Faculdade de Tecnologia
Universidade Estadual de Campinas

CET0621 – Aprendizado de Máquina na Análise de Dados
Lista de Exercícios 2

Nome do Aluno 1: **RA:**

Nome do Aluno 2: **RA:**

Nome do Aluno 3: **RA:**

Instruções

- Esta lista de exercícios deve ser desenvolvida em **trios**;
- As respostas devem estar em um arquivo PDF, juntamente com o **nome e RA dos membros do trio**;
- **Apenas um dos membros** deve postar o arquivo no Moodle;
- Esta lista contém questões teóricas e práticas.

Questões:

1) (3,0pt) Em uma clínica médica, especializada em uma determinada doença infecciosa, existe uma base de dados de pacientes que armazena informações referentes a **quatro atributos**: idade do paciente (valor numérico); nível de duas substâncias A e B no sangue (valores numéricos); e presença ou não de histórico da doença na família (valor binário). Cada amostra desta base pode ser de um paciente doente ou não (classes do problema), e tal informação também é dada. Diante deste cenário, pede-se:

   a) Considerando-se a utilização de uma rede neural do tipo **Perceptron Multicamadas** (MLP) como classificador de dados, supondo que sejam adotados **10 (dez) neurônios** na **única camada oculta** da rede e que **todos os neurônios tenham *bias***, indique o **número total de entradas** e **de saídas** que você definiria para esta rede e calcule o **número total de pesos** a serem ajustados durante a etapa de treinamento.

   b) Ao treinar a MLP indicada no item (a) você observou que, apesar da rede apresentar uma acurácia de 98% para os dados de treinamento, sua taxa de acertos cai para menos de 60% para os dados de teste. Diante disso pergunta-se: o que poderia estar acontecendo? Que medidas você adotaria para tentar sanar este problema?

   c) Considerando os atributos da base de dados descritos no enunciado e seu conhecimento sobre MLPs, quais **etapas de pré-processamento** você aplicaria neste cenário?

2) (2,0pt) A base de dados apresentada na Tabela 1 contém 14 amostras associadas a um problema de classificação. Na Tabela 1 são apresentados também os rótulos reais de cada amostra e os rótulos atribuídos por três classificadores. Neste contexto, pede-se:

   a) Apresente os rótulos que seriam atribuídos a cada amostra dos dados por um *ensemble*, baseado em **voto majoritário**, formado pelos três classificadores cujas saídas foram apresentadas;

   b) Apresente as matrizes de confusão construídas a partir dos resultados de cada classificador e do *ensemble*, em conjunto com as respectivas acurácias;

   c) Compare e discuta os resultados.

Tabela 1 - Base de dados para um problema de classificação de dados.

ID	Classe Real	Classificador 1	Classificador 2	Classificador 3	Ensemble
1	A	A	B	A	
2	A	B	A	A	
3	B	A	B	A	
4	B	B	B	A	
5	B	A	B	B	
6	A	B	B	B	
7	B	B	B	B	
8	A	A	A	A	
9	A	A	B	A	
10	B	B	B	A	
11	A	A	A	B	
12	B	B	B	B	
13	B	B	B	A	
14	A	A	B	B	

3) (5,0pt) Utilizando alguma **ferramenta computacional ou biblioteca de sua preferência** (como Weka, scikit-learn e Orange), realize um estudo comparativo entre o desempenho dos algoritmos *MLP*, *Ensemble* de árvores de decisão e *Ensemble* de MLPs (ambos construídos via estratégia de *bagging*) quando aplicados ao conjunto de dados conhecido como *Wine* (https://archive.ics.uci.edu/ml/datasets/Wine), disponível no *UCI Repository of Machine Learning Datasets*[1]. **Observação**: caso você opte por utilizar o Weka, os *ensembles* podem ser encontrados na categoria "META" de classificadores, enquanto as MLPs estão na categoria "FUNCTIONS".

   Para este estudo, pede-se:

   - Para avaliar cada algoritmo, adote a estratégia de validação cruzada com 10 pastas;
   - Descreva **detalhadamente** a metodologia experimental empregada. Apresente as etapas de pré-processamento utilizadas e, para cada algoritmo, os valores usados nos parâmetros.
   - Avalie se diferentes estratégias de pré-processamento levam a resultados diferentes.
   - Apresente as matrizes de confusão para cada algoritmo/experimento, aplicadas aos subconjuntos de teste, juntamente com as principais métricas de avaliação.
   - Discuta os resultados obtidos e compare-os com os dos classificadores treinados na Lista 01 da disciplina.

[1] Para facilitar a utilização do Weka, uma versão do conjunto de dados *Wine*, já em formato *.arff*, pode ser encontrada na pasta "*classification*" de https://github.com/renatopp/arff-datasets.