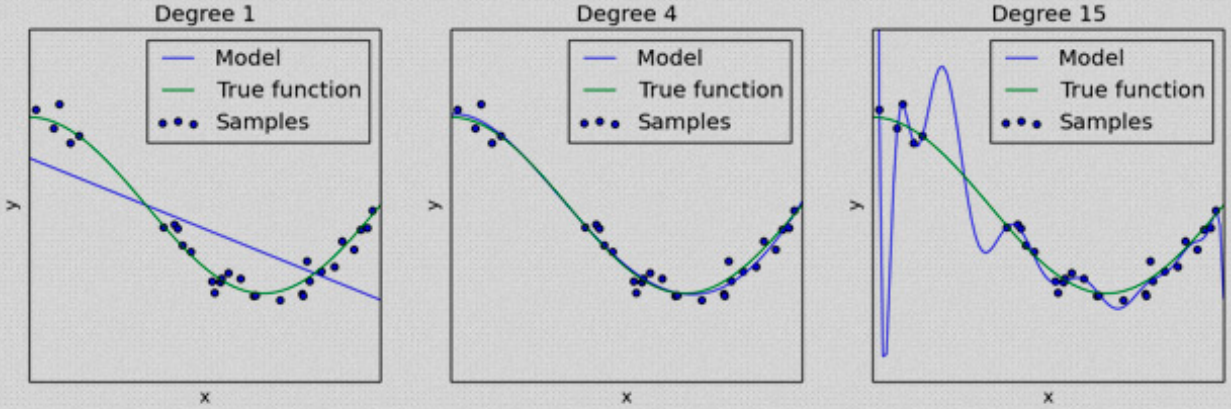
Técnicas essenciais e qualidade dos modelos

Metodologia de construção e avaliação de modelos

- A função linear de grau 1 não é suficiente para um bom treinamento
  resultando em um *underfittin*

- A função de grau 4 tem uma boa aproximação

- Funções de maior grau resultam em um *overfitting*
  Perda da capacidade de generalizar
  Modelo aprende até os erros e se torna muito específico



Técnicas essenciais e qualidade dos modelos

Metodologia de construção e avaliação de modelos

- Precisamos de dados para:

Treinamento

Criar o modelo

Teste

Avaliar o modelo

Validação

Avaliar a generalização do modelo

Técnicas essenciais e qualidade dos modelos

Metodologia de construção e avaliação de modelos

- A seguir, são apresentas duas estratégias para a geração e avaliação de modelos: *Holdout* e Validação cruzada

Holdout

- Na sua forma mais simples, a estratégia *holdout* pressupõe a criação de dois subconjuntos de dados disjuntos, a partir do conjunto de dados disponível para uso na indução do modelo.

- Um dos subconjuntos será usado para treinamento (indução) do modelo preditivo, e o segundo, para teste após o término do treinamento e, consequentemente, para aplicação das medidas de avaliação do modelo.

Técnicas essenciais e qualidade dos modelos

Metodologia de construção e avaliação de modelos

Holdout

- Tradicionalmente, os dois subconjuntos são gerados de forma que 70% dos exemplares do conjunto de dados sejam alocados para o subconjunto de treinamento

- Os 30% restantes são alocados no subconjunto de teste

- Alternativamente, as porcentagens 60% e 40% podem ser usadas

- Os exemplares a serem alocados em cada um dos subconjuntos devem ser escolhidos aleatoriamente.

Validação cruzada

- Na estratégia de validação cruzada, todos os exemplares farão parte, em algum momento, do conjunto de dados usado no teste do modelo preditivo

- Para implementar essa situação, o conjunto de dados será dividido em K subconjuntos disjuntos, com alocação aleatória de exemplares para cada subconjunto *(podendo ser aplicado aqui um controle referente à distribuição de classes, como já explicado)*

- Assim, o conjunto de dados D será dividido nos subconjuntos D1 ... Dk ... DK

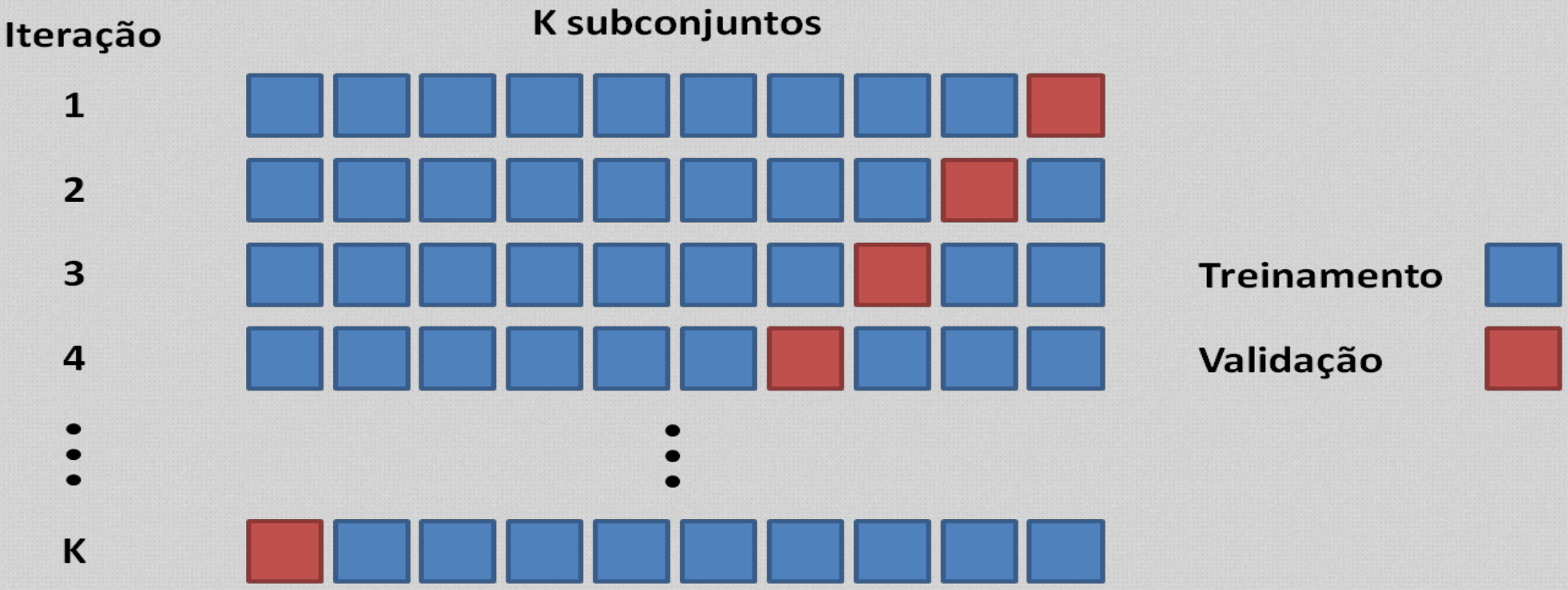
- Mantendo constante o conjunto de parâmetros livres da técnica de indução de modelo adotada, a geração de K modelos preditivos será realizada da seguinte forma:

    - Um dos subconjuntos será reservado para ser usado como conjunto de teste, e os K-1 restantes vão compor o conjunto de treinamento

    - Esse procedimento será repetido K vezes, alterando o subconjunto reservado para teste do modelo

Técnicas essenciais e qualidade dos modelos

Metodologia de construção e avaliação de modelos



Medidas de avaliação

Quando desenvolvemos sistemas, métodos ou testes que envolvem a detecção, diagnósticos ou previsão de resultados, é importante validar seus resultados de forma a quantificar seu poder discriminativo e identificar um procedimento ou método como bom ou não para determinada análise.

Referência: http://crsouza.com/2009/07/13/analise-de-poder-discriminativo-atraves-de-curvas-roc/

Matriz de confusão

Seu funcionamento é simples

- consideramos valores positivos que o sistema julgou positivos como verdadeiros positivos *(acerto)*

- valores positivos que o sistema julgou negativos como falsos negativos *(erro)*

- valores negativos que o sistema julgou como negativos como verdadeiros negativos *(acerto)*, e

- valores negativos que o sistema julgou positivos como falsos positivos *(erro)*

Técnicas essenciais e qualidade dos modelos

Metodologia de construção e avaliação de modelos

Matriz de confusão

		Valor Verdadeiro (confirmado por análise)	
		positivos	negativos
Valor Previsto (predito pelo teste)	positivos	**VP** Verdadeiro Positivo	**FP** Falso Positivo
	negativos	**FN** Falso Negativo	**VN** Verdadeiro Negativo

Acurácia

A proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo

ACC = (VP + VN) / (P + N)

(ACC = TOTAL DE ACERTOS / TOTAL DE DADOS NO CONJUNTO)

		Valor Verdadeiro (confirmado por análise)	
		positivos	negativos
Valor Previsto (predito pelo teste)	positivos	**VP** Verdadeiro Positivo	**FP** Falso Positivo
	negativos	**FN** Falso Negativo	**VN** Verdadeiro Negativo

Veja outras medidas em: http://crsouza.com/2009/07/13/analise-de-poder-discriminativo-atraves-de-curvas-roc/

A *Curva de Receiver Operating Characteristic* (ou curva ROC)

Tipo de gráfico que permite análise da qualidade de um modelo

Técnicas essenciais e qualidade dos modelos

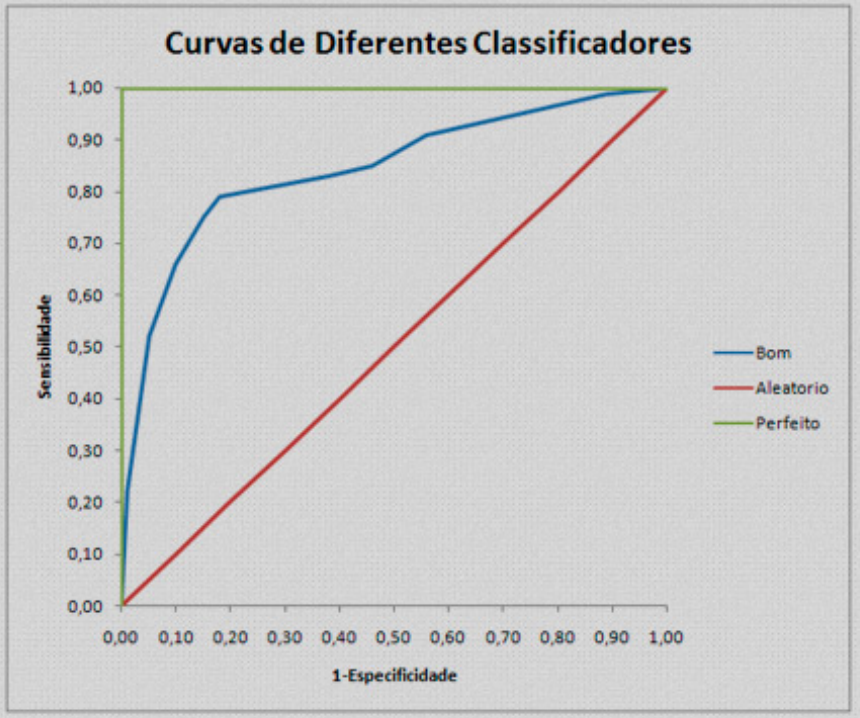
Metodologia de construção e avaliação de modelos

A Curva de *Receiver Operating Characteristic* (ou curva ROC)

- Um classificador perfeito corresponderia a uma linha horizontal no topo do gráfico

- Na prática, curvas consideradas boas estarão entre a linha diagonal e a linha perfeita, onde quanto maior a distância da linha diagonal, melhor o sistema

- A linha diagonal indica uma classificação aleatória



[Algumas funcionalidades do KNIME]

Obrigado!

Agente Educacional

Sérgio M. Dias

sergio.dias@serpro.gov.br | #31 6539

Demais agentes educacionais sobre o assunto

Marcelo Pita | marcelo.pita@serpro.gov.br | #81 8794

Gustavo Torres | gustavo.gamatorres@serpro.gov.br | #31 6950