

 UNIVERSIDADE FEDERAL DE SANTA CATARINA		Campus Araranguá Rodovia Governador Jorge Lacerda, 3201, Jardim das Avenidas Araranguá - Santa Catarina – Brasil / CEP 88900-000 www.ararangua.ufsc.br / +55 (48) 3721.6448		1/4
Avaliação II				
Disciplina	Tópicos Especiais III (DEC7553)		08655	2021.2
Aluno(a)s	Leonardo - Rian			

1) Relacione os conceitos da 1ª coluna com as respectivas definições na 2ª coluna – (0,75):


[1]	Descoberta de Conhecimento em Bases de Dados	[18]	Cria árvores de decisão a partir da seleção randômica de exemplos dos dados e obtém a predição de cada árvore selecionando a melhor solução.
[2]	Pré-processamento	[11]	Técnica de classificação que, através de um processo de otimização heurística e baseado na teoria da informação, seleciona iterativamente determinadas variáveis que promovem a melhor separação de classes de acordo com alguma função de custo, visando criar regiões disjuntas e estabelecendo uma fronteira de decisão.
[3]	Aprendizado de Máquina	[1]	Processo não trivial, interativo e iterativo, visando a identificação de padrões compreensíveis, que sejam válidos, novos, potencialmente úteis a partir de grandes conjuntos de dados.
[4]	Aprendizado Supervisionado	[8]	Objetiva descobrir um relacionamento entre os atributos precursores e o atributo meta, usando registros cuja classe é conhecida, para se construir um modelo que possa ser aplicado a objetos ainda não classificados.
[5]	Aprendizado Não Supervisionado	[13]	Processo pela qual determinado conjunto de dados de teste é dividido em N subconjuntos e a partir desses N subconjuntos são realizados N testes.
[6]	Avaliação	[16]	Expressa a distância, similaridade, correlação ou associação entre dois objetos quaisquer.
[7]	Pós-processamento	[15]	Proporção de exemplos de uma classe (neste caso valores positivos) que foram corretamente classificados pelos exemplos classificados incorretamente na classe - $TP / (TP + FP)$.
[8]	Tarefa de Classificação	[14]	Estrutura que, na tarefa de classificação envolvendo classes de valores discretos, permite estimar a qualidade dos resultados obtidos.
[9]	Tarefa de Agrupamento	[4]	Processo pelo qual determinado algoritmo de aprendizado (indutor) recebe um conjunto de exemplos de treinamento para os quais os rótulos da classe associada são conhecidos e, a partir disso, constrói um modelo visando a correta determinação da classe para novos exemplos ainda não rotulados.
[10]	Tarefa de Associação	[24]	Conjunto de ferramentas e técnicas com o objetivo de oferecer sugestões sobre determinado item para um usuário.
[11]	Árvore de Decisão	[22]	Técnica baseada na noção intuitiva de que determinados padrões de interação, representados na forma de um grafo, são importantes características capazes de descrever o comportamento de unidades de análise.
[12]	ID3	[10]	Objetiva descobrir relacionamentos importantes em um conjunto de dados, tal que, a presença de um item em uma determinada transação irá implicar na presença de outro item na mesma transação.

 UNIVERSIDADE FEDERAL DE SANTA CATARINA		Campus Araranguá Rodovia Governador Jorge Lacerda, 3201, Jardim das Avenidas Araranguá - Santa Catarina – Brasil / CEP 88900-000 www.ararangua.ufsc.br / +55 (48) 3721.6448		2/4
Avaliação II				
Disciplina	Tópicos Especiais III (DEC7553)		08655	2021.2
Aluno(a)s				

[13]	<i>Cross-validation</i>	[5]	Processo pelo qual o algoritmo indutor analisa os exemplos fornecidos e tenta determinar quais destes podem ser agrupados de alguma maneira, formando agrupamentos ou <i>clusters</i> .
[14]	Matriz de Confusão	[23]	Medida que avalia a frequência com que determinado nodo aparece no caminho mais curto entre dois nodos quaisquer.
[15]	Precisão (<i>Precision</i>)	[2]	Etapa que visa a preparação e a transformação de dados de modo que possam ser utilizados por algoritmos de mineração de dados e, deste modo, conduzir à descoberta de conhecimento.
[16]	Medidas de Similaridade	[19]	Conceito criado para unificar a estatística, análise de dados, aprendizado de máquina e seus métodos relacionados visando entender e analisar fenômenos reais com dados.
[17]	Centralidade de Proximidade	[6]	Processo pelo qual se objetiva aferir a qualidade da aplicação/utilização de algoritmos de Mineração de Dados sobre determinado conjunto de dados.
[18]	<i>Random Forest</i>	[9]	Objetiva servir como um passo anterior à tarefa de classificação quando não se possui um conjunto de dados previamente classificado, possibilitando a reunião de itens semelhantes em determinado grupo.
[19]	Ciência de Dados	[3]	Um programa aprende a partir da experiência E, em relação a uma classe de tarefas T, com medida de desempenho P, se seu desempenho em T, medido por P, melhora com E.
[20]	Dendrograma	[7]	Etapa do processo de KDD que consiste na avaliação dos resultados obtidos, isto é, analisa se o conhecimento descoberto é relevante ou não.
[21]	Rede	[25]	Determina a porcentagem de determinado <i>itemset</i> dentre todas as transações da Base de Dados.
[22]	Análise de Rede Social	[17]	Representa a distância natural entre todos os pares de nodos, definida pelo tamanho dos caminhos mais curtos.
[23]	Centralidade de Intermediação	[20]	Árvore que iterativamente divide o conjunto de dados em subconjuntos menores até que cada subconjunto consista de somente um objeto.
[24]	Sistemas de Recomendação	[12]	Algoritmo baseado na teoria da informação que constrói uma árvore de decisão onde cada vértice (nodo) corresponde a um atributo, e cada aresta da árvore a um valor possível do atributo.
[25]	Medida do Suporte	[21]	Conjunto de nodos conectados por <i>links</i> /arestas que podem ou não serem direcionados.

 UNIVERSIDADE FEDERAL DE SANTA CATARINA		Campus Araranguá Rodovia Governador Jorge Lacerda, 3201, Jardim das Avenidas Araranguá - Santa Catarina – Brasil / CEP 88900-000 www.ararangua.ufsc.br / +55 (48) 3721.6448		3/4
Avaliação II				
Disciplina	Tópicos Especiais III (DEC7553)		08655	2021.2
Aluno(a)s				

- 2) Considerado o conjunto `tae.csv` e, utilizando a linguagem Python e as bibliotecas apresentadas na disciplina, elabore uma árvore de decisão. Após isso reduza a profundidade da árvore. Apresente a acurácia inicial e após a redução da profundidade. Também apresente duas regras geradas pela árvore. O conjunto de dados representa avaliações de desempenho no ensino ao longo de alguns semestres e possui as colunas 'ta_native', 'course_instr', 'course', 'summer_regular', 'class_size' e 'label'. A coluna 'label' representa o atributo meta, ou seja, o objetivo da classificação - (1,0).
- 3) Utilizando o mesmo conjunto de dados (`tae.csv`) e, utilizando a linguagem Python e as bibliotecas apresentadas na disciplina, desenvolva um código com base no algoritmo de aprendizado de máquina do tipo *Random Forest*. Após aferir a acurácia realize algumas previsões com dados ainda não utilizados. Na sequência calcule a contribuição de cada característica (atributo) e realize novamente o treinamento e teste sem a característica menos importante. Ao final apresente as acurácias resultantes da primeira etapa (com todas as características) e da segunda etapa (desconsiderando a característica menos relevante) – (1,0).
- 4) Utilizando o mesmo conjunto de dados (`tae.csv`) e, utilizando a linguagem Python e as bibliotecas apresentadas na disciplina, elabore um algoritmo de aprendizado de máquina do tipo *k-NN* estabelecendo uma vizinhança (*k*) de 5 e 7 elementos/objetos. Calcule a acurácia para as duas situações e apresente os valores para os dois *ks*. Após aferir a acurácia realize algumas previsões com dados ainda não utilizados – (1,0).
- 5) Utilizando o mesmo conjunto de dados (`tae.csv`) e, utilizando a linguagem Python e as bibliotecas apresentadas na disciplina, elabore um algoritmo de aprendizado de máquina do tipo *Naive Bayes*. Calcule e apresente o valor da acurácia. Após aferir a acurácia realize algumas previsões com dados ainda não utilizados – (0,75).
- 6) Utilizando o conjunto de dados (`tae.csv`) e, utilizando a linguagem Python e as bibliotecas apresentadas na disciplina, elabore um algoritmo de aprendizado de máquina do tipo *k-means*. Calcule o valor da acurácia. Na sequência normalize os valores utilizando alguma transformação e calcule novamente a acurácia. Ao final apresente as duas acurácias calculadas, assim como os resultados de dois métodos de determinação do número de agrupamentos. Lembrando que a acurácia somente pode ser calculado devido ao *dataset* de classificação – (1,0).
- 7) Utilizando o conjunto de dados (`wine.csv`) e, utilizando a linguagem Python e as bibliotecas apresentadas na disciplina, elabore um algoritmo de aprendizado de máquina do tipo *Random Forest*. Na sequência calcule a contribuição de cada característica e realize novamente o treinamento e o teste sem algumas das características menos importantes. Considerando este novo conjunto de dados elabore o algoritmo *k-means* aplicando uma transformação nos dados. Ao final apresente as acurácias dos dois algoritmos, *Random Forest* e *K-means*. O conjunto de dados representa análises de vinho e possui as colunas 'label', 'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash', 'Magnesium', 'Total phenols', 'Flavanoids', 'Nonflavanoid phenols', 'Proanthocyanins', 'Color intensity', 'Hue', 'OD280/OD315 of diluted wines', 'Proline'. A coluna 'label' representa o atributo meta, ou seja, o objetivo da classificação. A coluna 'Proline' deve ser descartada para todas as análises – (1,25).
- 8) Utilizando o conjunto de dados '*breast cancer*' através do método **load_breast_cancer()** disponível na biblioteca `sklearn.datasets`, realize uma análise utilizando algum dos algoritmos de classificação ou agrupamento estudados. Ao final apresente a acurácia – (0,75).
- 9) Elabore um grafo com um conjunto de nodos (mais de 15) em que fique clara a separação em três grupos. Os grupos devem estar conectados somente por um dos nós de cada grupo. Após isso calcule as métricas de

 UNIVERSIDADE FEDERAL DE SANTA CATARINA		Campus Araranguá Rodovia Governador Jorge Lacerda, 3201, Jardim das Avenidas Araranguá - Santa Catarina – Brasil / CEP 88900-000 www.ararangua.ufsc.br / +55 (48) 3721.6448			4/4
Avaliação II					
Disciplina	Tópicos Especiais III (DEC7553)			08655	2021.2
Aluno(a)s					

centralidade de intermediação (*betweenness centrality*) e centralidade de proximidade (*closeness centrality*). Ao final apresente uma relação dos nós (5 nós, por exemplo) mais importantes ordenados dos mais relevantes para os menos relevantes, para as métricas de centralidade de intermediação e proximidade - (1,0).

10) Elabore um código em Python para um sistema de recomendação simples que integre as abordagens de filtragem colaborativa e baseada em conteúdo respeitando os seguintes passos – (1,5):

- Elabore uma matriz Termo X Documento (por exemplo, 15 termos e 10 documentos) com pesos já normalizados entre 0.0 e 1.0. Distribua os pesos de maneira esparsa;
- A partir da matriz calcule uma matriz de similaridades entre os documentos utilizando a métrica do cosseno;
- Elabore uma segunda matriz indicando usuários e documentos que estes usuários já leram, por exemplo:

Usuários	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
1	X		X		X		X		X	
2					X	X				
3	X	X		X	X			X	X	
4		X	X			X	X			X
5		X	X	X					X	X

- A partir desta matriz elabore uma matriz de similaridades entre os usuários. A similaridade pode ser obtida pela quantidade de documentos em comum que dois usuários quaisquer leram;
- Escolha um usuário qualquer (por exemplo, Usuário 1) e, a partir disso, obtenha o usuário mais similar (por exemplo, Usuário 3). Caso seja retornado mais de um usuário com a mesma similaridade máxima escolha o primeiro. Levando em conta o usuário recuperado (Usuário 3) identifique quais documentos este usuário leu e o Usuário 1 ainda não leu. Então, considerando estes documentos localize qual é o mais similar em relação aos documentos que o Usuário 1 já leu e recomende este documento.

Obs: Documente adequadamente cada um dos passos no código visando facilitar a correção.