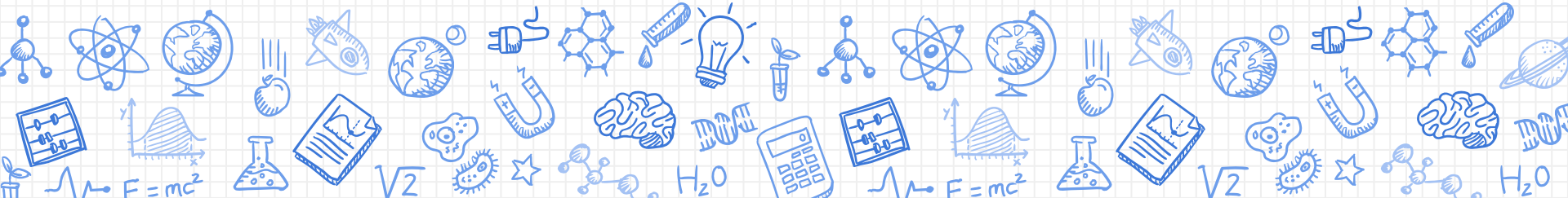
Análise de Dados e Inferência Estatística com Python

Aula de hoje!

- ✘ Introdução
- ✘ Matplotlib
- ✘ Probabilidade e algumas aplicações em Python

Introdução



Introdução

A Visualização de Dados é uma vertente essencial da Análise de Dados. Visualizações inteligentes permitem novas descobertas, *insights* e conhecimento através de gráficos e tabelas selecionados.

Existem muitas ferramentas para Visualização de Dados, e a cada dia novas ferramentas surgem. Nós iremos focar na ferramenta que é provavelmente a mais difundida para este fim no Python, que é o **matplotlib**.

Matplotlib

O matplotlib (http://matplotlib.org/) é uma biblioteca do Python para criação de gráficos em 2D, bastante utilizada para visualização de dados e que apresenta uma série de possibilidades gráficas, como gráficos de barra, linha, pizza, histogramas, entre muitos outros.

O matplotlib já está incluso na distribuição Anaconda (e em praticamente todas as distribuições de Python focadas em Análise de Dados e programação científica). Então, se você a está utilizando, não há necessidade de qualquer instalação.

Traçando os primeiros gráficos

Bem, vamos começar com um exemplo bem simples. Em primeiro lugar, é uma convenção usualmente adotada importar a coleção de comandos *pyplot* do matplotlib. De acordo com a documentação oficial, esta coleção de comandos faz com que o matplotlib funcione em estilo semelhante ao MATLAB.

Normalmente, importa-se esta coleção como plt. Feito isso, no matplotlib, cada comando executa uma alteração no gráfico, como criação da área, traçado dos pontos, mudança do label nos eixos, e o comando final exibe o gráfico.

Vamos passar uma lista para o comando plot() e depois usar o comando show() para exibir o gráfico.

Traçando os primeiros gráficos

Por padrão, ao receber uma lista, o comando plot() traça um gráfico de linha, com o índice da lista no eixo X e os itens no eixo Y.

Para o segundo exemplo, vamos definir os pontos do eixo X. Adicionalmente, um terceiro argumento passado à função plot() define o formato dos pontos e da linha do gráfico. Vamos fazer o gráfico com quadrados vermelhos e uma linha tracejada, passando como argumentos *color*, *linestyle* e *marker*. Usaremos "r" para a cor *red*, *linestyle* "–" para o tracejado e *marker* "s" para *square*, ou seja, quadrado. Por fim, vamos deixar a linha mais grossa com o parâmetro *linewidth*. Vamos também definir o eixo com a função *axis*, para melhorar um pouco a visualização. No *axis* você passa, dentro de uma lista, os valores para o eixo X e depois para o eixo Y, indicando os pontos iniciais e finais para cada eixo.

Gráfico de Barras

Outro gráfico popular presente no matplotlib é o gráfico de barras. Para o gráfico de barras, usamos a função bar(), onde definimos a posição das barras no eixo X e sua altura no eixo Y. Adicionalmente, também podemos configurar outras características, como a espessura das barras, cor, entre outros. O eixo X será um *range* com a mesma quantidade de itens do eixo Y. Vejamos um exemplo bem simples, onde vamos guardar as configurações em variáveis e então passá-los para a função.

Gráfico de Barras

Agora vamos usar o *Dataset* do Titanic para fazer um gráfico de barras empilhadas, onde poderemos visualizar, para cada sexo, os sobreviventes.

Primeiro usaremos a função pivot_table() do Pandas para criar uma tabela que agregue esses dados mencionados, passando qual deve ser a variável das linhas e colunas, a função que deve agregar os valores (pode ser soma, contagem, entre outros) e os valores aos quais esta função será aplicada, dividindo entre os valores já definidos nas linhas e colunas. Para quem já trabalhou com tabela dinâmica no Excel, esta função fica mais fácil de entender.

Gráfico de Barras

As barras, no nosso gráfico do Titanic, são criadas com a função bar(), onde definimos as posições do começo da barra no eixo X, a altura dos itens no eixo Y, a espessura da barra e sua cor. A função *xticks* define *labels* no eixo X. Primeiro definimos suas posições e depois um *Tuple* com os valores a serem incluídos no gráfico.

Porém, melhor ainda para fazer esta comparação seriam dois gráficos de pizza (*pie charts*), mostrando percentualmente quem sobreviveu e quem não sobreviveu para cada sexo.

Gráficos de pizza e gráficos múltiplos

Para gráficos múltiplos usamos a função subplots(). Ela cria o que o matplotlib chama de "figure", que seria o espaço onde os múltiplos gráficos podem ser criados. Os espaços onde são criados os gráficos funcionam como listas, ou *arrays*. Nos casos de uma linha ou coluna, os espaços são *arrays* de uma dimensão, e nos casos de mais de uma linha e coluna, os *arrays* são de duas dimensões.

Aqui criaremos uma *figure* com uma linha e duas colunas, onde colocaremos nossos gráficos de pizza lado a lado usando a função *pie*, e definindo nela os valores que irão compor o gráfico, *labels* para esses valores, vamos incluir a porcentagem de cada valor relativo ao total da pizza, junto com sua formatação e as cores de cada pedaço. Depois disso, incluiremos um título em cada *subplot* e a opção axis('equal') garantirá que os gráficos serão redondos (meramente estético).

Gráfico de dispersão (ou scatter plot)

Outro gráfico que pode ser útil é o gráfico de dispersão, ou *scatter plot*. Ele é bastante utilizado quando queremos ver se existe alguma relação entre duas variáveis. A função do *pyplot* utilizada é a scatter(). Vamos verificar, por exemplo, se existe entre os passageiros do Titanic alguma relação entre a idade e o valor pago na passagem. Para a função scatter(), passaremos os valores da primeira variável para o eixo X e da segunda para o eixo Y. Também definiremos um valor para "alpha", para que os pontos sejam transparentes e seja possível ver um pouco da transposição, e também vamos definir o parâmetro "c", para que as cores mudem de acordo com o valor pago no tíquete.

Também é possível configurar para que o tamanho dos pontos varie de acordo com o valor de uma dada variável, bastando passar a dada variável para o parâmetro 's', de *size*, ao invés do parâmetro 'c', de *color*.

Histogramas

Histogramas são úteis para visualizar a distribuição de uma série de valores. Vamos ver como se distribuem as idades dos passageiros do Titanic com um histograma. A função neste caso é a hist(). Nela, passaremos como parâmetros os valores de idade e a quantidade de barras, ou *bins*. A quantidade de barras é opcional, e se não passarmos nenhuma quantidade ela é definida em 10 por padrão.

Outros gráficos no Matplotlib

Existe ainda uma série de outros tipos de gráficos no matplotlib, como diagramas de caixa (*boxplot*), gráfico radar e outros.

Os gráficos apresentados nesta aula são alguns dos mais úteis e utilizados na análise de dados.

Probabilidade e algumas aplicações em Python

Probabilidade é a medida numérica que representa a chance de um determinado evento acontecer. Qual é a chance de eu jogar um dado e tirar um número 6? De eu jogar duas moedas para cima e as duas resultarem em Coroa? Essas perguntas (e muitas outras) são respondidas pela Probabilidade.

Para o cálculo básico da probabilidade de um evento, somamos todas as possibilidades que atendem este evento e dividimos pelo número total de possibilidades. Por exemplo, a probabilidade de rolar um 3 em um dado só é atendida pelo número 3. Entretanto, temos 6 possibilidades em um dado, os números que vão de 1 a 6 (o 3 inclusive). Desta forma, temos uma possibilidade que atende ao evento desejado dividido por 6 possibilidades no total, ou 1/6.

Eventos dependentes e independentes

Um primeiro conceito importante da Probabilidade é o conceito de eventos dependentes e independentes.

Os eventos independentes são eventos nos quais saber o resultado de um deles não nos dá qualquer informação sobre o possível resultado do segundo. O rolar de um dado não nos dirá nada sobre o resultado do rolar do segundo dado.

Analogamente, eventos dependentes são eventos nos quais o resultado de um nos dá mais informações sobre a probabilidade do outro. Neste caso, o rolar de um dado nos dá informações se queremos saber a probabilidade, por exemplo, de rolar 6 em dois dados.

Eventos dependentes e independentes

Para eventos independentes, a probabilidade de dois eventos acontecerem é igual a probabilidade de um multiplicada pela probabilidade do outro. Desta forma, se a probabilidade de rolar um determinado valor no dado é de 1/6, a probabilidade de rolar dois 6 seguidos é calculada da seguinte forma:

Sendo P(A) a probabilidade do evento A, P(B) a probabilidade do evento B e P(A,B) a probabilidade de ocorrerem os eventos A e B:

$$P(A,B) = P(A) \times P(B)$$

Eventos dependentes e independentes

Esta é a probabilidade de dois dados rolarem 6 seguidamente. Vamos usar a função fraction() do pacote *fractions*, para mostrar que este número equivale a 1/36, que é exatamente o resultado em forma de fração de 1/6 multiplicado por 1/6, veja no jupyter.

Evento Complementar

Eventos complementares são um conceito simples. Um evento complementar a um evento A é aquele que engloba todos os resultados que não atendem ao evento A. Novamente utilizando o rolar do dado como exemplo, se um evento A indica a probabilidade deste dado rolar um número menor ou igual a 2, que seria de 2/6, seu evento complementar seria a probabilidade do dado rolar um número maior que 2, ou 4/6. O cálculo da probabilidade de um evento complementar é bem simples, bastando subtrair de 1 a probabilidade do evento A.

Evento Complementar

Este conceito pode ser útil em casos onde é mais fácil calcular a probabilidade do complementar do que do evento em si.

Vejamos por exemplo, o seguinte caso. Qual a probabilidade de jogar dois dados e os dois números serem diferentes? Neste caso, a lista de resultados que atendem as condições entre os 36 possíveis são muitos. Mas o evento complementar deste é quando jogamos dois dados e os dois números rolados são iguais. Desta forma, os resultados que atendem a condição limitam-se a (1,1), (2,2), (3,3), (4,4), (5,5) e (6,6). Assim, podemos facilmente calcular a probabilidade do evento complementar em 6/36, e a probabilidade do evento desejada em 1 – 6/36, equivalente a 30/36 ou 5/6.

Eventos Mutuamente Exclusivos

Eventos mutuamente exclusivos ocorrem quando, entre dois eventos, apenas um pode acontecer. Ao jogar uma moeda, ela tem que resultar em cara ou coroa. Os dois não podem acontecer simultaneamente. Em um dado, o número rolado tem que ser par ou ímpar. Nenhum número atende às duas condições simultaneamente.

Eventos Mutuamente Exclusivos

Para dois eventos mutuamente exclusivos A e B, P(A,B), ou seja, a probabilidade que os dois ocorram, é igual a zero. Entretanto, podemos calcular a probabilidade de P(A ou B) que, para estes tipos de eventos, será igual a P(A) + P(B).

Para o caso dos dados, por exemplo, P(A ou B) será igual a 1, pois o número rolado necessariamente será par ou ímpar.

Para outro exemplo, em um baralho, consideremos a probabilidade de uma carta ser um valete (P(A)), igual a 4/52, ou 1/13, e a probabilidade dela ser um Ás (P(B)), também 1/13. Estas duas probabilidades dão-se por termos 4 de cada uma destas cartas em um baralho de 52 cartas. São eventos mutuamente exclusivos, pois uma carta não pode ser os dois. Desta forma, P(A,B) é igual a zero. Entretanto, a probabilidade dela ser um Valete ou um Ás (P(A ou B)) será igual à soma de P(A) e P(B), totalizando 1/13 + 1/13 = 2/13.

Probabilidade Condicional

Para eventos dependentes, o cálculo da Probabilidade muda. Vamos estabelecer que:

- P(A|B) -> Probabilidade condicional de A dado B, ou seja, probabilidade do evento A ocorrer, dado que ocorreu o evento B.
- P(A,B) -> Como já vimos, é a probabilidade dos dois eventos ocorrerem.
- P(A) e P(B) -> Também, como já vimos, é a probabilidade de cada evento acontecer.

Probabilidade Condicional

Para eventos dependentes, o cálculo é o seguinte:

P(A|B) = P(A,B)/P(B)

E algumas vezes, passamos P(B) para o outro lado da igualdade, e a equação fica assim:

P(A,B) = P(A|B) × P(B)

Vejamos um exemplo e um pouco de código. Consideremos um dado. Seja o primeiro evento tirar um número ímpar e o segundo tirar um 5 ou 6 no dado. Vamos calcular a probabilidade de A dado B.

P(A|B) = ??

P(A) = 3/6

P(B) = 2/6

P(A,B) = 2/6 × 3/6 = 0,16666 ou 16,66%, ou ainda 1/6 – Este resultado equivale a nossa única possibilidade, que é o número 5.

P(A|B) = (1/6)/(2/6)

Probabilidade Condicional

Vamos interpretar alguns destes resultados. Primeiramente, no cálculo da probabilidade de ambos os eventos, chegamos a um resultado de 1/6. Analisemos os dois eventos em separado. No caso do evento A, a probabilidade de um dado tirar um número ímpar, os resultados que atendem esta condição são os números 1, 3 e 5. Para o evento B, tirar um 5 ou um 6, os resultados que atendem, obviamente, são o 5 e o 6. Analisando conjuntamente, o único número que atende os dois eventos é o 5. Logo, 1/6, a chance de tirar um cinco no dado.

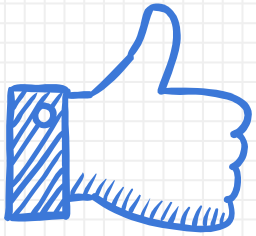
Probabilidade Condicional

Agora, analisemos a probabilidade de A dado B. Considerando que B aconteceu, sabemos que o número rolado é 5 ou 6. O resultado entre estes que atende o evento A é o 5. Dentro do universo de possibilidades de B, temos um resultado que atende nossa condição entre dois possíveis. Logo, 1/2. Assim, também comprovamos que os eventos são independentes, pois o cálculo de P(A,B) leva a uma conclusão que confere com a realidade de cada evento.

Vamos calcular e analisar rapidamente a probabilidade P(B|A):

Probabilidade Condicional

Neste caso, sabemos que o evento A, o dado rolar um número ímpar, já ocorreu. Então, temos 1, 3 ou 5. Entre estes, o único número que atende a condição do evento B, rolar um 5 ou um 6, é o próprio 5. Desta forma, dentre os 3 possíveis resultados A, apenas um atende a condição B, levando desta forma a uma probabilidade de B dado A igual a 1/3.



Obrigado!