**Relatório 5 - Estatística p/ Aprendizado de Máquina**

Guilherme Loan Schneider

**Descrição da atividade**

O módulo começa com uma introdução aos tipos de dados que podem ser encontrados, sendo eles numéricos, categóricos e ordinais. O primeiro deles é o mais comumente encontrado, se tratando de valores representados por números, sejam eles inteiros ou decimais, e podem ser classificados de duas formas, dados discretos (São contáveis e limitados a valores específicos) e dados contínuos (Podem assumir qualquer valor dentro de um intervalo, incluindo frações e decimais).

Em seguida, os valores categóricos, que como o próprio nome sugere, não tem relação com números, ou seja, no nosso cenário, se assimilarão as *strings*, como gênero, código de uma loja, estado, modelo de um carro, dentre outros.

Por fim, os valores ordinais, que indicam ordem de algo, ou seja, são os valores de avaliação de algum *videogame* por exemplo, partindo de 1 até 5, onde 1 é a pior avaliação e 5 a melhor. Esse tipo de dado é uma combinação dos citados anteriormente, utilizando valores inteiros ou decimais, no intuito de ordenar algo de certa forma.

Além dessa introdução aos tipos de dados, é explicado também a média, mediana e moda, que já é de familiaridade de todos.

* A média é a soma de valores dividido pela quantidade de valores somados.
* A mediana é ordenar os dados de forma crescente, e acessando o valor central. Caso seja um número ímpar de valores, é somado os dois elementos centrais e dividido por 2.
* A moda consiste em verificar qual o valor que mais se repete em um conjunto de dados.

Na figura abaixo, é mostrado o histograma do salário de dez mil indivíduos a fim de ser utilizado a moda, mediana e média. Os dados foram criados a partir da função np.random.normal(), utilizando a np.random.seed(150).

Gráfico, Histograma

Descrição gerada automaticamente

A mediana do salário desses indivíduos é aproximadamente a 26.795,176, e a média é de 26.726,182.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Agora, se nós adicionarmos apenas um indivíduo que ganha um salário absurdamente maior, a mediana não será alterada, no entanto, a média sofrerá uma grande diferença.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Como é possível perceber, a média é aproximadamente a 126.713,511 e a mediana se manteve a mesma.

Já na utilização da moda, será feita a importação da biblioteca scipy e a definição de um conjunto de dados das idades de 500 indivíduos utilizando a função np.random.randint().

Logotipo

Descrição gerada automaticamente com confiança média

Como resultado, obteve-se que o valor 45 repetiu 13 vezes nesse conjunto de dados.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Variância e Desvio Padrão

A variância é uma forma de valorar quão disperso estão os dados de determinado conjunto. Para isso, determina-se o valor médio das diferenças quadradas da média.

Exemplo:

Qual é a variância do seguinte conjunto de dados: (3,8,5,10,8,2).

1. Encontrar a média: (3+8+5+10+8+2) /6 = 6;
2. Encontrar as diferenças da média: (-3, 2, -1, 4, 2, -4);
3. Aplicar a diferença quadrada: (9, 4, 1, 16, 4, 16);
4. 2 = (9+4+1+16+4+16) /6 = 8,333.

O desvio padrão (DP) é calculado a partir da variância, ao ser a raiz quadrada desse parâmetro. Quando todos os valores de uma amostra são iguais, o desvio padrão é igual a 0. Sendo que, quanto mais próximo de 0, menor é a dispersão dos dados.

DP = 2

DP =

DP 2,8867

É importante salientar também que, caso o dataset não seja o conjunto total de dados, deve-se subtrair 1 do divisor na fórmula da variância.

Função de densidade de probabilidade

É uma função que descreve a probabilidade de uma variável aleatória contínua assumir um determinado valor.

Gráfico, Histograma

Descrição gerada automaticamente

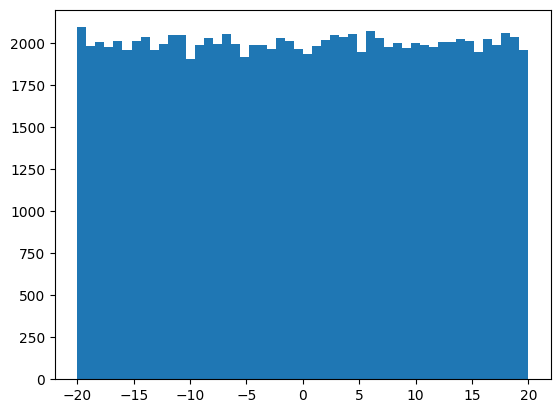
Função de massa de probabilidade

Gráfico, Histograma

Descrição gerada automaticamente

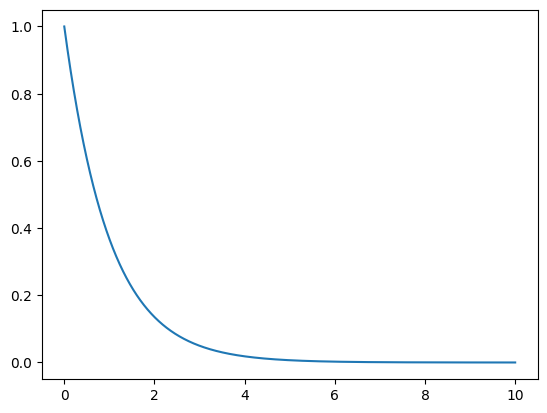
É um conceito fundamental da teoria das probabilidades que descreve a distribuição de probabilidades de variáveis aleatórias discretas.

Distribuição Uniforme



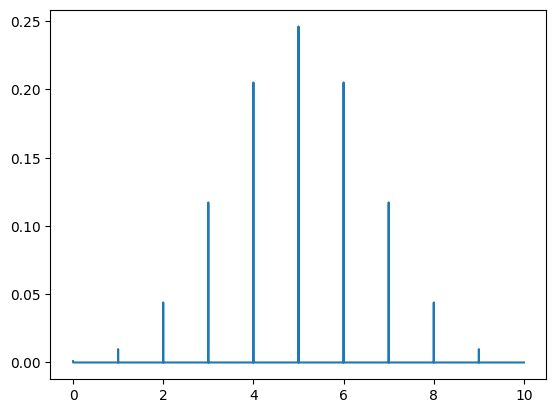
Nesse modelo, a probabilidade de gerar qualquer ponto em um intervalo contido no espaço amostral é proporcional ao tamanho do intervalo.

Distribuição Exponencial



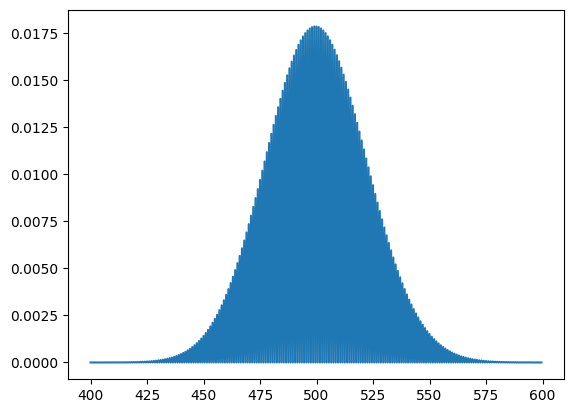
Essa distribuição contínua pode ser utilizada para descrever as probabilidades envolvidas no tempo que decorre para que um determinado evento aconteça.

Distribuição Binomial



É utilizada para calcular a probabilidade de um resultado específico, dado o número de tentativas e a probabilidade de sucesso em cada tentativa.

Distribuição de Poisson



Ela descreve a probabilidade de um evento ocorrer durante um determinado intervalo de tempo, quando a probabilidade de um evento ocorrer é bem baixa e a quantidade de tentativas é bastante grande.

Percentil

Tem-se uma pequena introdução à utilização de percentil na análise de dados, aplicado em alguns casos fictícios, como um histograma do salário anual de indivíduos, analisando o percentil de indivíduos que ganham acima de determinado valor.

Já no Python, é possível utilizar a função percentile(<conjunto\_de\_dados>, Y), onde Y é um valor, em porcentagem, e será retornado um valor que é maior que Y% dos dados.

Momentos de uma função

Os momentos de uma função são certas medidas quantitativas relacionadas à forma do gráfico da função. São divididos em quatro ordens:

Momento de primeira ordem: é a média ou valor esperado da função. No contexto de uma distribuição de probabilidade, o momento de primeira ordem indica o ponto médio dos dados.

np.mean(<conj\_dados>)

Momento de segunda ordem: é a variância, que mede a dispersão em torno da média. A variância é essencial para entender a "largura" ou "espalhamento" da função.

np.var(<conj\_dados>)

Momento de terceira ordem: mede a simetria ou assimetria da função (chamado de skewness em inglês). Um valor positivo indica uma cauda mais longa à direita, enquanto um valor negativo, uma cauda à esquerda.

import scipy.stats as sp

sp.skew(<conj\_dados>)

Momento de quarta ordem: mede a curtose, ou seja, o grau de "achatamento" ou "pontiagudez" da distribuição. Distribuições com alta curtose têm picos mais acentuados e caudas mais longas.

import scipy.stats as sp

sp.kurtosis(<conj\_dados>)

MatPlotLib

Nessa seção teve-se uma introdução a plotagem de diferentes tipos de gráficos, bem como alterações visuais, ajustes de eixos, dentre outros. Os principais tópicos tratados nessa seção foram os seguintes:

Os arquivos enviados no GitHub possuem maior detalhamento de cada tópico abordado

* Criar vários gráficos dentro de um só, como inserir a média de valores em um histograma de vendas de um produto;
* Salvar um gráfico gerado em um arquivo ‘.png’;
* Alteração das informações contidas nos eixos, bem como a escala do gráfico;
* Adicionar grade no fundo do arquivo;
* Alteração visual das linhas, como cor, estilo da linha, grossura;
* O estilo XKCD, que consiste como se fosse um gráfico desenhado à mão, como em um cartoon.
* Gráfico de pizza e barra, onde no primeiro foi utilizado a função explode, que destacou uma seção em específico;
* Gráfico de dispersão utilizando a função randn();
* Histograma, demonstrando o salário centrado em 27000 no exemplo da aula, com desvio padrão de 15000 e 10000 amostras.
* Por fim o Box e Whisker Plot, que é um método gráfico para representar a variação de uma variável ao longo de um conjunto de dados.

Seaborn

Todo o desenvolvimento dessa seção está detalhado nos arquivos enviados no GitHub, bem como imagens,comentários e atividades

Nessa seção teve-se uma introdução a plotagem de mais alguns gráficos utilizando a biblioteca Seaborn, que como o autor das aulas disse, é um MatplotLib++, onde é possível fazer visualizações mais agradáveis e personalizações nos gráficos. Abaixo estão listados os conteúdos abordados nessa seção do curso:

* Como primeiro item, utilizou-se um banco de dados de consumo de veículos.
* Em seguida, utilizou-se um gráfico de barras, já utilizando o Seaborn, para visualizar a quantidade de veículos com determinada quantia de marchas.
* Utilizou-se também um histograma com uma linha representando a média dos valores.
* Uma matriz de gráficos correlacionados chamado de pairplot, permitindo uma análise de várias informações combinadas.
* Gráfico de dispersão, que permite analisar a organização dos dados.
* O jointplot, que são gráficos combinados com um histograma, os gráficos podem ser de vários tipos, basta alterar o parâmetro “kind”.
* Lmplot, que basicamente consiste em um gráfico de dispersão com uma linha de regressão linear.
* O boxplot, que é o Box e Whisker Plot do MatplotLib. Tem o objetivo de analisar a variação de determinado dado.
* Swarmplot – Esse gráfico é um scatterplot que não deixa os pontos se sobreporem.
* Countplot - O countplot é utilizado para contar a quantidade de ocorrências de um determinado valor
* Heatmap – Como o próprio nome já diz, é um “mapa de calor”, onde é utilizado para visualizar a relação entre duas variáveis categóricas e uma variável numérica
* Por fim, o autor disponibilizou uma atividade no final da aula, permitindo a manipulação dos dados de marchas de veículos com o consumo combinado.

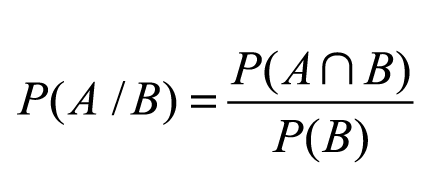
Covariância

A covariância permite analisar quão dependente uma variável é da outra, ou seja, se uma delas tente a aumentar e a outra tende a aumentar também, então, elas possuem covariância positiva, caso uma delas tente a aumentar e a outra não, a covariância é negativa.

Para se analisar a covariância, utiliza-se a correlação, que consiste em dividir a covariância pelo desvio padrão de ambos os conjuntos de dados. Se o valor obtido é igual a -1, é possível afirmar que a correlação entre esses valores é perfeitamente inversa, ou seja, quando um valor aumenta, o outro diminui. Caso o valor seja igual a 1, existe a correlação perfeita, onde quando um valor aumenta, o outro aumenta também. Por fim, caso seja próximo de 0, indica que não existe correlação entre os dados.

Probabilidade Condicional

Consiste em analisar a chance de algo ocorrer, contando que um outro acontecimento anterior já ocorreu (Chance de B acontecer, sendo que A já ocorreu).



É normalmente utilizada em sites de compras, avaliando a chance de um indivíduo comprar o produto B dado que ele está comprando um produto similar A, estudos de comportamento de usuários, determinando as chances de ele acessar algum tipo de recurso dentro do site, dentre outros.

Teorema de Bayes

Esse teorema descreve a probabilidade de um evento, baseado em um conhecimento inicial (a priori) que pode estar relacionado ao evento. Um exemplo clássico é o problema de Monty Hall (problema das portas).

Texto

Descrição gerada automaticamente com confiança média

É importante para verificar que a probabilidade de um evento B ocorrer dado um evento A, não implica necessariamente que um evento A ocorrer dado um B tem a mesma probabilidade.

**Conclusões**

Nesse módulo, foi possível compreender os conceitos fundamentais de análise de dados, abrangendo desde tipos de dados (numéricos, categóricos e ordinais) até conceitos estatísticos como média, mediana, moda, variância e desvio padrão. Também explora distribuições de probabilidade (uniforme, binomial, exponencial, entre outras), percentis e momentos estatísticos.

Além disso, detalha o uso de ferramentas como Matplotlib e Seaborn para visualização de dados, com exemplos de gráficos como histogramas, gráficos de dispersão, boxplots e heatmaps, destacando as possibilidades de personalização e análise visual. Foram abordados também conceitos de estatística avançada, como covariância, correlação e probabilidade condicional, culminando no Teorema de Bayes, usado para análise de eventos baseados em informações prévias.

**Referencias**

[001 🎥 Types of Data (Numerical, Categorical, Ordinal)](https://drive.google.com/open?id=1xkBJZSf_o0ts6lVr6aLJw6WNCpdSq5AM)

[002 🎥 Mean, Median, Mode](https://drive.google.com/open?id=1q7z7PNpE5GDTzXZIQncgGNlm9xa6_2v6)

[003 🎥](https://drive.google.com/open?id=1Szz6o6ZNxWqP7dOJejMVR-KuBmjxjKHV) **[Activity](https://drive.google.com/open?id=1Szz6o6ZNxWqP7dOJejMVR-KuBmjxjKHV)** [Using mean, median, and mode in Python](https://drive.google.com/open?id=1Szz6o6ZNxWqP7dOJejMVR-KuBmjxjKHV)

[004 🎥](https://drive.google.com/open?id=1zAnfhPQ4y3RNrgvkWVvq4wMYL0xebEVi) **[Activity](https://drive.google.com/open?id=1zAnfhPQ4y3RNrgvkWVvq4wMYL0xebEVi)** [Variation and Standard Deviation](https://drive.google.com/open?id=1zAnfhPQ4y3RNrgvkWVvq4wMYL0xebEVi)

[005 🎥 Probability Density Function; Probability Mass Function](https://drive.google.com/open?id=12g0_cIFv_ONYCFaksLoDMjhWbZTEmcw1)

[006 🎥 Common Data Distributions (Normal, Binomial, Poisson, etc)](https://drive.google.com/open?id=1vVoF7oFF-bFgdYuEe4PKgt7S_jXkKoTN)

[007 🎥 **Activity** Percentiles and Moments](https://drive.google.com/open?id=1cTLYlXYs8ZfeoWOUZ_BgFo0b16mqY3qk)

[008 🎥](https://drive.google.com/open?id=1Ef1zw9_Ls0u5YVNTD537tpajhz5V8g_P) **[Activity](https://drive.google.com/open?id=1Ef1zw9_Ls0u5YVNTD537tpajhz5V8g_P)** [A Crash Course in matplotlib](https://drive.google.com/open?id=1Ef1zw9_Ls0u5YVNTD537tpajhz5V8g_P)

[009 🎥](https://drive.google.com/open?id=1-UJroO48iLpnjD3ORa8tYJ8T9zAxbiMJ) **[Activity](https://drive.google.com/open?id=1-UJroO48iLpnjD3ORa8tYJ8T9zAxbiMJ)** [Advanced Visualization with Seaborn](https://drive.google.com/open?id=1-UJroO48iLpnjD3ORa8tYJ8T9zAxbiMJ)

[010 🎥](https://drive.google.com/open?id=1UJJCK-4tWkTsX6BSdz2FnFWP8cleOsX3) **[Activity](https://drive.google.com/open?id=1UJJCK-4tWkTsX6BSdz2FnFWP8cleOsX3)** [Covariance and Correlation](https://drive.google.com/open?id=1UJJCK-4tWkTsX6BSdz2FnFWP8cleOsX3)

[011 🎥 **Exercise** Conditional Probability](https://drive.google.com/open?id=1EgFDVNDDIas7ud_fuKecQBL6diLUHJwM)

[012 🎥 Exercise Solution Conditional Probability of Purchase by Age](https://drive.google.com/open?id=1x3DQgCihVeNw_iWzsLsYE1dPa3gdyLAG)

[013 🎥 Bayes' Theorem](https://drive.google.com/open?id=1tu4NQCTifK4SsQk9zqAcgT1DydXD0zZg)