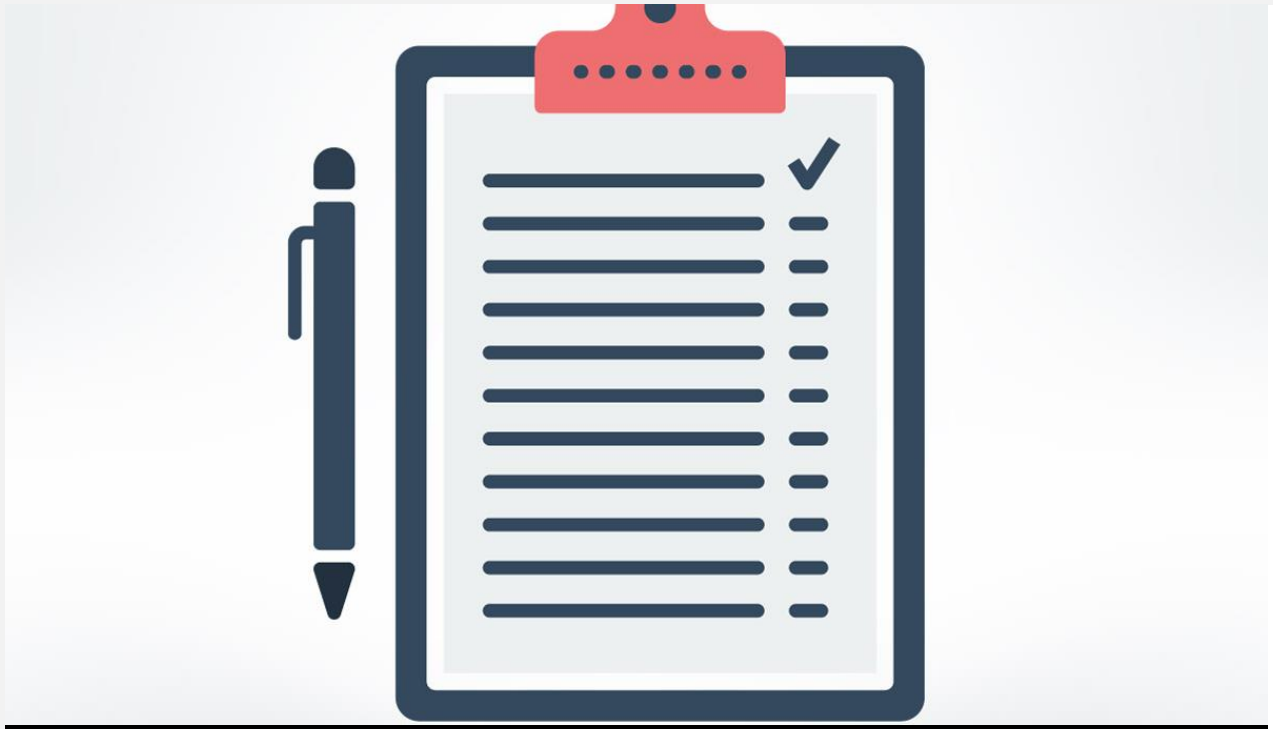
Classificação

Eduardo Silva – easilva91@gmail.com



Agenda

Classificação

1. Classificação de Dados
2. Supervisionado & Não Supervisionado
3. Processo de Predição de Dados
    1. Como avaliar um modelo?

Classificação de Dados

- Muitos problemas práticos possuem registros históricos relacionando situações específicas com determinados resultados.
  - Administradoras de cartões de crédito possuem registros de transações passadas e a informação se forma fraudulentas ou não.
  - Empresas possuem registros de funcionários com seu perfil e desempenho no trabalho.
- Quando cada registro possui um rótulo de classe ou um valor de saída associado que representa o resultado histórico de registros passados, o objetivo da análise é, quase invariavelmente, construir um modelo que possa ser usado para predizer qual seria essa saída para novos registros, ou seja, registros cuja classe ou valor de saída é desconhecido.
- Esse tipo de tarefa é chamado genericamente de predição e pode ser de dois tipos: discreta, denominada classificação; ou contínua, denominada estimação.

Os rótulos auxiliam no treinamento de algoritmos de classificação (algoritmos supervisionados)
1 = sim ; 0 = não

Var 1	Var 2	Var 3	Var 4	Rótulo
12	34	56	78	1
89	56	74	15	0

Supervisionado & Não Supervisionado

	Não Supervisionado	Supervisionado

Contínuo	• Agrupamento & Redução de Dimensionalidade
	  • SVD (redução)
	  • PCA (redução)
	  • K-means
		• Regressão
		  • Linear
		  • Polinomial
		• Arvore de Decisão
		• Random Forest

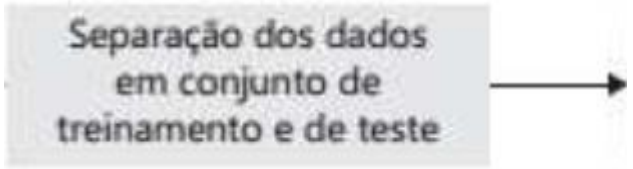
Categórico	• Análise de Associação
	  • Apriori
	  • FP-Growth
	• Modelo de Markov
		• Classificação
		  • KNN
		  • Tress
		  • Regressão Logistica
		  • Naive-Bayes
		  • SVM
		  • Redes Neurais
		  • Perceptron

Processo de Predição de Dados

Figura 5.4 Fluxo do processo de construção e aplicação de um modelo preditivo

Introdução à Mineração de dados: Conceitos Básicos – 1º Edição

Processo de Predição de Dados



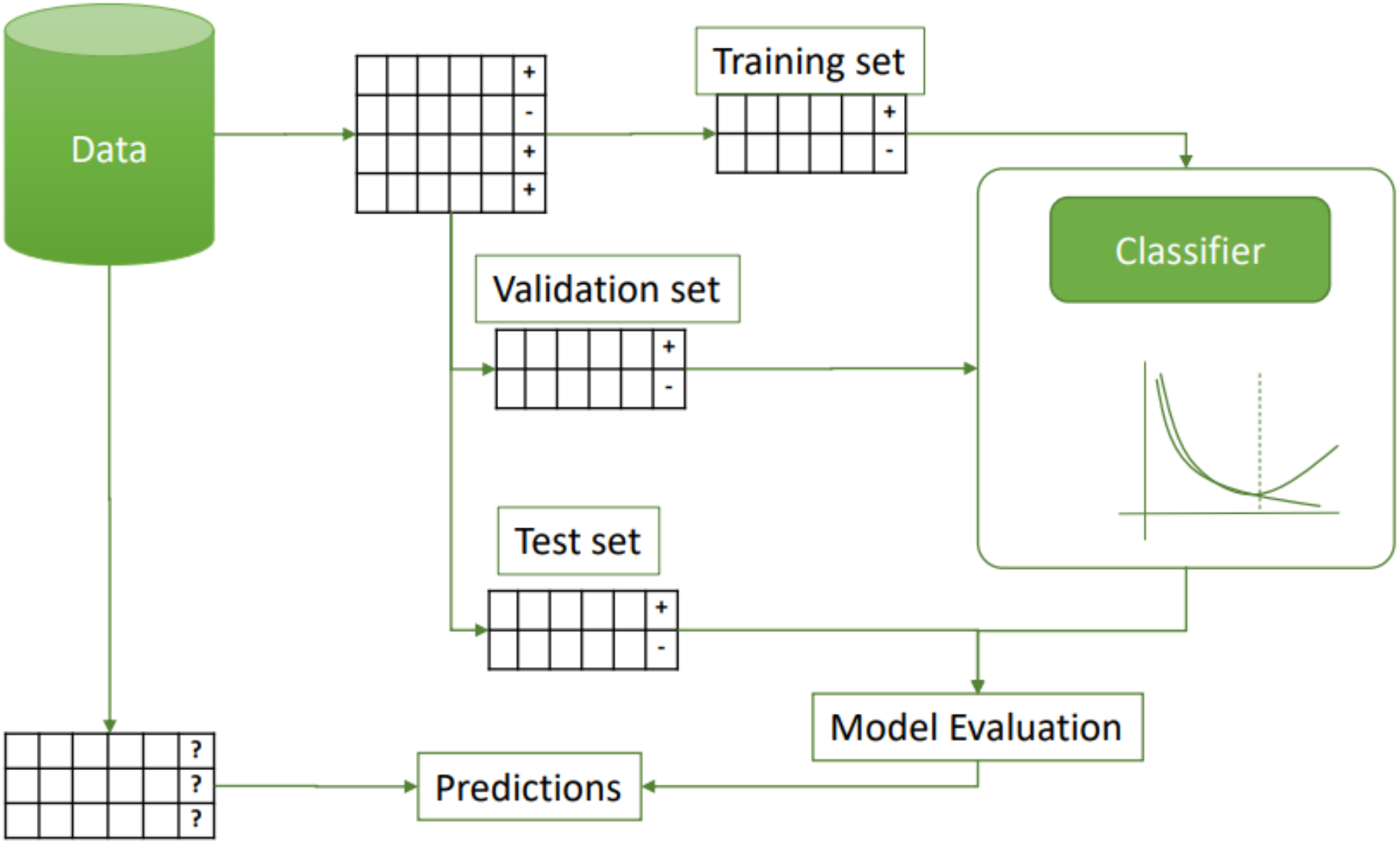
• Separação dos Dados

- Durante a separação dos dados é preciso levar em consideração alguns fatores, normalmente os dados são separados em treino e teste.
  - Ao fazer a separação é interessante que se faça uma divisão onde 70% dos dados sejam de treino e 30% de teste.
  - Caso seus dados contenham mais de 10 mil observações, faça uma divisão de validação, assim sendo 70% treino, 15% teste, 15% validação.

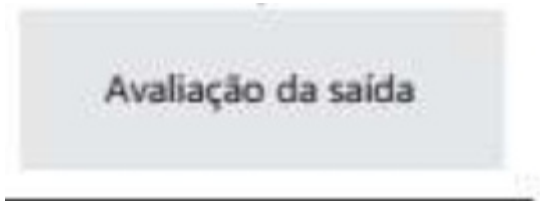
- Mas como fazer a separação?
  - Normalmente com uma ou duas linhas de código

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=
1)
```

Processo de Predição de Dados



Processo de Predição de Dados



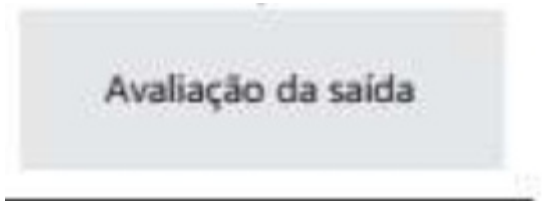
- Como avaliar o resultado de um Modelo?

- Existem algumas métricas já conhecidas, que fazem uso da matrix de confusão.
  - Acurácia/Accuracy – proporção de eventos corretamente identificados (positivo ou negativo) em todos os eventos.
  - Precisão/Precision – mede a proporção de eventos positivos corretamente classificados como positivos.
  - Revocação/Recall – mede a proporção de quantos eventos retornados são positivos.
  - Sensitivity – mede a proporção de eventos identificados como positivos em todos os eventos positivos.
  - Specificity – mede a proporção de eventos identificados como negativos em todos os eventos negativos.

- Compreendendo a matriz de confusão
  - TP/VP (true positive/verdadeiro positivo) – positivos verdadeiros
    - Nº de exemplos classificados positivos e que são positivos. (corretamente classificados)
  - FP (false positive/falso positivo) – positivos falsos
    - Nº de exemplos classificados positivos que são negativos – (incorretamente classificados)
  - TN/VN (true negative/verdadeiro negativo) – negativos verdadeitos
    - Nº de exemplos classificados negativos que são negativos – (corretamente classificados)
  - FN (false negative/falso negativo) – negativos falsos
    - Nº de exemplos classificados negativos que são positivos. – (incorretamente classificados)

		PREDITO	
		Classe A	Classe B
VERDADEIRO	Classe A	VP	FN
	Classe B	FP	VN

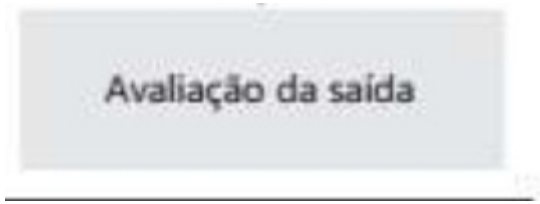
Processo de Predição de Dados



- Como avaliar o resultado de um Modelo?
  - Acurácia/Accuracy – (VP+VN)/(P+N) (acurácia nem sempre é a melhor métrica, depende da situação)
  - Precisão/Precision – VP/(VP+FP)
  - Revocação/Recall – VP/(VP+FN)
  - Sensitivity – VP/P
  - Specificity – VN/N
  - Erro – (FP+FN)/(P+N)

		PREDITO	
		Classe A	**Classe B**
VERDADEIRO	**Classe A**	VP	FN
	Classe B	FP	VN

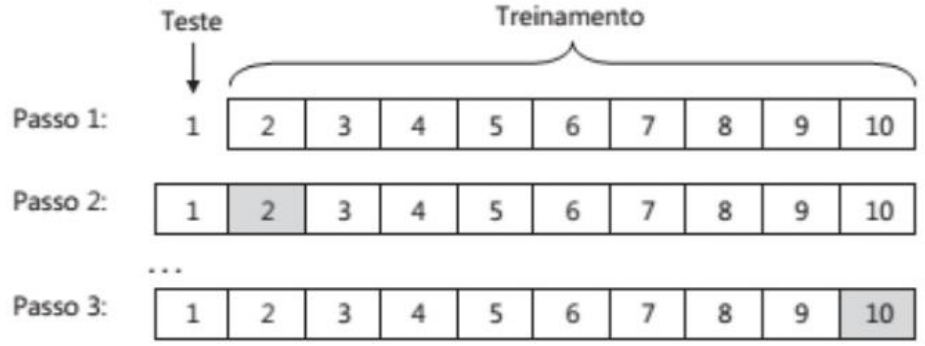
Processo de Predição de Dados



- Como avaliar o resultado de um Modelo?

- As métricas de avaliação mostradas anteriormente apresentam como o modelo está aprendendo, e de que forma está se dando as predições, no entanto, existem algumas formas mais especificas de avaliar o modelo, tentando assim não deixar brechas.
- K-Folds
  - Uma forma bastante comum de validação cruzada em mineração de dados é a chamada validação cruzada em **k-pastas/folds**, que consiste em dividir a base de dados em **k** subconjuntos, sendo **k-1** pastas para treinamento e 1 pasta para teste.
  - Esse processo de treinamento e teste é repetido com todos os **k** subconjuntos, e a média dos desempenhos para as bases de treinamento e as bases de teste é adotado como indicador de qualidade do modelo.

Figura 5.5 Validação cruzada do tipo 10-pastas

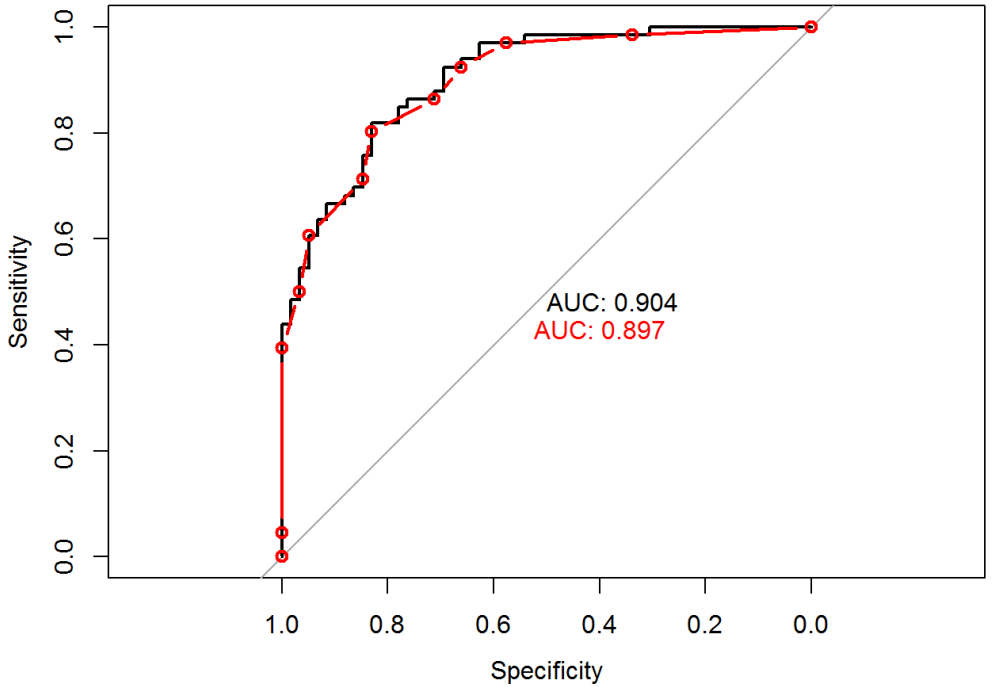


Processo de Predição de Dados

Avaliação da saída

- Como avaliar o resultado de um Modelo?

- As métricas de avaliação mostradas anteriormente apresentam como o modelo está aprendendo, e de que forma está se dando as predições, no entanto, existem algumas formas mais especificas de avaliar o modelo, tentando assim não deixar brechas.
- ROC e AUC (área sobre a curva/ area under the curve) (caso o AUC seja menor que 0.5, o modelo não está bom)



Observando na prática

- https://www.youtube.com/watch?v=DZR5vzm4T5Y
- https://www.kaggle.com/nirajvermafcb/support-vector-machine-detail-analysis
- https://www.kaggle.com/kanncaa1/roc-curve-with-k-fold-cv

- SVM
- https://www.vooo.pro/insights/support-vector-machine-simplificado/
- https://lamfo-unb.github.io/2017/07/13/svm/

- Compilado interessante:
- https://www.vooo.pro/insights/fundamentos-dos-algoritmos-de-machine-learning-com-codigo-python-e-r/

Obrigado

Real Python