Relatório sobre a atividade “5 - Prática: Estatística p/ Aprendizado de Máquina (I)”

Lucas Gabriel Arenhardt

1. Introdução

Nesse módulo são abordados variados conceitos envolvendo probabilidade e estatística e suas aplicações em Python.

2. Tipos de dados

2.1. Numéricos

São informações quantitativas que podem ser medidas e expressas em números. Pode ser dividido em duas categorias: contínuo e discreto. Os dados contínuos representam valores que podem assumir qualquer valor dentre um intervalo, ou seja, podem possuir casas decimais. Os dados discretos representam valor que são contáveis e não podem assumir valores fracionários.

2.2. Categóricos

Esse tipo de dado representa categorias ou grupos, não possuindo um valor numérico que permita operações aritméticas (não é quantitativo). Os números podem ser usados apenas para identificar as categorias.

2.3. Ordinais

É um intermediário entre os tipos numérico e categórico. Há itens separados por categorias, porém há uma hierarquia, ou seja, é possível realizar uma comparação entre os valores. Um exemplo seriam as “tier lists” onde os dados são organizados em categorias de melhor ou pior e seus intermediários.

3. Média, Mediana e Moda

3.1. Média

É a soma de todos os valores dividido pela quantidade de elementos.

3.2. Mediana

É o valor que separa a metade inferior da metade superior em um conjunto de dados ordenados. Caso o número de elementos no conjunto seja par, a mediana será a média entre os dois elementos centrais.

3.3. Moda

É o valor que aparece com mais frequência em um conjunto de dados.

4. Variância, desvio padrão e covariância

4.1. Variância

A variância é uma medida que indica o quanto os valores de um conjunto de dados estão próximos ou distantes da média. Valores mais baixos indicam um conjunto mais homogêneo. Por outro lado, valores mais altos de variância indicam maior heterogeneidade dos dados. Seu cálculo é dado por $\sigma^2 = \frac{\sum_{1}^{N}(xi-\bar{x})^2}{N}$, onde N é o número de elementos, x é a variável, $\overline{x}$ é a média e x_i corresponde ao elemento de índice i .

4.2. Desvio padrão

Também indica o quanto os valores de um conjunto de dados estão distantes da média, porém expressa a dispersão dos dados na mesma unidade de medida dos valores originais. Por exemplo, se os valores no conjunto estão expressos em metros, logo o desvio padrão também será em metro. O seu cálculo é simplesmente a raiz quadrada da variância.

4.3. Covariância

A covariância é uma medida estatística que indica a extensão e a direção da relação linear entre duas variáveis. Em outras palavras, a covariância ajuda a entender como duas variáveis variam juntas.

5. População e amostra

População refere-se ao conjunto completo de todos os elementos que estão sendo analisados, enquanto a amostra é uma parte selecionada dessa população.

6. Gráficos

6.1. Histograma

Um histograma é uma representação gráfica da distribuição de dados quantitativos. Nele, os valores são agrupados em intervalos, e a altura das barras indica quantos valores estão em cada intervalo.

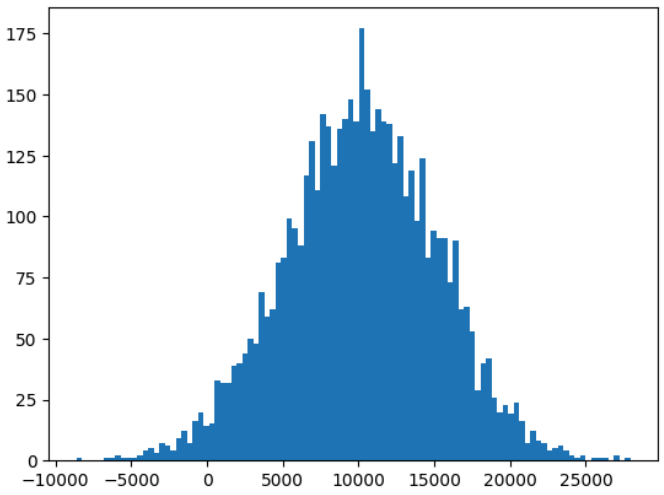

Imagem 1 - Exemplo de histograma.

6.2. Gráfico de distribuição normal

A distribuição normal, também conhecida como distribuição gaussiana, é uma das mais importantes na estatística. Ela possui um formato característico em forma de sino. O que a torna especial é que a média, a mediana e a moda possuem o mesmo valor nessa distribuição.

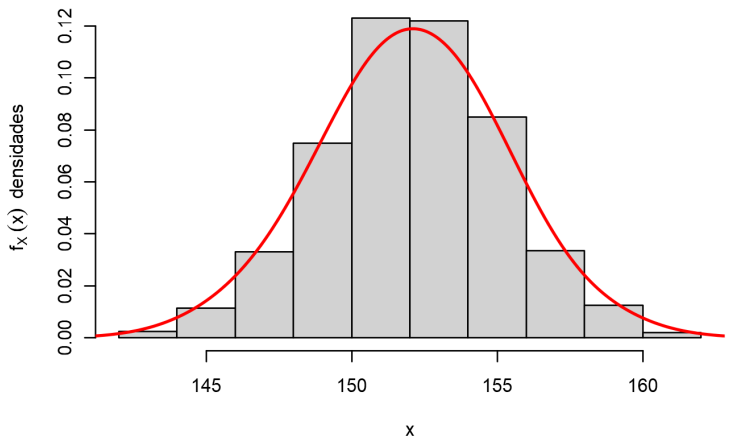

Imagem 2 - Exemplo de gráfico de distribuição normal ou gaussiana.

7. Percentil

Os percentis são os valores que dividem um conjunto de dados ordenados em cem partes iguais. Por exemplo, o valor 35 do percentil diz respeito a 35% dos menores dados do conjunto.

8. Momentos

O primeiro momento é a média. O segundo momento é a variância. O terceiro momento mede a assimetria da distribuição. O quarto momento é a curtose, responsável por medir o “achatamento” da função.

9. Teorema de Bayes

O Teorema de Bayes descreve a probabilidade de um evento baseado no conhecimento prévio de condições relacionadas ao evento.

A probabilidade de A com B é a probabilidade de A vezes a probabilidade de B com A, dividido pela probabilidade de B.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$