Resolvendo Problemas de Machine Learning utilizando Estatística em Python e R

Laércio Jorge de Oliveira
Gerencia de Arquitetura Corporativa
GAC

prodemge

Agenda

1. Funções R e Python
2. Estatística e Probabilidade
3. Distribuição Normal
4. Teorema do Limite Central
5. Implementação
    1. KNN (K Nearest Neighbors)
    2. Utilizando Probabilidade

Funções R e Python (Média)

- Em R

```
media <- function(numeros) {
   sum(numeros)/length(numeros)
}
```

- Em Python

```
def media(numeros):
        return sum(numeros)/len(numeros)
```

Funções R e Python (Distância)

- Em R

```
distancia <- function(v1,v2) {
   dim <- length(v1); soma <- 0
   for(i in seq_len(dim -1)) {
      soma <- soma + (as.numeric(v1[i]) - as.numeric(v2[i]))^2
   }
   sqrt(soma)
}
```

- Em Python

```
def distancia(v1,v2):
   dim, soma = len(v1), 0
   for i in range(dim -1):
      soma += math.pow(float(v1[i]) - float(v2[i]),2)
   return math.sqrt(soma)
```

Estatística e Probabilidade

- Prever o resultado de uma eleição
- Prever se uma pessoa está doente ou não está doente
- Generalizando: prever a que classe irá pertencer um determinado elemento

Estatística e Probabilidade

- **Média**

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{X} = \sum_{i=1}^{n} x_i p(x_i)$$

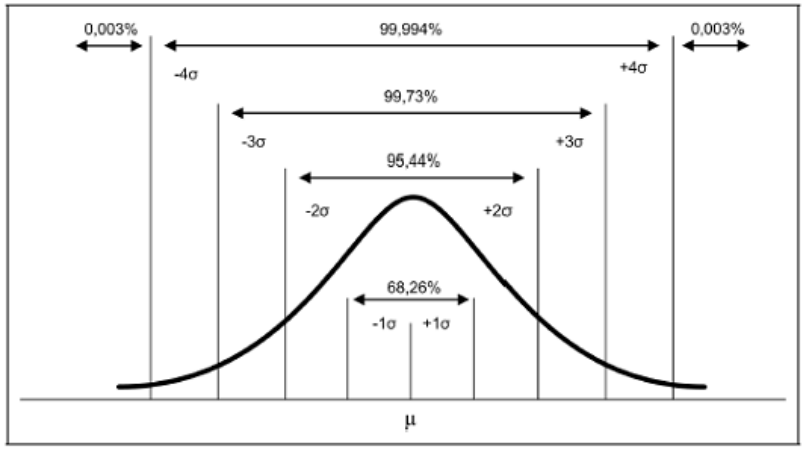
- **Variância e Desvio Padrão**

$$S^2 = \frac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$$

$$Sd(X) = Var(X)^{\frac{1}{2}}$$

Estatística e Probabilidade

- Distribuição Normal



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Teorema do Limite Central

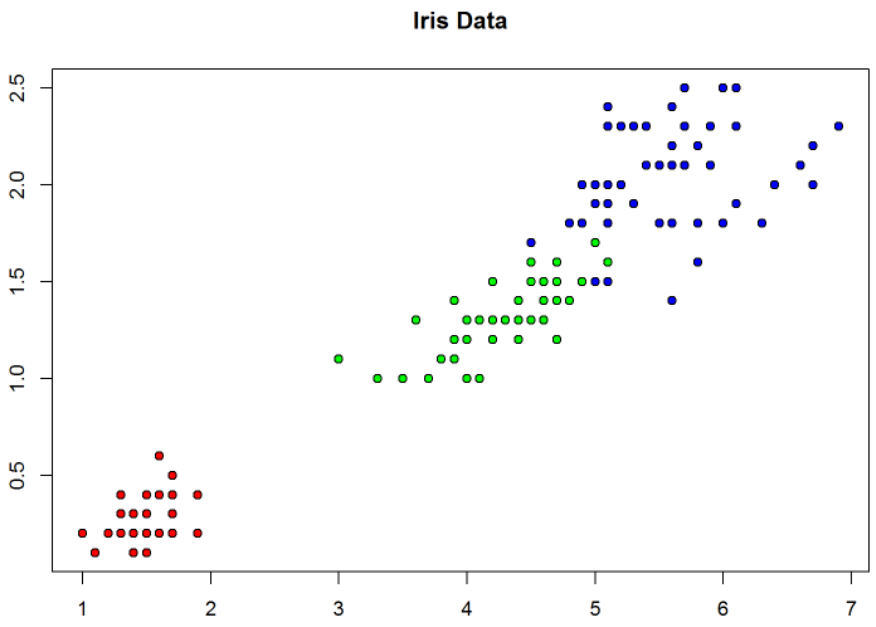
- **O Teorema do Limite Central afirma que a distribuição das médias se torna uma distribuição normal padrão à medida que o tamanho da amostra aumenta.**

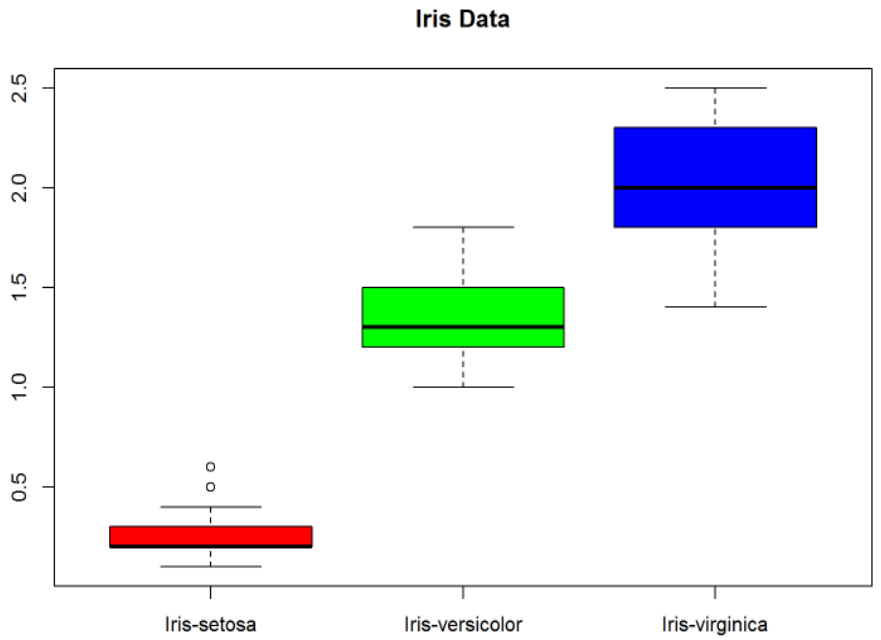
Estatística e Probabilidade

- Exemplo: Número de clicks diários de anúncios tem media=1020; sd=50. Qual é a probabilidade de mais de 1160 clicks?
  - **0.00255513**

- Exemplo: media=1020; sd=50. Qual o número diário de clicks que representa 75% dos dias que têm menos clicks?
  - **1053.724**

Análise analítica de dados





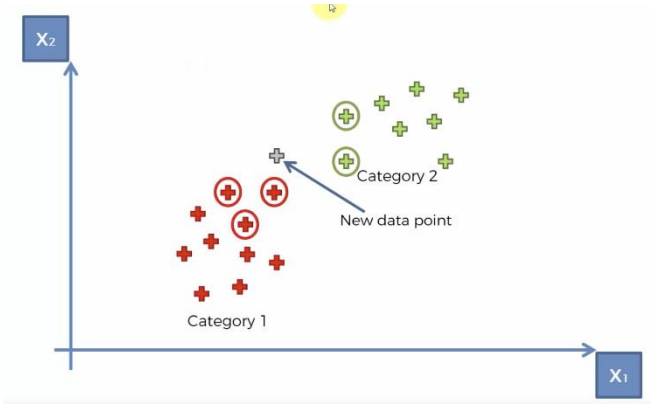
Implementação

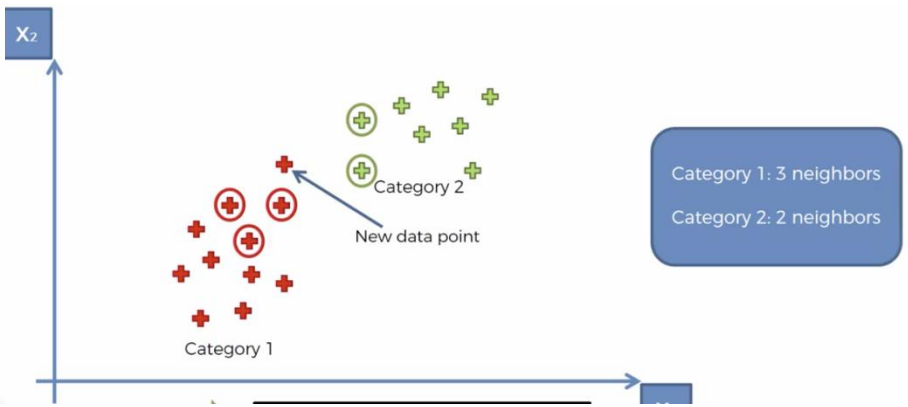
- **KNN (K Nearest Neighbors)**
  - **Utilizaremos a distância euclidiana**

- **Probabilidades**
  - **Utilizaremos a fórmula da Distribuição Normal e o Teorema do Limite Central**

Implementação

- KNN (K Nearest Neighbors)





Implementação

- **Probabilidades**
  - **Calcular as médias e desvios padrão de cada característica do Conjunto de Treinamento**
  - **Para cada entrada do conjunto de Testes, calcular a probabilidade de cada característica utilizando os cálculos acima**
  - **Multiplicar estas probabilidades para encontrar a probabilidade total**
  - **Seleciona a maior probabilidade**

Obrigado