Processo Seletivo Itaú-Unibanco

Avaliação sobre Aprendizado de Máquina e Ciência de Dados

Observações:

- A interpretação das questões é parte integrante da avaliação.
- Sempre que julgar apropriado, você pode usar softwares (e.g., R, Python, Weka, etc.) para resolver as questões.
- O tempo de realização da prova será usado em sua avaliação. Inicie a prova imediatamente após o recebimento e envie suas respostas o mais rapidamente possível. O tempo máximo de prova é de 3 horas.
- Nas questões de Verdadeiro/Falso, cada questão errada anula uma certa. Em caso de dúvidas deixe em branco.
- Todos os arquivos CSV possuem cabeçalho com o nome das colunas e campos separados por vírgula ",".
- Todas as questões têm pesos iguais e valem 1 ponto.

Módulo 1 – Agrupamento

Questão 1) Considerando os dados presentes no arquivo `clus01.csv`, execute o algoritmo *K-Means* com o posicionamento inicial dos centroides sendo: [1, 1, 1, 1], [-1, -1, -1, -1] e [1, -1, 1, -1]. Qual é o posicionamento final de cada centroide após 10 iterações?

Questão 2) 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.

( ) O número mínimo de grupos no algoritmo *K-Means* é igual à raiz quadrada do número de elementos da base.

( ) Devido à inicialização aleatória, o resultado final do *K-Means* sempre será o mesmo independentemente do critério de convergência.

( ) O algoritmo *K-Median* é mais sensível a outliers do que os algoritmos *K-Means*, *K-Mode* e *K-Medoid*.

( ) O algoritmos *Single-Linkage, Complete-Linkage e Average-Linkage* são exemplos de algoritmos hierárquicos.

( ) Não é possível identificar *outliers* com algoritmos hierárquicos e, por este motivo, tais algoritmos requerem que os *outliers* sejam removidos numa etapa necessária de pré-processamento.

( ) Para conjuntos de dados com milhões de linhas é aconselhada a utilização de algoritmos hierárquicos em vez de algoritmos particionais pois aqueles calculam somente uma única matriz de distância.

( ) O Rand Index pode ser utilizado para mensurar a aderência entre o agrupamento obtido e um agrupamento de referência (por exemplo, dado por um especialista do domínio).

( ) O *DBScan* é um exemplo de algoritmo de agrupamento baseado em densidade.

( ) Para acelerar o tempo de processamento dos algoritmos de agrupamento baseados em densidade, pode-se calcular a distância de cada elemento para o centroide de cada um dos grupos.

( ) O número mínimo de variáveis necessárias para realizar um agrupamento é 1.

Módulo 2 – Classificação

Questão 3) 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 para a base de treino? Qual o valor da acurácia para a base de validação? Faça o mesmo treinamento com a metodologia *Leave-One-Out*. Qual o valor da acurácia média para a base de treino? Qual o valor da acurácia média para a base de validação?

Questão 4) 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 10 *folds* não estratificados. Considere que a primeira pasta de validação seja formada pelas primeiras 10% linhas do arquivo, que a segunda pasta de validação seja formada pelas 10% linhas seguintes, e assim por diante, até atingir a última pasta, formada pelas 10% linhas finais da base. Qual o valor médio da acurácia 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.

( ) No método *KNN*, o melhor valor de k é igual a n (n = número de exemplos), porém dado o custo computacional, valores pequenos de k são preferidos.
( ) As folhas de uma Árvore de Decisão são obtidas objetivando-se minimizar a pureza entre classes.
( ) Uma Árvore de Decisão completa (profundidade máxima possível) tem maior chance de fazer *overfitting* nos dados do que uma árvore com profundidade limitada.
( ) Para utilizar uma rede neural em um problema multiclasse é necessário o uso de estratégias como *One-Vs-One* e *One-Vs-Rest*.
( ) *Root Mean Squared Error* (RMSE) é uma medida adequada para avaliar classificadores.
( ) Todo algoritmo de classificação possui um viés que dita a forma como cada método explora o espaço de busca pela hipótese que melhor se ajusta aos dados.
( ) O grau do polinômio do algoritmo *Support Vector Machine* (SVM) deve ser definido para todos os *kernels*.
( ) Uma vantagem dos algoritmos baseados em árvores é que, em sua maioria, a presença de valores faltantes (*missing*) não inviabiliza sua execução.
( ) Utiliza-se validação *out-of-time* separando um conjunto de dados de forma aleatória independentemente do tempo.
( ) O Algoritmo MLP (*Multi-Layer Perceptron*) é conceitualmente baseado em uma visão abstrata e simplificada de um neurônio biológico

---

Módulo 3 – Regressão

Questão 6) Considerando os dados presentes no arquivo `reg01.csv`, obtenha um modelo de regressão linear com regularização L1 (*LASSO* com parâmetro de regularização igual a 1) utilizando a metodologia *Leave-One-out*. Qual o valor médio do Root *Mean Squared Error* (*RMSE*) para a base de treino? Qual o valor médio do *RMSE* para a base de validação?

Questão 7) 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 $k = 10$. 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 8) 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 efetuarmos regressão com o algoritmo *KNN*, é aconselhado 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.
( ) A F1 é uma medida adequada para avaliar algoritmos de regressão.
( ) Em todos os modelos de regressão, a métrica r^2 é igual ao quadrado da correlação de pearson entre o valor predito e o observado.
( ) No algoritmo *Random Forest*, uma possibilidade simplista para obtenção do valor final é calcular a média dos valores encontrados em cada árvore.

Módulo 4 – Estatística

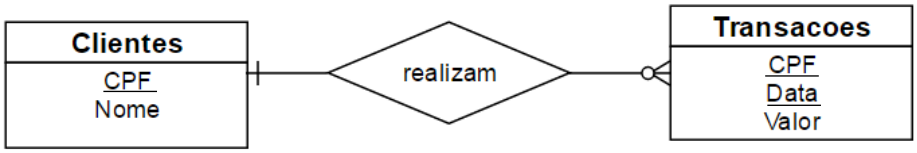
Questão 9.a) Calcule a distância máxima entre as funções de distribuição acumulada empírica das seguintes amostras (5, 3, 3, 11, 8, 7, 1, 5, 4, 9) e (2, 1, 1, 4, 3, 1, 1, 1, 3, 2).
Questão 9.b) Um analista possui as seguintes informações a respeito dos valores de uma amostra de 30 observações:

- A média de todos os valores é igual a 2.96
- A soma dos quadrados dos valores é igual a 268

Calcule o desvio padrão amostral.

Módulo 5 – SQL

Questão 10) Suponha a existência de duas tabelas de dados conforme o Modelo Entidade Relacional (MER) abaixo:



1. Assinale a(s) alternativa(s) correta(s) que retornam os 100 clientes que transacionaram mais dinheiro.

   a) `SELECT a.* FROM (SELECT cpf, SUM(valor) as acumulado FROM transacoes GROUP BY cpf ORDER BY acumulado DESC) a limit 100;`

   b) `SELECT SUM(Transacoes.Valor), clientes.CPF FROM Clientes INNER JOIN Transacoes ON Clientes.CPF=Transacoes.CPF GROUP BY clientes.CPF ORDER BY SUM(Transacoes.Valor) DESC LIMIT 100;`

   c) `SELECT cpf, valor FROM transacoes order by cpf limit 100;`

   d) `SELECT clientes.cpf, transacoes.valor FROM clientes LEFT JOIN transacoes ON clientes.cpf = transacoes.cpf ORDER BY transacoes.valor DESC LIMIT 100;`

2. Assinale a(s) alternativa(s) correta(s) que retorna(m) a tabela Clientes com mais uma coluna. Esta coluna contem a marcação 1 para aqueles que transacionaram mais do que R$10.000,00 ao longo do tempo e 0 caso contrário.

   a) `SELECT clientes.*, case when acumulado is not null then 1 else 0 END as gdes_valores FROM clientes left join (SELECT * FROM (SELECT cpf, sum(valor) as acumulado FROM transacoes group by cpf) trans where acumulado>10000) transf on transf.cpf=clientes.cpf;`

   b) `SELECT clientes.CPF, clientes.Nome, CASE WHEN SUM(Transacoes.Valor) > 10000 THEN 1 ELSE 0 END as gdes_valores FROM Clientes INNER JOIN Transacoes ON Clientes.CPF=Transacoes.CPF GROUP BY Clientes.CPF, Clientes.nome;`

   c) `SELECT clientes.cpf, clientes.nome, case when valor>10000 then 1 else 0 END as gdes_valores FROM clientes left join transacoes on transacoes.cpf=clientes.cpf;`

   d) `SELECT /*+ MAPJOIN(clientes) */ clientes.*, case when acumulado is not null then 1 else 0 END as gdes_valores FROM clientes left join (SELECT * FROM (SELECT cpf, sum(valor) as acumulado FROM transacoes group by cpf) trans where acumulado>10000) transf on transf.cpf=clientes.cpf;`