Relatório 5 - Estatística p/ Aprendizado de Máquina (I)

Yuri Vacelh Zamulhak Zdebski

Descrição da atividade

Assistir ao módulo Statistics and Probability Refresher and Python Practice do curso Machine Learning, Data Science and Generative AI with Python do instrutor Frank Kane, realizar as atividades propostas e desenvolver códigos autorais.

O curso pode ser dividido em 3 partes principais, estatística básica, aplicações utilizando python e por fim um pouco de probabilidade. O fato de começar pela estatística e depois ir para a probabilidade me chamou a atenção (foi o caminho contrario ao que tive contato até agora).

Estatistica Basica

Media Mediana e Moda

O curso começa classificando dados 3 tipos, numéricos, categóricos e ordinais. Os numéricos são dados quantitativos, indicam coisas como altura, tempo, preços ,que podem ser divididos em contínuos e discretos (os contínuos tem precisão infinita enquanto os discretos tem amostragem finita). Os categóricos são dados qualitativos, que não tem significado matemático intrínseco, literalmente categorias (raça, estado em que nasceu, sexualidade). Por fim os ordinais misturam os 2 primeiros, dando significado matemático (ou peso) para dados categóricos, como por exemplo dar notas a filmes de 1 a 5, sendo 1 que odiou e 5 que amou.

Logo em sequencia o curso já mergulha direto na estatística, definindo os conceitos de média, mediana e moda. A média é definida pela soma de todos os valores do conjunto divididos pelo tamanho do conjunto, por mais que eu queira dizer que é o valor do "meio", em verdade ela não é, tendo casos em que é deslocada por conta de valores extremos (muito maiores que os do conjunto), podendo levar a analises imprecisas. A formula para média é definida da seguinte forma:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A mediana é definida pelo valor do meio do conjunto ordenado, sendo assim, mais difícil de se alterar por discrepâncias do dataset, existem 2 casos para a mediana, em um numero impar de elementos, a mediana é o elemento central, para um numero par de elementos é calculada a média entre os 2 elementos centrais do conjunto. E por fim a moda é a medida estatística que indica qual valor mais se repete, não é relevante em dados contínuos. Vale lembrar que todas essas funções já estão implementadas nas bibliotecas estudadas.

Desvio padrão e variância

Seguimos então para o desvio padrão e variância, começando pela variância, que é definida pela seguinte formula:

$\sigma^2 = E[[X - E[X]]^2]$

A variância é uma medida da dispersão estatística dos dados ou seja, o quão longe eles vão, ou o quão diferentes eles são, entre si. Já o desvio padrão também denota o quanto os dados se dispersam, porém em relação a média, ele é definido da seguinte forma:

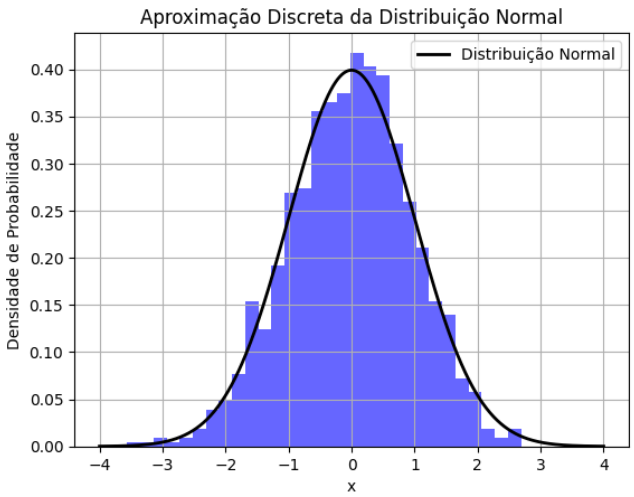
$\sigma = \sqrt{\sigma^2}$

É possível calcular essas métricas tanto para populações (o conjunto completo de dados) ou para amostras (subconjunto da população, literalmente uma amostra.

Função de densidade de probabilidade e de massa de probabilidade

A função densidade de probabilidade (PDF) é uma representação contínua que descreve a distribuição de probabilidades de uma variável aleatória. Por outro lado, a função massa de probabilidade (PMF) é a versão discreta da PDF, utilizada para descrever a distribuição de probabilidades de variáveis aleatórias discretas, na imagem abaixo isso fica mais claro

Figura 1: PDF/PMF



Fonte: autoria própria

Distribuições de dados comuns

Na estatística, distribuições de dados descrevem o comportamento de variáveis aleatórias (continuas e discretas), algumas das principais são:

Uniforme

Distribuição que, como o nome indica, a probabilidade de cada caso é igual.

Gaussiana

Provavelmente a distribuição mais famosa (e inclusive a que utilizamos até agora), seus dois parâmetros mais importantes são o μ (onde está centralizada, ou a média) e σ (desvio padrão). Tem pdf definida como:

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

Exponencial

Novamente, como o nome indica, é uma distribuição continua que tem como curva uma exponencial, pode crescer ou decrescer.

Binomial

A distribuição binomial modela situações onde há exatamente dois resultados possíveis para cada experimento, geralmente classificados como "sucesso"e "fracasso". Ela descreve o número de sucessos em uma sequência de n experimentos independentes, onde cada experimento tem uma probabilidade fixa p de resultar em sucesso. Essa distribuição é chamada de "binomial"devido à sua natureza dicotômica, refletindo os dois cenários possíveis em cada experimento.

Poisson

A distribuição de Poisson é uma distribuição de probabilidade discreta que descreve a probabilidade de um número fixo de eventos ocorrer em um intervalo de tempo ou espaço, assumindo uma taxa média constante de ocorrência e que os eventos são independentes entre si.

Porcentagens e Momentos

A questão da porcentagem é relativamente simples, em dada distribuição, qual o valor x em que x é maior que $n\%$ dos dados, o exemplo dado no curso é o da distribuição de renda, a partir de quantos dólares uma pessoa está acima dos 90% da população.

Os momentos são certas medidas quantitativas relacionadas à forma do gráfico da função de probabilidade, sendo eles:

- Primeiro momento: a média
- Segundo momento: a variância
- Terceiro momento: a assimetria do grafico (mais especificamente o quanto ele se distorce, para direita ou para a esquerda)
- Quarto momento: a curtose (o quão alto e pontuda é uma distribuição, em relação a uma curva normal)

Plots basicos com python

Nessa parte do curso, são apresentadas as bibliotecas gráficas Matplotlib e Seaborn e como as utilizar. Ambas fazem praticamente a mesma coisa, porém a Seaborn é uma "repaginação"da Matplotlib, trazendo elementos mais bonitos e modernos para seus gráficos (além claro de algumas funções extras).

Probabilidade básica

Na parte final do curso, nos é dada uma revisão de probabilidade, trazendo primeiramente a formula da probabilidade condicionada, que indica a probabilidade de um evento B sabendo que outro evento A aconteceu, definida por:

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

Sendo que,

- P(A) a probabilidade de A acontecer
- P(A,B) a probabilidade de A e B acontecerem
- P(B|A) a probabilidade de B sendo que A já aconteceu

Também somos apresentados ao teorema de Bayes que é um dos princípios fundamentais da probabilidade e da estatística, descreve a probabilidade de um evento com base em um conhecimento a priori que pode estar relacionado a esse evento, ele é definido por:

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

Sendo que,

- P(A|B) é a probabilidade de A ocorrer sendo que B ocorreu
- P(B|A) é a probabilidade de B ocorrer sendo que A ocorreu
- P(A) a probabilidade de A ocorrer
- P(B) a probabilidade de B ocorre

Conclusões

Neste curso foram mostradas aplicações das bibliotecas anteriormente introduzidas, possibilitando uma melhor utilização delas. Com os conhecimentos adquiridos, apliquei as ferramentas aprendidas em uma análise de um conjunto de dados climáticos obtido no Kaggle (o mesmo dataset utilizado no módulo 4). Senti que a analise ficou mais fluida do que no módulo 4, realmente conseguindo pensar em possíveis relações entre os dados que antes não eram tão claras.