AGENDA

Problema de regressão

- Dado um conjunto de instâncias em que cada uma possui um atributo alvo de valor contínuo que se deseja predizer tem-se então um problema de regressão.
- Cada instância desse conjunto possui atributos que podem ser avaliados de

modo que possa ser encontrada uma função que possa ser utilizada para predizer o atributo alvo.

- Alguns algoritmos que são voltados para regressão: Árvore de decisão, SVM, MLP, SDG e Floresta Aleatória.

Árvore de decisão

- A árvore de decisão para regressão funciona da mesma maneira quando utilizada para realizar uma classificação, contudo, nesta abordagem as folhas da árvore contém um valor numérico contínuo que será a predição do atributo alvo.
- Os nós folhas são valores contínuos médios obtidos através dos atributos de

entrada.

- Ao longo do treino do algoritmo avalia-se o Erro quadrático médio de forma a minimizá-lo em busca de predições mais acuradas

Árvore de decisão

- Sua implementação para regressão segue o mesmo caminho que para a classificação.
- A função DecisionTreeRegressor(max_depth=7).fit(X_treino, Y_treino) retorna um regressor treinado utilizando o dataset fornecido.
- O parâmetro *max_depth* determina a profundidade máxima da árvore, para regressão recomenda-se uma maior profundidade devido ao incremento das

possibilidades de predição. A profundidade ideal para um dado dataset é obtida com base em experimentação.

SVM - Support Vector Machine

- Para a regressão o objetivo é minimizar o erro de uma função de regressão de modo que ele se torne menor que um determinado limite ��.
- Para isso uma região de tamanho ?? ao redor do hiperplano é definida juntamente com uma função de perda que deve ser minimizada com o intuito de encontrar uma região reduzida que contenha o maior número de instâncias de treinamento.
- Uso dessa função faz com que o algoritmo penalize predições que distam

mais que da saída desejada.

- Para implementação do regressor utiliza-se a função SVR().fit(X_treino, Y_treino) que retorna um regressor treinado com base no dataset fornecido.

MLP - Multilayer Perceptron

- Da mesma forma que a rede neural Perceptron pode ser utilizada para predição de valores discretos(classificação) ela também pode ser treinada para realizar a predição de valores contínuos.
- Diferentemente da MLP para classificação neste caso a função de ativação na saída da rede que é responsável por discretizar os valores é removida para que

a saída contínua possa ser utilizada.

- O regressor é implementado por meio da função MLPRegressor().fit(X_treino, Y_treino)

Floresta Aleatória

- Nesta abordagem um conjunto de árvores de regressão(apresentadas anteriormente) são treinadas de forma conjunto para se obter um regressor mais preciso.
- Sua implementação é realizada por meio da função RandomForestRegressor(n_estimators=100), com o parâmetro *n_estimators* sendo o número máximo de árvores de decisão utilizadas no treinamento.

Gradiente Descendente Estocástico

- Técnica de otimização bastante utilizada em algoritmos de regressão linear que realiza a busca de parâmetros para a minimização de uma função de custo.
- Consiste em uma variação do Gradiente Descendente comum, em que a cada iteração são selecionadas apenas algumas amostras aleatórias do conjunto de dados para otimizar a função de custo.
- Nesta abordagem o tempo de cada iteração é consideravelmente diminuído por conta da utilização de poucos exemplos tornando possível sua execução de maneira eficiente.

Gradiente Descendente Estocástico

- A implementação desse modelo pode ser efetuada por meio da função SGDRegressor(max_iter=1000, tol=1e-3, loss='epsilon_insensitive').fit(X_treino, Y_treino).
  - max_iter: Número máximo para iterações para treinamento do modelo.
  - loss: Função de perda.

Métricas para avaliação

- **Erro Quadrático Médio(MSE)**: calcula a média do quadrado da diferença entre os valores observados e os valores preditos.

- **Raiz Quadrada do Erro Quadrático Médio(RMSE)**: Mede a diferença ou resíduo entre o valor real e o valor predito.
- **Erro Absoluto Médio(MAE)**: medida que nos fornece o quão distante dos valores reais estão as predições realizadas pelo modelo.
- **Variância Explicada(Ev)**: medida utilizada para medir a porção da variância presente nos dados que é explicada pelo modelo estimado.

Métricas para avaliação

- **Coeficiente de Determinação(R2)**: medida utilizada para avaliar o quanto um modelo de regressão estimado consegue explicar dos valores observados. O valor da medida varia entre 0 e 1, sendo que o valor 1 indica que a variável

independente pode ser predita com base na variável independente sem nenhum erro e o valor 0 indica que não é possível realizar a predição.

$$MAE = \frac{1}{N}\sum_{i}^{N}|x_i - \hat{x}_i| \quad RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N}(x_i - \hat{x}_i)^2} \quad MSE = \frac{1}{N}\sum_{i}^{N}(x_i - \hat{x}_i)^2$$

$$Ev(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{Y\}} \quad R^2 = \left(\frac{1}{N}\sum_{i}^{N}\frac{[(x_i - \bar{x})(y_i - \bar{y})]}{\sigma_x \sigma_y}\right)^2$$