Inteligência Artificial Computacional

Trabalho Computacional 1: 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 de pares de amostra e valor observado. 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.

Para o problema de regressão solicita-se que faça o acesso ao conjunto de dados disponibilizado na plataforma AVA. A descrição do que são as variáveis dependente e independente é disponibilizada na própria plataforma. Após o download, faça o que se pede:

1. Faça uma visualização inicial dos dados através do gráfico de espalhamento. Nessa etapa levante hipóteses 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 de dimensão $\mathbb{R}^{N\times p}$. Faça o mesmo para o vetor de variáveis observadas, organizando em um vetor de dimensão $\mathbb{R}^{N\times 1}$.

3. Para validar os modelos utilizados na tarefa de regressão, é necessário definir uma quantidade específica de rodadas de treinamento e teste dos modelos. Assim, defina essa quantidade de rodadas com o valor 1000.

4. Os modelos a serem implementados nessa etapa serão: **MQO tradicional**, **MQO regularizado** (Tikhonov) e **Média de valores observáveis**.

5. Como o modelo regularizado depende da definição do valor λ, é de interesse encontrar aquele que tem o valor médio mínimo de EQM. Discuta qual foi o valor encontrado.

6. Para validação de tais modelos, em cada rodada deve-se embaralhar as amostras do conjunto de dados e em seguida realizar o particionamento em 80% dos dados para treinamento e 20% para teste.

7. Os dados selecionados para teste, são utilizados para validar o modelo. Assim, é necessário computar o Erro Quadrático Médio (EQM) e armazenar essa medida em uma lista/vetor que representa o EQM em cada uma das rodadas.

8. Ao final das 1000 rodadas calcule para cada modelo utilizado, compute a média, desvio-padrão, valor maior, valor menor de cada EQM. Coloque esses valores em um gráfico ou tabela e discuta os resultados obtidos.

Extra - Regressão

Os resultados obtidos pelo EQM, EQM regularizado e Média, são satisfatórios? 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.

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). Neste presente conjunto de dados, tem-se 50 000 observações para os dois sensores, em classes totalmente balanceadas, ou seja, 10 000 para cada classe.

O arquivo fornecido via AVA, trata-se de um .csv contendo 10 rodadas de aquisições dos sinais de EMG. Para cada rodada, as aquisições dos sinais foram realizadas na seguinte ordem: 1000 dados do gesto **neutro**, 1000 dados do gesto **sorrindo**, 1000 dados do gesto **aberto**, 1000 dados do gesto **surpreso** e 1000 dados do gesto **Rabugento**.

Pede-se inicialmente que faça a identificação de p (número de preditores), N(Quantidade de amostras) e c(Quantidade de classes). Em seguida, após acessar os dados no arquivo deve-se organizar a massa de dados $\mathbf{X}$ e $\mathbf{Y}$ da seguinte maneira:

$$\mathbf{X} \in \mathbb{R}^{N \times P} \qquad \mathbf{Y} \in \mathbb{R}^{N \times C}$$

Após o download, faça o que se pede:

1. Faça uma visualização inicial dos dados através do gráfico de espalhamento. Nessa etapa levante hipóteses sobre quais serão as características de um modelo que consegue separar as classes do problema.

2. Para validar os modelos utilizados na tarefa de classificação, é necessário definir uma quantidade específica de rodadas de treinamento e teste dos modelos. Assim, defina essa quantidade de rodadas com o valor 100.

3. Os modelos a serem implementados nessa etapa serão: **MQO tradicional**, **MQO regularizado** (Tikhonov), Classificador $k-$Vizinhos mais Próximos ($k-$NN) e Distância Mínima ao Centróide (DMC).

4. Como os modelos de regularização necessitam da definição de seus hiperparâmetros, é de interesse encontrar aquele que tem o valor médio maior de Acurácia. Discuta qual foi o valor encontrado para OLS (Tikhonov) e k elementos para o $k-$NN.

5. Para validação de tais modelos, em cada rodada deve-se embaralhar as amostras do conjunto de dados e em seguida realizar o particionamento em 80% dos dados para treinamento e 20% para teste.

6. Os dados selecionados para teste, são utilizados para validar o modelo. Assim, é necessário computar a acurácia de cada modelo e armazenar essa medida em uma lista/vetor que representa a acurácia em cada uma das rodadas.

7. Ao final das 100 rodadas calcule para cada modelo utilizado, compute a média, desvio-padrão, valor maior, valor menor e moda de cada acurácia. Coloque esses valores em um gráfico ou tabela e discuta os resultados obtidos.

Observações

- O trabalho pode ser desenvolvido em equipe de no máximo dois alunos.
- A nota do trabalho é divida da seguinte maneira:

1. 40% Relatório E implementações.
2. 60% Arguição.

- O total referente aos 40% (Relatório e implementações) é dividido da seguinte maneira:
  - Título (1%).
  - Resumo (15%).
  - Introdução (20%).
  - Desenvolvimento (27%).
  - Resultados (27%).
  - Referências (5%).
  - **Implementações** (5%).
- 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%).**