CARD 5 - Prática: Estatística p/ Aprendizado de Máquina (I)

Este card tem por objetivo revisar conceitos estatísticos importantes para o mundo do Machine Learning, além de ensinar a aplicação desses conceitos no Python e apresentar o Matplotlib, uma biblioteca de visualização de dados em Python amplamente utilizada para criar gráficos e figuras. Ela permite a criação de gráficos de linhas, barras, dispersão, histogramas e muitos outros tipos de visualizações, oferecendo uma interface fácil de usar e altamente customizável.

Abaixo segue um resumo dos principais conceitos estatísticos aprendidos no curso:

- **Mean (Média):** é a soma de um conjunto de valores dividida pelo número total de valores. É uma medida de tendência central que representa o valor típico ou central de um conjunto de dados.
- **Median (Mediana):** é uma medida de tendência central que representa o valor do meio de um conjunto de dados ordenados. Para encontrar a mediana, você deve organizar os dados em ordem crescente ou decrescente e identificar o valor central (não sofre tanta influência de um ponto discrepante nos dados diferente da média).
- **Mode (Moda):** é o valor que aparece com maior frequência em um conjunto de dados. É uma medida de tendência central que indica o valor mais comum ou recorrente.
- **Variation (Variação):** é a medida de quão dispersos ou espalhados estão os valores de um conjunto de dados. A variação quantifica a diferença entre os valores dos dados em relação à média do conjunto.
- **Standard Deviation (Desvio Padrão):** é uma medida estatística que indica o grau de dispersão ou variabilidade dos valores em um conjunto de dados em relação à média. Em outras palavras, o desvio padrão mostra o quanto os dados se desviam da média, em média.
- **Probability Density Function (Função Densidade de Probabilidade):** é uma função que descreve a probabilidade de uma variável aleatória contínua assumir um determinado valor.
- **Probability Mass Function (Função Massa de Probabilidade):** é uma função que descreve a probabilidade de uma variável aleatória discreta assumir cada um dos seus possíveis valores.
- **Normal Data Distribution (Distribuição Normal):** descreve como os valores de uma variável aleatória se distribuem de forma simétrica em torno da média.
- **Binomial Data Distribution (Distribuição Binomial):** descreve o número de sucessos em um número fixo de tentativas ou experimentos, onde cada tentativa tem apenas dois resultados possíveis: sucesso ou fracasso.
- **Poisson Data Distribution (Distribuição de Poisson):** é usada para modelar a contagem de eventos que ocorrem em um intervalo de tempo ou espaço específico, quando esses eventos ocorrem de forma independente e a uma taxa média conhecida.
- **Percentile (Percentil):** indica a posição de um valor específico em relação a uma série de dados ordenados. Em outras palavras, o percentil informa qual porcentagem dos dados está abaixo de um determinado valor.
- **Skew (Assimetria):** refere-se à medida que descreve a falta de simetria da distribuição dos dados. Em outras palavras, a assimetria indica se os dados estão distribuídos de maneira simétrica em torno da média.
- **Kurtosis (Curtose):** é uma medida que descreve a forma da distribuição dos dados em relação à sua concentração na cauda (Extremidades).

- **Covariance (Covariância):** descreve o relacionamento linear entre duas variáveis aleatórias. Em resumo, a covariância indica como duas variáveis se comportam juntas: se elas variam juntas (quando uma aumenta, a outra também aumenta ou diminui) ou se elas variam de maneira oposta (quando uma aumenta, a outra diminui).
- **Correlation (Correlação):** é uma medida que descreve o grau e a direção do relacionamento linear entre duas variáveis.
- **Conditional Probability (Probabilidade Condicional):** é uma medida que descreve a probabilidade de um evento ocorrer, dado que outro evento já ocorreu ou é conhecido.
- **Bayes' Theorem (Teorema de Bayes):** trata-se de um princípio fundamental na teoria da probabilidade que descreve como atualizar probabilidades quando novas evidências são obtidas. Ele é usado para calcular a probabilidade condicional de um evento, dado que outro evento já ocorreu.

Em resumo, estes são os principais conceitos apresentados no curso. Vale ressaltar que no vídeo também foram ensinadas as aplicações desses conceitos dentro do Python, utilizando as bibliotecas NumPy e Matplotlib.

Conclusão:

Ao longo deste card, exploramos a importância dos conceitos estatísticos fundamentais para o entendimento e a modelagem de dados no contexto do Machine Learning. A análise estatística não apenas proporciona insights valiosos sobre os dados, mas também fundamenta as decisões de modelagem e predição. Além disso, a utilização eficaz da biblioteca Matplotlib se revelou crucial para visualizar e comunicar resultados complexos de forma clara e informativa. Assim, tanto a compreensão dos princípios estatísticos quanto o domínio das ferramentas de visualização como Matplotlib são essenciais para o avanço e a aplicação bem-sucedida de técnicas de Machine Learning em diversos campos.