Inteligência Artificial Computacional

Trabalho Computacional Modelos de Regressão e Classificação .

Professor: Prof. Msc. Paulo Cirillo Souza Barbosa

Introdução.

O presente trabalho é composto por duas etapas em que deve-se utilizar os conceitos de IA baseados em modelos preditivos que realizam seu processo de aprendizagem através da minimização de uma função **custo** (***loss function***). Em ambas etapas do trabalho, tais modelos utilizam o paradigma supervisionado para aprender a partir dos pares, vetor de características(variáveis regressoras) e variável dependente. Contudo, a tarefa da primeira etapa trata-se do desenvolvimento de um sistema que faz predições quantitativas (**regressão**), ao passo que a segunda etapa é caracterizada pelo desenvolvimento de um sistema que realiza predições qualitativas (**classificação**).

Tarefa de Regressão. [3,0 pts]

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. Os modelos a serem implementados nessa etapa serão: **MQO tradicional**, **MQO regularizado** (Tikhonov) e **Média de valores observáveis**.**Obs:** lembre-se que todos os modelos também estimam o valor do intercepto.

4. Para o modelo regularizado, há a dependência da definição de seu hiperparâmetro λ. Assim, sua equipe deve testar o presente modelo para os seguintes valores de lambda:

$$\lambda = \{0, 0.25, 0.5, 0.75, 1\}$$

   . Assim, ao todo, existirão 6 estimativas diferentes do vetor $\boldsymbol{\beta} \in \mathbb{R}^{p+1\times 1}$

5. Para validar os modelos utilizados na tarefa de regressão, sua equipe deve projetar a validação utilizando as simulações por Monte Carlo. Nessa etapa, defina a quantidade de rodadas da simulação igual a $R = 500$. Em cada rodada, deve-se realizar o particionamento em 80% dos dados para treinamento e 20% para teste. A medida de desempenho de cada um dos 5 modelos diferentes deve ser a soma dos desvios quadráticos (RSS - *Residual Sum of Squares*) e cada medida obtida deve ser armazenada em uma lista.

6. Ao final das R rodadas calcule para cada modelo utilizado, média aritmética, desvio-padrão, valor maior, valor menor de cada RSS. Coloque os resultados obtidos em uma tabela e **discuta os resultados obtidos**. **Obs:** O resultado não precisa ser limitado a tabela, como pode ser expresso via gráficos!

Modelos	Média	Desvio-Padrão	Maior Valor	Menor Valor
Média da variável dependente				
MQO tradicional				
MQO regularizado (0,25)				
MQO regularizado (0,5)				
MQO regularizado (0,75)				
MQO regularizado (1)				

Tarefa de Classificação [7,0 pts]

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. Os modelos a serem implementados nessa etapa serão: **MQO tradicional**, **Classificador Gaussiano Tradicional**, **Classificador Gaussiano Com Covariâncias Iguais**, **Classificador Gaussiano com Matriz Agregada**, **Classificador Gaussiano Regularizado (Friedman)** e **Classificador de Bayes Ingênuo**.

4. Para o classificador gaussiano regularizado, há a dependência da definição de seu hiperparâmetro λ. Assim, sua equipe deve testar o presente modelo para os seguintes valores de lambda:

$$\lambda = \{0, 0.25, 0.5, 0.75, 1\}$$

.

5. Para validar os modelos utilizados na tarefa de classificação, sua equipe deve projetar a validação utilizando as simulações por Monte Carlo. Nessa etapa, defina a quantidade de rodadas da simulação igual a $R = 500$. Em cada rodada, deve-se realizar o particionamento em 80% dos dados para treinamento e 20% para teste. A medida de desempenho de cada um dos 5 modelos diferentes deve ser a **acurácia** (taxa de acerto), e cada medida obtida deve ser armazenada em uma lista.

6. Ao final das R rodadas calcule para cada modelo utilizado, média aritmética, desvio-padrão, valor maior, valor menor das acurácias obtidas para cada modelo. Coloque os resultados obtidos em uma tabela (exemplo fornecido) e **discuta os resultados obtidos**. **Obs:** O resultado não precisa ser limitado a tabela, como pode ser expresso via gráficos!

Modelos	**Média**	**Desvio-Padrão**	**Maior Valor**	**Menor Valor**
MQO tradicional				
Classificador Gaussiano Tradicional				
Classificador Gaussiano (Cov. de todo cj. treino)				
Classificador Gaussiano (Cov. Agregada)				
Classificador de Bayes Ingênuo (Naive Bayes Classifier)				
Classificador Gaussiano Regularizado (Friedman λ =0,25)				
Classificador Gaussiano Regularizado (Friedman λ =0,5)				
Classificador Gaussiano Regularizado (Friedman λ =0,75)				

Extra - Regressão[0,5 pts]

Para a tarefa de regressos, os resultados obtidos poderiam ser melhores do que os obtidos pelos modelos utilizados? Existe um modelo que consegue entender melhor as relações das variáveis? Um modelo não linear poderia resolver esse problema? Se sim, defina um valor do polinômio não linear e construa um sistema não linear de equações que minimize a soma dos desvios quadráticos. Com esse modelo implementado, faça sua inclusão no processo das 1000 rodadas de treinamento e teste. Discuta os resultados obtidos. **Essa pontuação extra apenas será fornecida caso o trabalho seja entregue por completo, bem como não seja usado nenhuma biblioteca que não sejam as permitidas. Além disso, todos os membros da equipe devem saber explicar o que foi desenvolvido**.

5) Relatório.

Além das implementações, o presente trabalho deve ser entregue em modelo de relatório. Este deve possuir as características descritas nos slides de apresentação do curso. Desta maneira, deve possuir:

1. Título (2,5%).
2. Resumo (2,5%).
3. Metodologia (42,5%).
4. Resultados (42,5%).
5. Conclusões (10%).

O modelo para trabalho pode ser encontrado neste **LINK**

6) Observações.

- Obs1: O envio das implementações é **obrigatório**. Caso a equipe não realize esta entrega, será atribuído nota **zero** para os respectivos alunos.
- Obs2: A data estipulada para entrega do trabalho, também é um critério avaliativo. Assim, caso haja atraso na entrega do trabalho, será aplicada: **de 00:15h até 24h: penalidade de 20% ; 24:15h até 48h: penalidade de 40% ; acima de 48h: penalização máxima (100%).**

- Obs3: A apresentação do trabalho no formato de arguição, é obrigatória. A data da apresentação está definida no AVA. Se a equipe não participar deste momento, a nota do trabalho será ZERO.
- Ob4: Os trabalhos e implementações serão enviadas a um software anti-plágio. Qualquer caracterização de plágio ocasionará em nota zero para ambas equipes.
- Obs5: Para o presente trabalho, não será permitido o uso de bibliotecas que tenham as implementações prontas dos modelos de qualquer algoritmo requisitado no trabalho. Caso sua equipe faça o uso de tais bibliotecas, a nota atribuída será zero.