Análise Exploratória de Dados – Roteiro de atividades 01 – 2024/01

Prof. Hugo Carvalho

14/05/2024

1 Introdução

Nessa atividade, nós iremos trabalhar com o conjunto de dados `iris`, que está embutido no `R`. Esse é um conjunto de dados multivariados[1] introduzido pelo estatístico e biólogo britânico Ronald Fisher em seu trabalho de 1936 "*O uso de múltiplas medições em problemas taxonômicos, como um exemplo de análise discriminante linear*". O conjunto de dados consiste em 50 amostras de cada uma das três espécies de flores *Iris* (*Iris setosa*, *Iris virginica* e *Iris versicolor*). Quatro variáveis foram medidas em cada amostra: o comprimento e a largura das sépalas e pétalas, em centímetros. Com base na combinação dessas quatro características, Fisher desenvolveu um modelo discriminante linear para distinguir as espécies umas das outras. Com base no modelo discriminante linear de Fisher, esse conjunto de dados se tornou um caso de teste típico para muitas técnicas de classificação estatística em aprendizado de máquina.[2] Exploraremos esse e outros métodos de classificação na disciplina de Aprendizado de Máquina. Aqui iremos fazer uma análise exploratória dessa base de dados.

2 Relembrando algumas quantidades

Aqui iremos relembrar algumas quantidades que vimos em sala de aula, adequadas para resumir dados numéricos. No que se segue, $x = \{x_1, \ldots, x_n\}$ irá denotar uma sequência de observações independentes da mesma característica para um total de n indivíduos – informalmente, x conterá uma coluna de um conjunto de dados.

2.1 Estatísticas de ordem

São simplesmente as posições das observações após a ordenação dos dados. Se $x_1, \ldots, x_n$ representam os dados na ordem em que aparecem na tabela (de cima para baixo, por exemplo) denotamos sua ordenação crescente por $x_{(1)}, \ldots, x_{(n)}$, ou seja, temos que $x_{(1)} \leq \cdots \leq x_{(n)}$. Assim, $x_{(1)}$, a menor observação, é dita a *primeira estatística de ordem*; $x_{(2)}$, a segunda menor observação, é dita a *segunda estatística de ordem*. De forma geral, $x_{(k)}$, a k-ésima menor observação, é dita a *k-ésima estatística de ordem*. O conjunto de dados x devidamente ordenado é obtido no `R` através do comando `sort(x)`, e a k-ésima estatística de ordem é obtida por `sort(x)[k]`, para $k = 1, \ldots, n$.[3]

2.2 Medidas de centralidade

- **Média:** $\overline{x}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$.

  A média representa o valor em torno do qual o conjunto de dados x está "orbitando". É sensível a observações discrepantes. Calculado no `R` através do comando `mean(x)`.

- **Mediana:** $\text{med}(x) = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ é ímpar;} \\ \frac{1}{2}\left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right], & \text{se } n \text{ é par.} \end{cases}$

  A mediana representa um valor tal que metade das observações estão abaixo dela e a outra metade está acima dela. É pouco sensível a observações discrepantes. Calculado no `R` através do comando `median(x)`.

2.3 Medidas de dispersão

- **Variância:** $\text{v}(x) = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$.

[1] Contendo múltiplas medições (colunas) para cada indivíduo (linha).
[2] Fonte: Wikipedia.
[3] Lembre-se que o `R` começa a contar do 1, enquanto que o Python começa a contar do 0.

A variância representa um *desvio médio das observações em relação à média*, sendo esses "desvios" calculados como a observação subtraída da média, posteriormente elevada ao quadrado, penalizando assim "grandes desvios" em relação à média. Como seu cálculo é uma média de quantidades, é sensível a observações discrepantes. Ademais, se as observações em x estão medidas na unidade $[U]$, então a variância é medida em $[U]^2$, uma unidade que pode não fazer sentido. Calculada no `R` através do comando `var(x)`.

- **Desvio padrão:** $\mathrm{dp}(x) = \sqrt{\mathrm{v}(x)}$

  O desvio padrão "normaliza" o problema da unidade de medida da variância, tendo a mesma unidade dos dados observados. Também é sensível a observações discrepantes, já que é uma métrica derivada da variância. Calculada no `R` através do comando `sd(x)`.

- **Desvio absoluto médio:** $\mathrm{dma}(x) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} |x_i - \overline{x}_n|$.

  Semelhante à variância, o desvio absoluto médio também representa um desvio médio das observações em relação à média, porém agora esses "desvios" são a observação subtraída da média, posteriormente sendo tirado o seu módulo. Assim, "grandes desvios" em relação à média não são penalizados. Sua unidade de medida é a mesma dos dados observados. No `R` não há uma função "simples" para calcular tal quantidade, mas ela pode ser obtida através do comando `mean(abs(x - mean(x)))`.

3 Atividade

O objetivo da atividade é estudar as medidas descritivas acima para os quatro atributos de interesse (largura e comprimento da pétala e sépala) dentro de cada uma das três espécies de flores (*setosa*, *virginica* e *versicolor*), com o objetivo de identificar qual variável pode ser mais útil na identificação de uma espécie. Para isso, siga o roteiro abaixo.

a) No `R` (pode ser local ou no Colab), carregue a base de dados com o comando `data(iris)`. Visualize o início da base através do comando `head(iris)`, para entender como a base está estruturada.

b) Como todos os quatro atributos são contínuos, faça histogramas de cada um deles para cada espécie de flor. O objetivo disso é determinar se algum dos quatro atributos é significativamente diferente para cada espécie de flor, de modo a auxiliar na identificação. Você pode acessar a coluna i para a espécie `especie` através do comando: `iris[iris$Species == especie, i]`, onde `especie` pode ser `'setosa'`, `'virginica'` ou `'versicolor'`. Qual dos quatro atributos você achou visualmente mais diferente dentre as três espécies de flor?

   Obs.: Pode ser conveniente fazer os três gráficos lado-a-lado para compará-los melhor. Isso pode ser feito através da sequência de comandos abaixo:

```
par(mfrow = c(3, 1))
hist(iris[iris$Species == 'setosa', i])
hist(iris[iris$Species == 'virginica', i])
hist(iris[iris$Species == 'versicolor', i])
```

   onde i representa alguma das colunas da tabela. Repare que essa visualização ***não*** *fica das mais belas, especialmente no Google Colab, mas dá para entender o que está acontecendo.*

c) Vamos verificar a intuição ganha no item b) através de medidas numéricas. Primeiro, calcule a média e a mediana de cada um dos atributos para cada espécie de flor. Há algum atributo para o qual a média ou mediana seja substancialmente diferente dentre as três espécies de flor?

d) Vamos refinar a comparação realizada no item c), através das medidas de dispersão. Calcule o desvio padrão e o desvio absoluto médio de cada um dos atributos para cada espécie de flor. Para o atributo que você identificou no item c), o que você pode dizer após calcular as medidas de dispersão?