
Tera

Aula #24

Execução de Projetos e Detecção de Anomalias

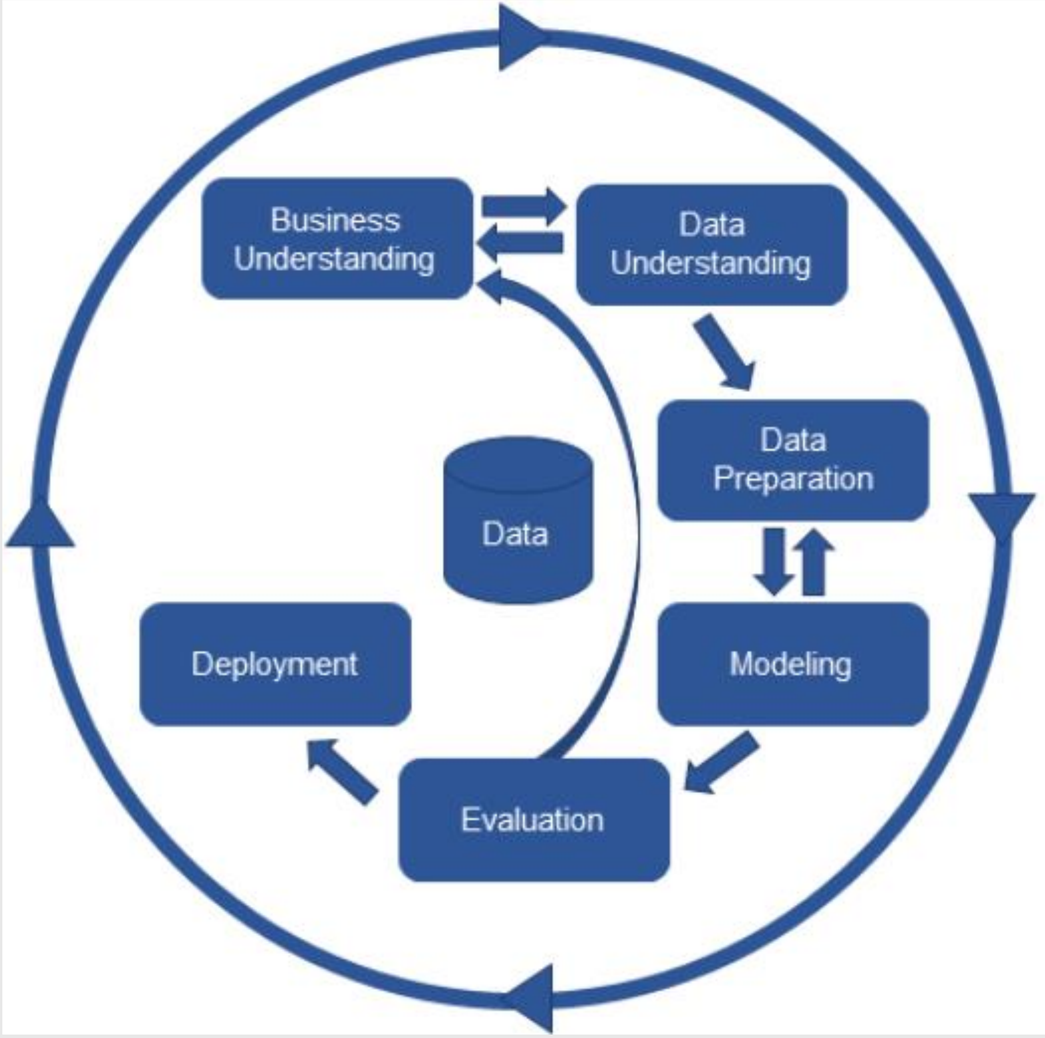
Case

Desenvolver um projeto de modelo Supervisionado

1. **Analise Exploratória**
2. **Propor um tratamento para amostragem**
3. **Treinar Modelos Preditivos**
4. **Avaliar Qualidade de Ajustes**
5. **Propor Tratamentos e um Desenho do Modelo “em Produção”**

- Base de dados de 2 dias de transações de cartões de crédito (Europa)
- Variáveis explicativas são componentes principais
- Somente Valor e Time estão com “formatação” original

Antes de Começar

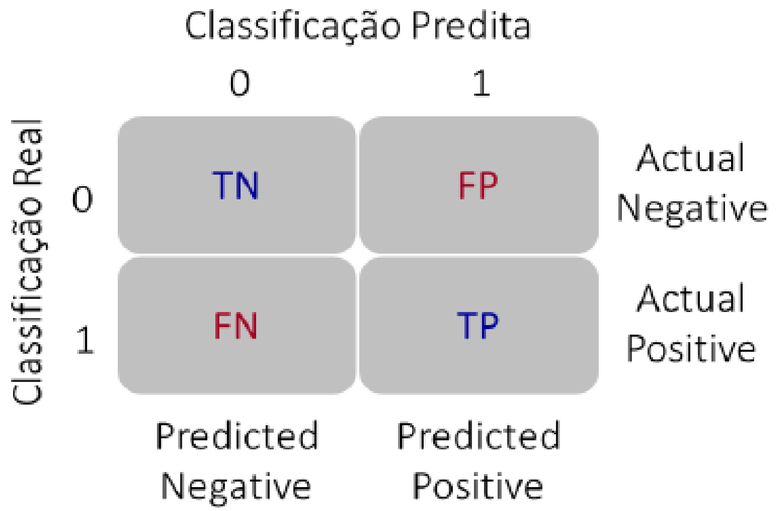


CRISP-DM

1. **Entendimento do Problema**
2. **Entendimento dos Dados**
3. **Preparação de Dados e Feature Engineering**
4. **Modelagem/ Treinamento dos Algoritmos supervisionados**
5. **Avaliação da Qualidade do Modelo**
6. **Deployment – “Colocar em produção”**

Conceito

Métricas de Acurácia – Modelos de Classificação



TN: True Negative
TP: True Positve
FN: False Negative
FP: False Positive

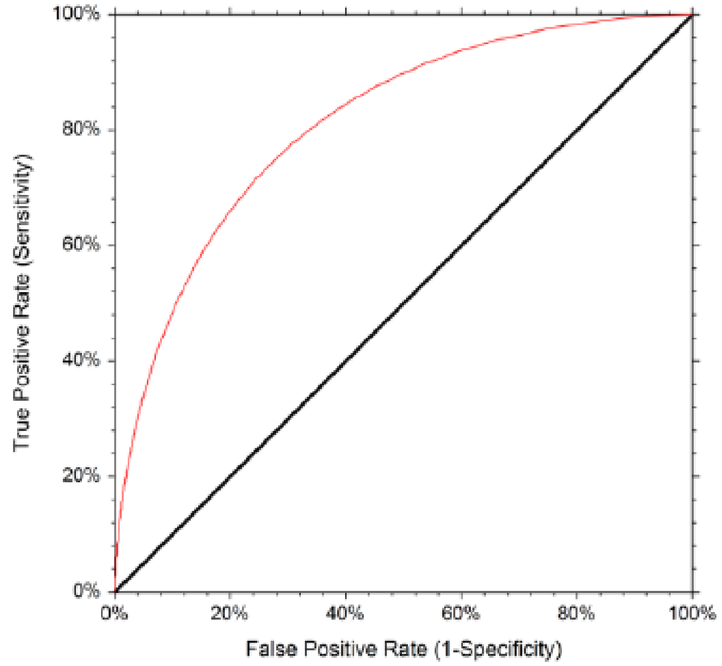
- ***Métricas*** para avaliar a qualidade do ajuste do modelo
  - **Missclassification** $= \frac{FP+FN}{Total\ de\ casos}$
  - **Acurácia** $= \frac{TP+TN}{Total\ de\ casos}$
  - **Precision** $= \mathrm{P} = \frac{TP}{TP+FP}$
    - Altos valores de precision estão relacionados a baixa taxa de FP
  - **Recall** $= \mathrm{R} = \frac{TP}{TP+FN}$
    - Altos valores de recall estão relacionados a baixa taxa de FN

- ***Conclusões:***
  - Alto recall e Baixo precision -> prejudica o cliente, pois o cliente era bom (0) e foi classificado como ruim (1).
  - Baixo recall e Alto precision -> beneficia o cliente, pois o cliente era ruim (1) e foi classificado como bom (0).
  - Altos valores de precision e recall são indicativos de um modelo bem ajustado

Conceito

Métricas de Acurácia – Modelos de Classificação

A curva ROC, mede, fração a fração, quantos 1's foram capturados (taxa de true positive) vs quantos 0's foram capturados (taxa de false positive).



- ***Métricas***

  - $\boldsymbol{Sensibilidade = Recall = \frac{TP}{TP+FN}}$
  - $\boldsymbol{Especificidade = \frac{TN}{TN+FP}}$

Prática

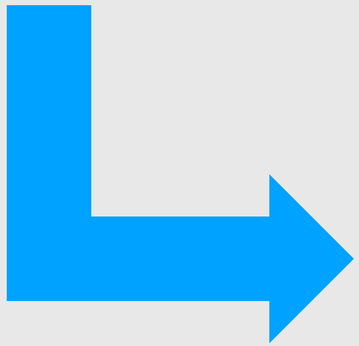
Fraude Transacional de Cartões de Crédito

- Analisar Evento e Variáveis Explicativas
- Ajustar Modelos de Classificação – KNN e Regressão Logística (opcional Decision Tree)
- Críticar métricas de acurácia

Conceito

Como lidar com o “Class Imbalance”

1. **Existe uma quantidade considerável de Clientes Maus ?**
2. **Diminuir sem viés a amostra da categoria dominante**
3. **Utilizar métricas de perfomance adequadas ou mais robustas**
4. Algoritmos Baseados em árvores de decisão (“Ensemble Trees”)



a) **Amostra Sintética - SMOTE - Synthetic Minority Over-sampling Technique**

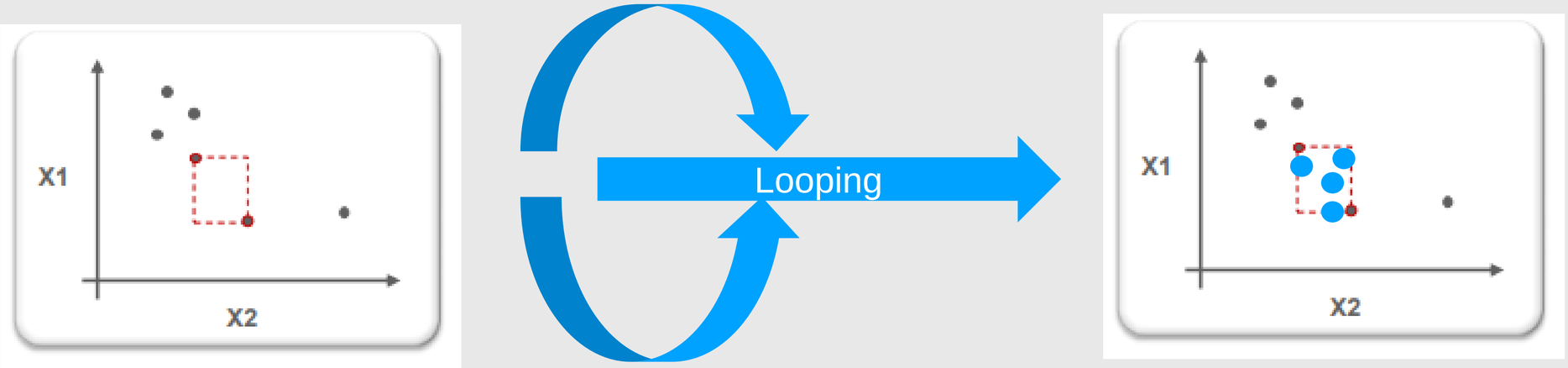
b) Algoritmos com parametros de penalização

c) Algoritmos de detecção de Anomalias

Conceito

SMOTE – Amostra Sintética

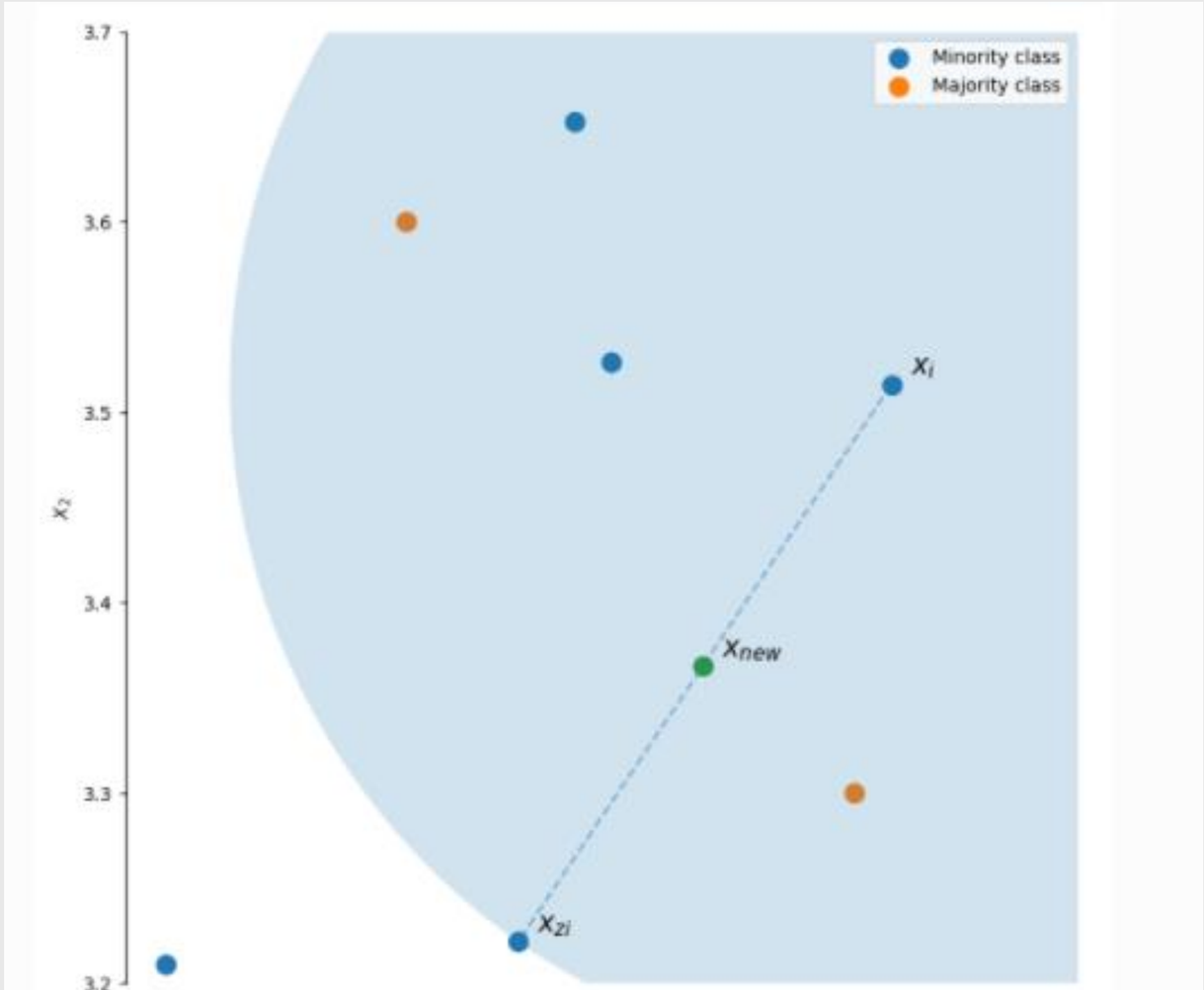
1. **Oversampling**
2. **Filosofia semelhante ao Método KNN**
3. **Novos dados da classe minoritária são gerados baseando se na relação das características/features da classe minoritária**



Conceito

SMOTE – Amostra Sintética – Algoritmo Regular

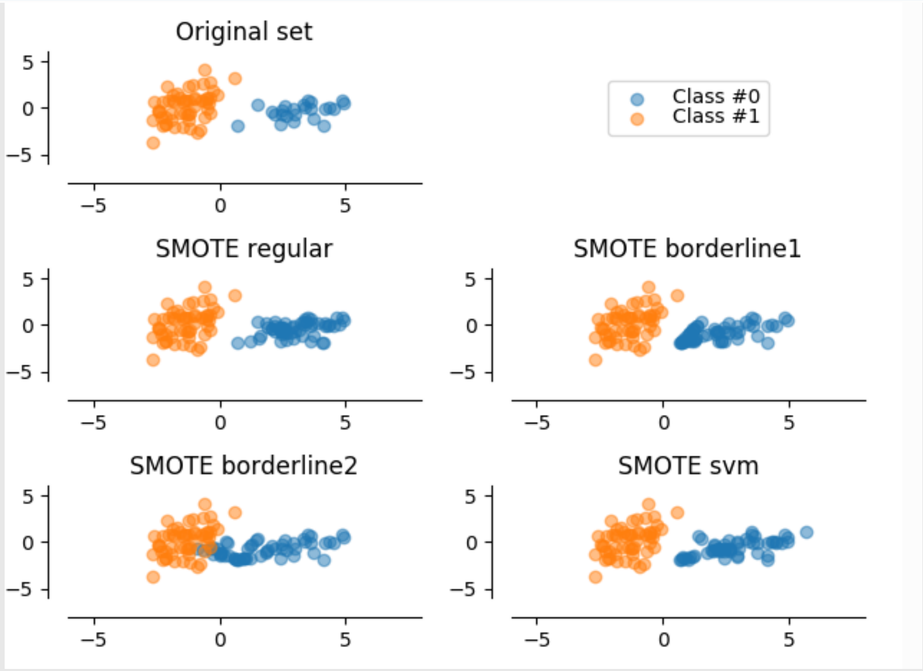
1. **Buscar K vizinhos mais próximos**
2. **Sortear aleatóriamente um Fator** λ
3. X_novo= $V_1 + \lambda(V_1 - V_2)$ (2 vizinhos)



Referêcia

SMOTE – Imbalanced Learning

class **imblearn.over_sampling.SMOTE**(*ratio='auto', random_state=None, k=None, k_neighbors=5, m=None, m_neighbors=10, out_step=0.5, kind='regular', svm_estimator=None, n_jobs=1*) [source] [source]



http://contrib.scikit-learn.org/imbalanced-learn/stable/generated/imblearn.over_sampling.SMOTE.html

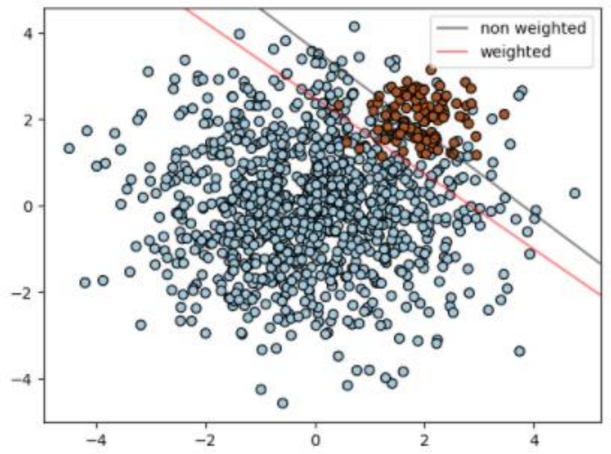
Conceito

Algoritmos com parâmetros de penalização

- **SVM**
- **SVC**
- **Logistic Regression**

In problems where it is desired to give more importance to certain classes or certain individual samples keywords `class_weight` and `sample_weight` can be used.

`SVC` (but not `NuSVC`) implement a keyword `class_weight` in the `fit` method. It's a dictionary of the form `{class_label : value}`, where value is a floating point number > 0 that sets the parameter `C` of class `class_label` to `C * value`.

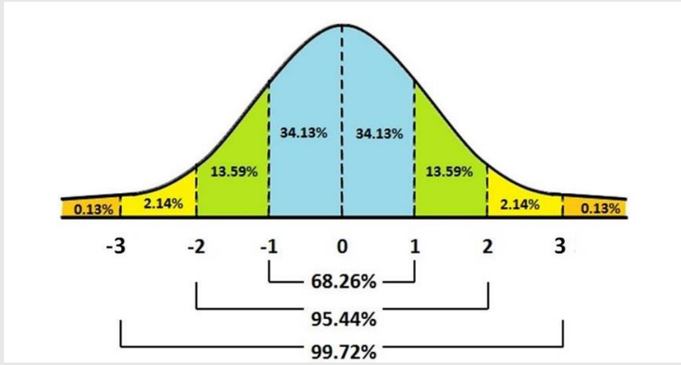


http://scikit-learn.org/stable/modules/svm.html
http://scikit-learn.org/stable/auto_examples/svm/plot_separating_hyperplane_unbalanced.html

Conceito(1/2)

Detecção de Anomalias

- **Cada Variável Explicativa (Feature) possuí uma média e desvio padrão**
- **Comparar a Probabilidade Normal da observação com a média –j e desvio padrão-j da Feature -j**
- $P_j(X_i = x_i) = \frac{1}{\sqrt{2\pi\sigma_j}} \times e^{\frac{-(x_i-\mu_j)}{2\sigma_j^2})}$
- **Para n-variáveis -> Prob =** $\mathrm{P}_{j=1} \times P_{j=2} \times \cdots \times P_{j=n}$
- **Existe uma fronteira vísivel** $Prob < \varepsilon$**?**
- **Alternativa -> Z-Score (Aula 2)** $\rightarrow Z_j = \frac{x_j-\mu_j}{\sigma_j}$ **-> Extremos valores de Zj são outliers**
- **Abordagens avançadas : Cálculo da Normal Multivariada**



http://colingorrie.github.io/outlier-detection.html
https://www.coursera.org/learn/machine-learning/lecture/Rkc5x/anomaly-detection-vs-supervised-learning

Conceito(2/2)

Mahalanobis Distance

```
from scipy.spatial.distance import mahalanobis
import scipy as sp
import pandas as pd

x = pd.read_csv('IrisData.csv')
x = x.ix[:,1:]

Sx = x.cov().values
Sx = sp.linalg.inv(Sx)

mean = x.mean().values

def mahalanobisR(X,meanCol,IC):
    m = []
    for i in range(X.shape[0]):
        m.append(mahalanobis(X.ix[i,:],meanCol,IC) ** 2)
    return(m)

mR = mahalanobisR(x,mean,Sx)
```

Detecção multivariada de outliers

- Distância de Mahalanobis

$$D^2 = (x_i - vetor_{médias})^T \Sigma^{-1} (x_i - vetor_{médias})$$

- Métodos de Cluster : analisar elementos fora dos clusters
- Métodos de regressão : ajuste linerar e busca pelos maiores erros ou gráfico de residuos
- Conselho Prático : Foque mais em outliers univariados

https://stackoverflow.com/questions/29817090/is-there-a-python-equivalent-to-the-mahalanobis-function-in-r-if-not-how-can

Prática

Fraude Transacional de Cartões de Crédito

- Detecção de Anomalias
- Calculem o Z-Score das Variáveis
- Filtre os valores anomalos de Saldo
- Criem 3 valores previstos através do Z-score (sugestão: |Zscore| > 1.96, |Zscore|> 2 e |Zscore| >3)
- Comparem Matriz de Confusão
- Multivariado, Selecionem 3 variáveis e repitam o processo acima

Conceito

Modelos “em produção”

- **Escrever a Regra de Escoragem sem risco operacional?**
- **PMML**
- **Feature Engineering em produção -> Cuidado com variáveis relativas**
- **Monitorar Distribuição das Variáveis Explicativas**

http://dmg.org/pmml/pmml_examples/index.html