Lista 2 - Exercícios Geral:

Prof. Dr. Jodavid Ferreira

Discente:

* Gabriel D'assumpção de Carvalho

Data: 27/07/2024

---

1. **Quais são os três principais conceitos do Machine Learning e e como funciona cada um deles?**

    A. **Aprendizado Supervisionado**: O aprendizado supervisionado é um processo de aprendizado em que os dados estão no formato estruturado, portanto se tem o conhecimento da caracteristica de todas as variáveis de entrada (X) e de saida (y). Portanto, o modelo vai ser treinado com os dados de entrada e saida $D = \{X, y\}$ e o algoritmo vai tentar encontrar uma função em $g \sim f(X, y)$

    B. **Aprendizado Não Supervisionado**: Diferente do modelo supervisionado, no não supervisionado os dados de entrada e saida não são estruturado, por tanto a ideia do modelo é criar grupos das observações com base na semelhança entre os dados. Entretanto, o algoritmo no caso do não supervisionado vai tentar agrupar os dados de entrada **X** (features) em grupos (clusters).

    C. **Aprendizado Por Reforço**: Já no aprendizado por reforço se tem um modelo onde vamos ter uma função objetivo e a cada iteração do modelo com o ambiente pode se ter uma recompensa ou perda na função objetivo.

2. **O que é um modelo de Machine Learning?** O modelo de Machine Learning é um algoritmo que tem a capacidade de aprender as relações entre os dados, então se temos umas grande quantidade de dados o algoritmo tem uma grande capacidade de aprender o padrão existente entre eles.

3. **O que é um conjunto de treinamento?** O conjunto de treinamento é um conjunto de dados onde o modelo vai ser treinado e o algoritmo vai tentar estimar uma função onde com a inputação dos features (X) o algoritmo vai nos retornar a saida (y) no caso dos modelos supervisionados.

4. **O que é um conjunto de teste?** O conjunto de teste é a base de dados que vai avaliar o desempenho do modelo com base na estimativa realizada atraves do conjunto de

treinamento, essa separação é feita para que o modelo se torne mais preciso e tenha uma maior capacidade de generalizar, sendo capaz de prever novos dados.

5. **Considerando que estamos trabalhando com dados estruturados, como é a estrutura de um conjunto de dados para um estudo de Análise Supervisionada?** Cite exemplos de problemas que podem ser aplicados Análise Supervisionada. Para um dados estruturados o conjunto de dados vão ser separado em features (**X**) e targets (**y**), onde os features pode ser idade, peso, altura, raça, circunferência da cintura e etc, e o target pode ser a probabilidade da pessoa ter diabetes ou nao.

6. **Quais são as etapas do Processo de Aprendizagem de um Modelo?** O processo de aprendizado deve ter as seguintes etapas:

    A. Coleta dos Dados
    B. Analise Exploratoria dos Dados
    C. Preparação dos Dados
    D. Separação dos Dados de Treinamento e Teste ou Treinamento, Validação e Teste.
    E. Treinamento do Algoritmo ou Treinamento e Validação do Algoritmo
    F. Teste do Algoritmo
    G. Avaliação do Modelo

7. **O que é Underfitting e Overfitting?** Como podemos identificar esses problemas em um modelo de Machine Learning?

    - **Underfitting**: É o nome que se da quando o modelo não tem a capacidade de compreender o padrão ou relação existente aos dados, sendo um modelo muito simples.
    - **Overfitting**: Ao contrario do under, o overfitting se da quando o algoritmo é bem complexo e ele aprende demasiadamente o padrão dos dados do conjunto de treinamento e perde a capacidade de generalizar para novos dados.

8. **Como podemos definir o erro total de previsão de um modelo?** O erro total de previsão pode ser decomposto em 3 erros, sendo eles:

    A. **Erro de Viés**: Que seria o erro quando o modelo é simples e não consegue capturar o padrão dos dados.
    B. **Erro de Variância**: Este erro acontece quando temos um algoritmo complexo e ele acaba perdendo a capacidade de generalizar para novas observações.
    C. **Erro irredutivel**: Já o erro irredutivel se da pelo fato do ruido apresentado nos dados, portanto o medelo não é capaz de prever corretamente esse erro.

9. **Qual a regra geral de um bom modelo?** Um bom modelo precisa ter uma boa precisão, ou seja, conseguir fazer boas previsões; também precisa ter uma boa interpretabilidade, podendo verificar quais variáveis são mais importantes; ter uma boa capacidade de generalização para novas observações; o modelo precisa ser escalável, sendo capaz de trabalhar com grande quantidade de dados e, ao mesmo tempo, ser eficiente, gerando previsões em um tempo rápido e consumindo o mínimo de recursos.

10. **Cite duas métricas e coloque a equação descrevendo as variáveis de solução que podemos utilizar para avaliar quantitativamente o desempenho do modelo de Regressão?**

A. **Distância Euclidiana**: A distância euclidiana mede a diferença entre os valores verdadeiros e os valores previstos. É calculada como a raiz quadrada da soma dos quadrados das diferenças individuais entre os valores verdadeiros e previstos:

$$\mathrm{d}(y, \hat{y}) = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

B. **Erro Quadrático Médio (MSE)**: O Erro Quadrático Médio é a média da soma dos quadrados das distâncias entre os valores verdadeiros e os valores previstos. É uma medida comum para avaliar a precisão de um modelo de regressão:

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

11. **Cite duas métricas e coloque a equação descrevendo as variáveis de solução que podemos utilizar para avaliar quantitativamente o desempenho do modelo de Classificação?**
Diferente da regressão, agora estamos interessado de saber se o modelo fez a classificação certa, portanto so vamos ter dois estados ou o modelo acerta a classificação ou ele erra.

A. **Acurácia**: Esta métrica é importante porque nos permite ver a proporção de instâncias que o modelo classifica corretamente. A fórmula da acurácia é:

$$\mathrm{Acurácia} = \frac{TP + TN}{P + N}$$

onde:

- (TP) (True Positive) é o número de verdadeiros positivos.
- (TN) (True Negative) é o número de verdadeiros negativos.
- (P) (Positive) é o número total de positivos.
- (N) (Negative) é o número total de negativos.

B. **Taxa de Erro**: A taxa de erro é uma boa estatística porque nos permite ver a proporção de instâncias que o modelo classifica incorretamente. A fórmula da taxa de erro é:

$$\mathrm{Taxa\ de\ Erro} = \frac{FP + FN}{P + N}$$

onde:

- (FP) (False Positive) é o número de falsos positivos.
- (FN) (False Negative) é o número de falsos negativos.
- (P) (Positive) é o número total de positivos.
- (N) (Negative) é o número total de negativos.