SIMULADO DE PROVA

APRENDIZADO DE MÁQUINA SUPERVISIONADO

Conceitos Gerais

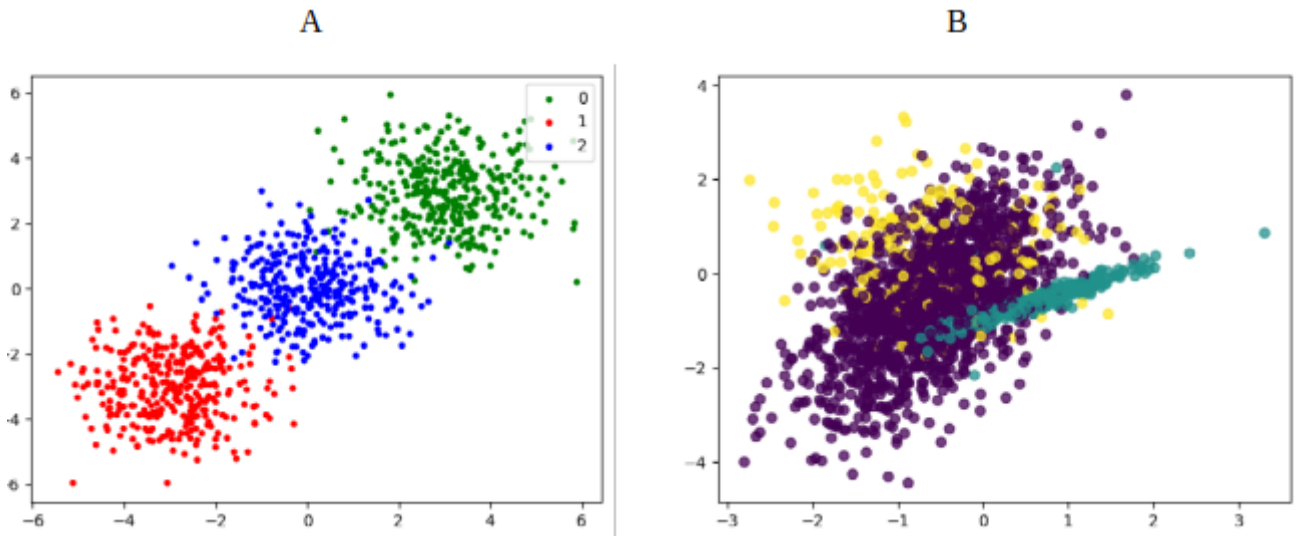
- O que é o aprendizado de máquina supervisionado?
- O que é aprendizado não-supervisionado?
- O que são atributos e classes?
- De um exemplo de uma instância (amostra) de um problema de classificação qualquer. Por exemplo, como você classificaria carros? E caes e gatos?
- O que significa a anotação dos dados?
- O que significa representativade em termos de características ?
- Dado o dataset abaixo, determine o que são atributos e o que são classes:

Feature 1	Feature 2	Feature 3	Cor	Tipo de Veículo
1500	110	30	Vermelho	Moto
2500	150	50	Azul	Carro
3500	220	70	Verde	Caminhão
1200	80	20	Preto	Moto
2800	160	60	Branco	Carro
4200	240	90	Vermelho	Caminhão
800	60	15	Preto	Moto
2000	120	40	Azul	Carro
3000	200	80	Verde	Caminhão

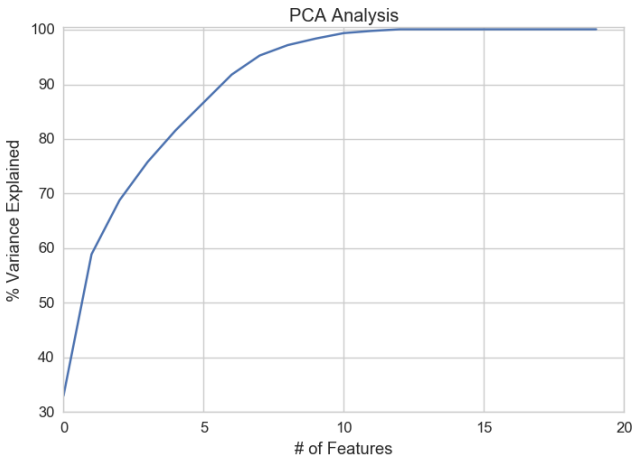
- Quais etapas possui um fluxo (pipeline) de aprendizado de máquina?
- O que é classificação binária e multi-classes?
- Como abordar classificação multi-classes a partir de modelos binários?

Análise Exploratória

- O que significa análise explotória dos dados? O que desejamos verificar com isso

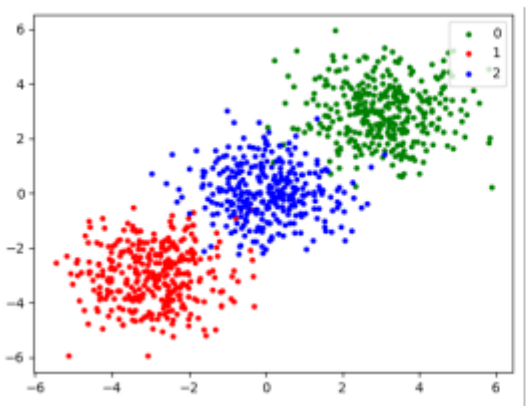
- O que são dados categóricos ou dados númericos?
- Como converto um dado categórico em númerico? De um exemplo
- Na questão 6 acima, converta o atributo 'Cor' em número
- O que é um dataset desbalanceado? O desbalanceamento é em termos de atributos ou classes?
- Que técnicas podem minimizar o impacto de dados desbalanceados? Quando utiliza-las?
- Na técnica Oversampling, o que significa interpolar as amostras? Ilustre um exemplo
- Análisando as distribuições abaixo (A e B):
    - Qual aprensenta as fronteiras decisão mais definidas?
    - Qual apresenta desbalanceamento de classes e porque?
- 

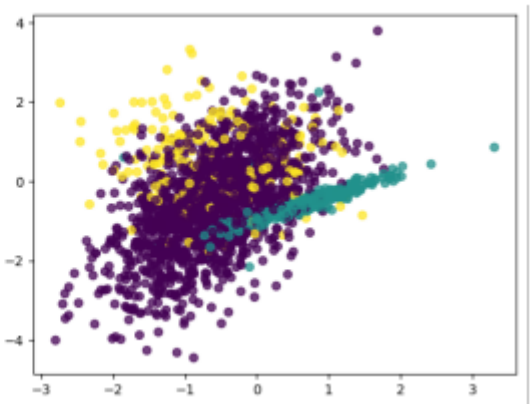
- O que é a redução de atributos, e quando podemos aplica-la?
- O que significa a correlação de atributos? De exemplos
- O que faz o algoritmo PCA?
- Como interpretar o grafico da variancia (PCA) abaixo ?



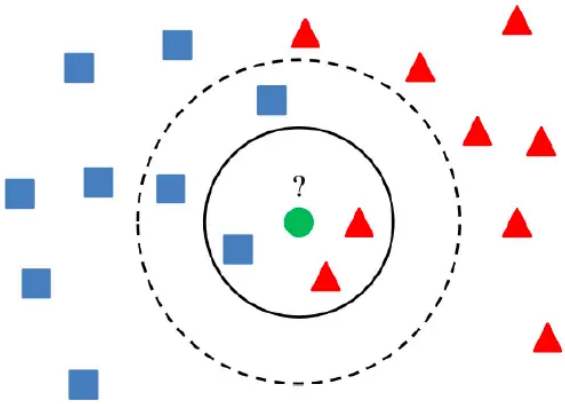
Algoritmo KNN

- Descreva em poucas linhas o algoritmo KNN. Se preferir, faça um desenho auxiliar e explique.
- O KNN funciona somente para 2 classes (binário) ?
- Qual a desvantagem do KNN em datasets grandes ? Por exemplo, com 100 mil amostras?
- O KNN reduz o espaço de características? E o espaço de busca? Porque?
- O que é o parametro K, do KNN? Como ele impacta na classificação?
- Analisando as seguintes distribuições, qual o algoritmo KNN deve performar melhor? Porque? Qual o impacto de um K maior e menor? O que você indicaria ?





- Qual a classe da amostra de teste para K=3 e K=5, abaixo?



- Considere o seguinte dataset e a amostra de teste abaixo:

Exemplo	Característica 1	Característica 2	Característica 3	Classe
1	2	3	1	A
2	1	2	0	A
3	3	0	2	B
4	0	1	3	B
5	2	2	3	A

Amostra de Teste	Característica 1	Característica 2	Característica 3
1	2	1	2

- Utilizando a distância Euclidiana abaixo, qual o resultado da amostra para K=1 e K=3?

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

Algoritmo Naive Bayes

- Explique de maneira sucinta como funciona o Naive Bayes?
- Explique sucintamente o teorema de bayes. Cite exemplos
- O que é uma probabilidade a posteriori e a priori ? Como isso é aplicado no Naive Bayes?
- Análise as matrizes de confusão abaixo e discuta cada caso, quanto a:
- Dado dataset abaixo:

ID	Cor	Tamanho	Tipo de fruta	Classe
1	Verde	Pequeno	Limão	Ácido
2	Amarelo	Médio	Abacaxi	Doce
3	Laranja	Médio	Laranja	Ácido
4	Amarelo	Pequeno	Banana	Doce
5	Verde	Grande	Melancia	Doce
6	Vermelho	Pequeno	Morango	Doce
7	Amarelo	Médio	Pêra	Doce
8	Laranja	Grande	Tangerina	Ácido
9	Amarelo	Pequeno	Limão	Ácido
10	Verde	Médio	Maçã	Doce
11	Verde	Pequeno	Limão	Ácido
12	Amarelo	Médio	Abacaxi	Doce
13	Laranja	Médio	Laranja	Ácido
14	Amarelo	Pequeno	Banana	Doce
15	Verde	Grande	Melancia	Doce
16	Vermelho	Pequeno	Morango	Doce
17	Amarelo	Médio	Pêra	Doce
18	Laranja	Grande	Tangerina	Ácido
19	Amarelo	Pequeno	Limão	Ácido
20	Verde	Médio	Maçã	Doce

    - Aplique o algoritmo Naive Bayes para determinar a probalidade e classes das amostras abaixo

ID	Cor	Tamanho	Tipo de fruta	Classe
1	Verde	Pequeno	Limão	Ácido
2	Vermelho	Médio	Maçã	Doce

- Como aplicar o modelo Naive Bayes em datasets númericos, como atributos como peso, altura, salario, etc? De exemplos.

Caso C:

Algoritmo Decision Tree

Elaboração das questões ainda em andamento

Análise Crítica

- O que é acuracia, recall?
- Porque a acuracia geral não é uma boa métrica? Dê um exemplo
- Calcule a accurácia do modelo a partir da matriz de confusão abaixo:

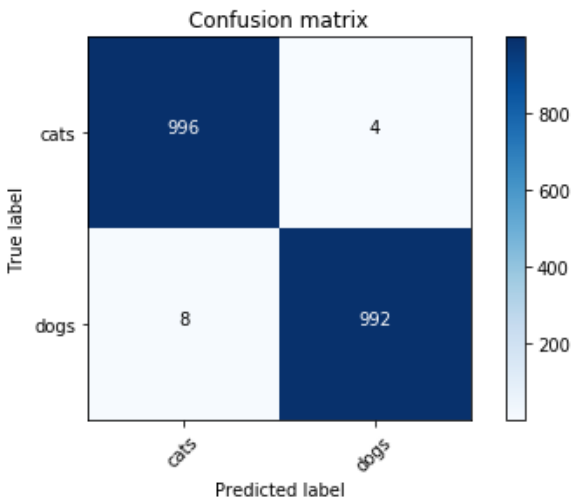
		Prediction	
		Cat	Dog
Actual	Cat	15	35
	Dog	40	10

- A seguir analise os casos individualmente quanto:
  - Qual a acurácia global ?

    O modelo está bem ajustado ou existe overfitting?

    O dataset pode ser considerado balanceado?

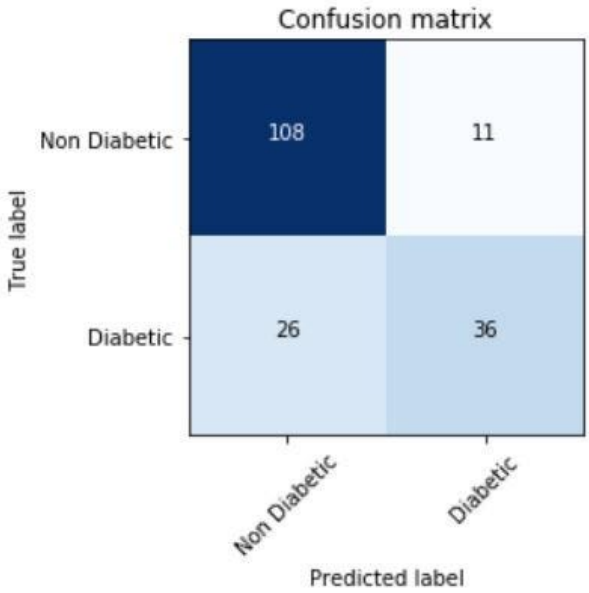
    Caso A:



    Caso B

n=192	Predicted: 0	Predicted: 1
Actual: 0	118	12
Actual: 1	47	15

Caso C:



Caso D:

	Actual positive (1)	Actual negative (0)
Predicted positive (1)	5	50
Predicted negative (0)	10	10000