Machine Learning

Trabalho Vivencial.

Professor: Prof. Msc. Paulo Cirillo Souza Barbosa

Introdução.

O presente trabalho é composto por duas etapas em que deve-se utilizar os conceitos de machine learning baseados em modelos preditivos que realizam seu processo de aprendizagem através dos paradigmas supervisionado e não supervisionado. A primeira parte do trabalho envolve a resolução de duas tarefas: classificação e regressão. A segunda etapa envolve apenas a classificação.

Aprendizado Supervisionado.

Tarefa de Regressão

Para o problema de regressão solicita-se que faça o acesso ao conjunto de dados disponibilizado na plataforma AVA, chamado aerogerador.dat. A variável independente é uma medida de velocidade do vento, e a variável dependente é uma observação de potência gerada pelo aerogerador.

1. Faça uma visualização inicial dos dados através do gráfico de espalhamento. Nessa etapa, faça discussões sobre quais serão as características de um modelo que consegue entender o padrão entre variáveis regressoras e variáveis observadas.

2. Em seguida, organize os dados de modo que as variáveis regressoras sejam armazenadas em uma matriz ($\mathbf{X}$) de dimensão $\mathbb{R}^{N \times p}$. Faça o mesmo para o vetor de variável dependente ($\mathbf{y}$), organizando em um vetor de dimensão $\mathbb{R}^{N \times 1}$.

3. Deve-se implementar um modelo de regressão linear cuja estimativa dos parâmetros, é realizada pelo método dos mínimos quadrados ordinários. A implementação pode utilizar bibliotecas segue link referência: Linear Regression (scikit-learn)

4. Para medir o desempenho do modelo, utilize o processo de validação por amostragem aleatória (com 500 rodadas). Em cada rodada, deve-se realizar o particionamento em 80% dos dados para treinamento e 20% para teste. As medidas de desempenho devem ser Mean Squared Error (MSE), bem como o Mean Absolute Error (MAE). A cada cálculo, deve-se armazenar a métrica em uma lista.

5. Ao final das 500 rodadas calcule para cada métrica, a média aritmética, desvio-padrão, valor maior, valor menor. Coloque os resultados obtidos em uma tabela e discuta os resultados obtidos. **Segue exemplo de tabela:**

Modelos	**Média**	**Desvio-Padrão**	**Maior Valor**	**Menor Valor**
MSE				
MAE				

Tarefa de Classificação

No ambiente virtual AVA, está disposto um conjunto de dados referente aos sinais de eletromiografia, captados nos músculos faciais: Corrugador do Supercílio (Sensor 1); Zigomático Maior (Sensor 2). O presente conjunto de dados foi obtido através de um grupo de sensores chamados ***Myoware Muscle Sensor***, em conjunto com

um microcontrolador NODEMCUESP32. As aquisições foram realizadas numa taxa de amostragem de 1Khz e a resolução do ADC do microcontrolador é de 12 bits (0 – 4095). Os sensores foram posicionados em duas regiões diferentes da face de uma única pessoa, em que o primeiro sensor se encontra na região do Corrugador do Supercílio e o segundo foi posicionado no músculo Zigomático Maior. As aquisições foram realizadas seguindo um roteiro de expressões faciais forçadas com a seguinte ordem: neutro; sorriso; sobrancelhas levantadas; surpreso; rabugento. Este roteiro se repetiu 10 vezes e cada gesto foi posto durante 1 segundo.

O arquivo **EMGDataset.csv** possui $N = 50000$ amostras, $p = 2$ características e $C = 5$ classes. Na primeira linha da matriz, existem os dados obtidos pelo sensor posicionado no Corrugador do Supercílio. Na segunda linha da matriz, existem os dados obtidos pelo sensor posicionado no Zigomático Maior. A terceira linha são informações referentes da categoria para cada amostra, rotuladas da seguinte maneira:

1 – Neutro

2 – Sorriso

3 – Sobrancelhas levantadas

4 – Surpreso

5 – Rabugento

Após o download, faça o que se pede:

1. Organize os dados do arquivo em variáveis $\mathbf{X}$ e $\mathbf{Y}$, de modo que elas sejam matrizes (numpy array) com as seguintes dimensões

$$\mathbf{X} \in \mathbb{R}^{N\times p} \qquad \mathbf{Y} \in \mathbb{R}^{N\times C} \text{ Para o modelo que estima seus parâmetros via método MQO}$$
$$\mathbf{X} \in \mathbb{R}^{p\times N} \qquad \mathbf{Y} \in \mathbb{R}^{C\times N} \text{ Para os modelos gaussianos bayesianos}$$

2. Faça uma visualização inicial dos dados através do gráfico de espalhamento (destacando as categorias). Nessa etapa levante hipóteses sobre quais serão as características de um modelo que consegue separar as classes do problema. (são linearmente separáveis ou não, por exemplo).

3. O modelo a ser implementados nessa etapa é o $K-$NN. Este pode ser consumido via biblioteca: KNN

4. Pede-se inicialmente que utilize a validação chamada de $k-$*fold cross-validation* para encontrar o hiperparâmetro K do modelo KNN. Assim, sua equipe deve aplicar o modelo para diferentes valores de K, por exemplo, $K = \{1, 7, 11, 17, 23, 39, 101, 501, 1001\}$. Para a validação, segue o link referência de qual método utilizar: k-fold cross validation.

5. Com a definição de K, aplique a validação cruzada por amostragem aleatória, e faça a composição da matriz de confusão, bem como calcule a acurácia para cada treino e teste. Neste caso, utilize a mesma estratégia de particionamento adotada para a tarefa de regressão. Ao final das rodadas extraia as mesmas métricas estatísticas e insira no relatório a matriz de confusão que está associada ao maior valor de acurácia, bem como o menor valor.

Modelos	**Média**	**Desvio-Padrão**	**Maior Valor**	**Menor Valor**
Acurácia				

1 Paradigma Não-Supervisionado

O conjunto de dados disponibilizado no AVA, tratam-se de 640 imagens com rostos de diferentes pessoas. A tarefa está associada a fazer com que o modelo encontre padrões específicos e classifique os rostos de pessoas.

1.1 Redução de Dimensionalidade

- Como o conjunto de dados é composto por imagens de dimensões 128×120, existem 15.360 variáveis independentes. Assim, é importante analisar o conjunto de dados sob a perspectiva de alguns métodos para aplicar a redução de dimensionalidade.
- Inicialmente, pede-se que aplique o método $t - SNE$ para visualizar a projeção dos dados na segunda dimensão. A aplicação do método, pode te dar percepções sobre o grau de separabilidade dos grupos existentes. Neste caso, como a estamos falando de um caso não supervisionado, essa análise visual pode te fornecer o ponto de partida dos grupos a serem encontrados com os modelos $K-means$ e $K-medoids$. O $t - SNE$ pode ser utilizado via biblioteca: tsne
- Em seguida, deve-se aplicar a redução de dimensionalidade de duas maneiras diferentes: pelo Método PCA (link: PCA; pelo método UMAP (link: UMAP.
- Crie algumas variações da utilização de tais métodos. Por exemplo, no caso do PCA armazene a saída para o caso em que utiliza-se as componentes que representem diferentes variâncias (90%, 80% e 75%). Para o UMAP escolha a redução para dimensões diferentes (3, 15, 55, 101).

1.2 Algoritmos de Clusterização

- Antes de aplicar os métodos $K - means$ e $K - medoids$, pede-se que faça a visualização dos resultados obtidos na etapa anterior. Assim, se o método de redução escolhido lhe proporcionou uma quantidade de características (variáveis independentes) maior que 3, então deve-se utilizar as estratégias adotadas no material do terceiro percurso de aprendizagem.
- Em seguida, faça a aplicação dos métodos $K-means$ e $K-medoids$ para diferentes valores de K. Estes valores devem ser escolhidos por sua equipe, com base nos resultados obtidos anteriormente.
- Para cada K avaliado, deve-se computar o índice de Dunn, afim de definir o valor ideal dentre aqueles testados. Faça a discussão dos resultados obtidos.

5) Critérios Avaliativos.

- A equipe deve realizar um envio de uma pasta compactada contendo:
  1. As implementações realizadas no trabalho. A pontuação deste é (0 até 1,25).
  2. Documento relatório Que possui pontuação de 0 até 2,5. Este deve ser escrito utilizando o modelo de artigo científico disponibilizado neste **LINK**. A avaliação do relatório será realizada da seguinte forma (descrita em proporção):
     - Título (5%).
     - Metodologia (47,5%).
     - Resultados (47,5%).
  3. Além das implementações, sua equipe deve gravar um vídeo de no máximo 10 minutos, no qual cada membro da equipe deve ter um papel protagonista na explicação do trabalho. Assim, é **obrigatório** que no início de fala de cada membro seja dito qual nome e matrícula. A pontuação também será de (0 até 1,25), porém, esta será condicionada a apresentação realizada pelo aluno.

6) Observações.

- Obs1: O trabalho é de equipe composta de 6 alunos no máximo.
- Obs2: A data estipulada para entrega do trabalho, também é um critério avaliativo. Assim, a nota da equipe será **zero** caso o envio não seja realizado.
- Ob3: Os relatórios e implementações serão enviadas a um software anti-plágio. Qualquer caracterização de plágio ocasionará em nota zero para ambas equipes.