Relatório_parteI

February 4, 2022

Todos os exercícios que exigiam códigos foram resolvidos na linguagem de Python, no ambiente Jupyter Notebook. Os algoritmos vistos em sala estão implementados no código.

Na **Questão 1**, utilizou-se a base de dados Car Evaluation. Como é uma base com dados categóricos, tais dados foram transformados em dados numéricos com a função `transformar_dados`. Para a **Questão 1a**, como é comum, antes da aplicação do PCA, calculou-se a média dos atributos e diminuiu-se na base de dados original. O PCA foi calculado com as funções `np.cov` e `np.linalg.eig` da biblioteca *numpy*.

Para a **Questão 1b**, utilizou-se os autovetores correspondentes às duas primeiras componentes principais encontradas para obtenção do *factor loadings*. Os autovetores foram normalizados e plotados em um gráfico 2d.

Na **Questão 1c** criou-se uma função para calcular a entropia de um atributo e uma função para calcular a informação mútua entre dois atributos. O cálculo da entropia é feito após a normalização das probabilidades do atributo. Para obter as probabilidades, a função que calcula a informação mútua utiliza a função `np.histogram` (para um único atributo) e `np.histogram2d` (para dois atributos).

A **Questão 2a** consistiu de apenas aplicar o algoritmo t-SNE na base de dados Haberman. Para a implementação do algoritmo SMOTE, na **Questão 2b**, utilizou-se o algoritmo de vizinhos mais próximos. Após a identificação da classe minoritária, aplicou-se o SMOTE sobre a base de dados e adicionou-se a nova base na base original. Por fim, calculou-se o t-SNE da nova base balanceada.

A **Questão 3** foi resolvida com base no que foi estudado em sala de aula.

Por fim, na **Questão 4** implementou-se o algoritmo de regressão linear para ser utilizado na base Auto MPG. Entretanto, antes de ser utilizada, a base passou por um processamento, que consistia em separar corretamente os valores correspondentes às colunas. Além disso, removeu-se a última coluna (que era não era do tipo numérico) e valores com "?", que também não são numéricos. A implementação da regressão linear foi feita em duas partes. A primeira obtinha os coeficientes da base de dados e a segunda fazia a predição com base nos coeficientes calculados. Assim, para a **Questão 4a** dividiu-se a base em treino e teste e calculou-se o RMSE do modelo de Regressão Linear.

Na **Questão 4b**, implementou-se os algoritmos de RSS e RSS_0, para identificação de quais atributos menos relevantes para a base de dados. Assim, removeu-se os atributos identificados e realizou-se uma nova execução do algoritmo de Regressão Linear. Por fim, comparou-se o RMSE do modelo com os atributos reduzidos com o RMSE do modelo com todos os atributos.

[ ]: