Lista de Exercícios - Inteligência Artificial (PCS3438)
Prof. Eduardo R. Hruschka

Observações:

- Pode-se usar softwares (e.g., R, Python, Weka etc.) para resolver as questões.
- Todos os arquivos CSV possuem cabeçalho com o nome das colunas e campos separados por vírgula ",".
- Todas as questões têm pesos iguais e valem 2 pontos.
- Data máxima para entrega: 11/11 no horário da aula.

Questão 1) Considerando os dados presentes no arquivo `class01.csv`, treine o algoritmo *Naive Bayes* Gaussiano utilizando a metodologia de validação cruzada *holdout* (utilize para treino as 350 primeiras linhas e para validação as demais). Qual o valor da acurácia a base de treino? Qual o valor da acurácia a base de validação?

Questão 2) Considerando os dados presentes no arquivo `class02.csv`, treine o algoritmo *10-Nearest Neighbors* (*KNN* com $k = 10$ e distância Euclidiana), utilizando a metodologia de validação cruzada *k-fold* com 5 *folds* não estratificados. Considere que a primeira pasta de validação seja formada pelas primeiras 20% linhas do arquivo, que a segunda pasta de validação seja formada pelas 20% linhas seguintes, e assim por diante, até atingir a última pasta, formada pelas 20% linhas finais da base. Qual o valor médio da acurácia para a base de validação?

Questão 3) Considerando os dados presentes no arquivo `reg01.csv`, obtenha um modelo de regressão linear com regularização L1 (*LASSO* com $\alpha = 1$) utilizando a metodologia *Leave-One-out*. Qual o valor médio do Root *Mean Squared Error* (*RMSE*) para a base de treino e para a base de validação?

Questão 4) Considerando os dados presentes no arquivo `reg02.csv`, treine uma árvore de regressão (sem realizar podas) com quebras baseadas no erro quadrático médio (do inglês *MSE - Mean Squared Error*) utilizando a metodologia de validação cruzada *k-fold* com $k = 5$. Qual o valor do *Mean Absolute Error (MAE*) para a base de treino? Qual o valor médio do *MAE* para a base de validação?

Questão 5) Assinale as alternativas com V ou F para Verdadeiro ou Falso respectivamente. **Atente para o fato que uma questão errada anula uma certa.** Em caso de dúvidas, deixe em branco.

( ) Quando ajustamos um modelo linear, geralmente supomos que os erros tem distribuição normal e são independentes e identicamente distribuídos (i.i.d.).
( ) Quando ajustamos um modelo de regressão, podemos utilizar os valores preditos e os resíduos do modelo para avaliar se o modelo se adequa bem aos dados.
( ) O coeficiente de determinação (r^2) indica, em termos percentuais, quanto da variabilidade da variável resposta é explicada pelas covariáveis do modelo.
( ) Os modelos de regressão não são afetados por observações atípicas (*outliers*) e valores faltantes.
( ) Considerando um modelo de regressão simples, temos que o coeficiente associado à covariável representa o grau de inclinação da reta.
( ) Para efetuar regressão com o algoritmo *KNN*, pode-se fazer uma votação simples dos valores dos k vizinhos encontrados.
( ) Para melhor desempenho da árvore de regressão, pode-se utilizar regressões lineares em suas folhas para previsão do valor final.
( ) No algoritmo *Random Forest* para regressão, o valor predito é obtido pela média dos valores encontrados em cada árvore.