Inteligência Artificial
Trabalho Prático

CCENS UFES – Departamento de Computação
Prof. M. Sc. Jacson Rodrigues Correia da Silva

Introdução ao Aprendizado de Máquina

O trabalho consiste na aplicação e análise de três métodos de aprendizado de máquina para classificar corretamente um conjunto de dados. As equipes já foram divididas em sala de aula e seus conjuntos de dados (*datasets*) já foram sorteados, sendo:

1. Marcos e Leonardo (*dataset* 7)
2. Rafael e Thiago (*dataset* 3)
3. Brenno e Guilherme (*dataset* 2)
4. Andrei e Jeferson (*dataset* 8)
5. Erik e Arthur (*dataset* 4)
6. José e Wagner (*dataset* 5)
7. Natália e João Paulo Borges (*dataset* 6)
8. João Paulo, Jessé e Breno Batista (*dataset* 1)

Os *datasets* (conjuntos de dados) homologados estão disponíveis em: <https://goo.gl/P8XxFu>

Os cálculos de variância, covariância, média, autovalores, autovetores, etc., não precisam ser feitos a mão. Podem ser utilizados o Octave, ou o sklearn, ou outro programa/biblioteca.

O que deve ser feito

Árvore de Decisão

Passos:

1. Inicialmente, obtenha seu *dataset* e faça uma análise da variância de seus atributos. (Ajudinha com variância: <https://goo.gl/c9gMU5>)
2. Ordene-os do maior para o menor e inicie (faça somente com dois ou três atributos) a construção da árvore de decisão com *N* (essa quantidade é de sua escolha) elementos de seu conjunto.
3. Agora, ordene-os do menor para o maior e inicie (faça com a mesma quantidade de atributos do passo 2) a construção da árvore de decisão com os mesmos *N* elementos de seu conjunto.
4. Analise qual dos dois passos (2 ou 3) foi melhor para esse algoritmo.
5. Termine a árvore de decisão continuando o passo 2 ou o passo 3.
6. Construa um relatório contendo informações sobre:
   i. o que pôde ser concluído com a variância do *dataset*;
   ii. como a análise da variância pode ajudar na construção da árvore de decisão;
   iii. a ilustração da árvore de decisão criada;
   iv. as informações do *dataset* que ficaram visíveis na árvore de decisão.

Pré-processamento do *dataset*

1. Transforme os dados simbólicos de seu *dataset* para numéricos.
2. Após aleatorizar o *dataset*, divida-o em: Teste (30% dos dados); e Treino (70% dos dados).
3. Obtenha a média (μ) e o desvio padrão (σ) dos dados de Treino.
4. Aplique a padronização (*standardization*) sobre todos os dados:

$$x_{novo} = \frac{x_{velho} - \mu}{\sigma}$$

5. Sobre a matriz de covariância do conjunto de Treino, encontre os autovalores e autovetores.
   (Ajudinha com autovalores e autovetores: <https://goo.gl/Caq4rg>)
6. Escolha os atributos $\{a_i, \ldots, a_k\}=N$ que tem maior importância no conjunto de dados.
7. Reduza a quantidade de dimensões para a *N*, multiplicando os todos os dados obtidos no passo 4 pelos autovetores dos atributos escolhidos.
8. Salve os dados gerados em todos os passos, pois eles serão necessários depois.
9. Utilize esse *dataset* preprocessado para fazer a Rede Neural e também o k-NN.
10. Adicione ao relatório as informações:
    i. como a quantidade de dimensões do problema pôde ser reduzida;
    ii. qual o novo conjunto de dados;
    iii. se possível, plote o gráfico do novo conjunto de dados;
    iv. como preprocessar novos dados que ainda não fazem parte do conjunto de dados atual.

Rede Neural Artificial (RNA)

1. Esboce uma RNA no papel, definindo e escolhendo:
    i. a "camada" de entrada para seu *dataset*;
    ii. a quantidade de neurônios ocultos que será testada.
    Você deve testar com várias quantidades de neurônios ocultos.
    Então, faça uma tabelinha com seus chutes e depois vá anotando quais foram os resultados.
    iii. a quantidade de neurônios de saída.
    iv. A função de ativação da camada oculta e da camada de saída.
    As funções também devem ser modificadas para ter diferentes resultados.
    Então, também faça uma tabelinha com seus chutes e depois vá anotando os resultados.
    As funções de ativação podem ser: <https://goo.gl/Aat9k7>
2. Utilizando a biblioteca FANN (*Fast Artificial Neural Network*), crie uma rede neural para seus dados.
3. Faça o treinamento da rede neural. Para isso, treine várias vezes, alternando entre as configurações (quantidade de neurônios ocultos e funções de ativação) escolhidos no passo 1.
4. Vá anotando os resultados obtidos no passo 3.
5. Agora, escolha a rede com o melhor resultado, treine e teste ela 10 vezes.
   Para cada uma das 10 vezes, anote sua acurácia (taxa de acertos) sobre o conjunto de testes.
6. Calcule a média e o desvio padrão das acurácias do passo 5.
7. Adicione ao relatório as informações:
    i. qual foi o esboço da RNA e quais os parâmetros que foram definidos no passo 1;
    ii. quais foram os resultados obtidos no passo 3;
    iii. um gráfico com todas acurácias obtidas no passo 5;
    iv. qual foi a melhor rede encontrada e qual sua acurácia (média e desvio padrão);

Obs.: se a RNA tiver mais de um neurônio de saída, você deve obter os resultados da rede e escolher o maior dentre eles. Exemplo de RNA com três saídas:
Valores de entrada: $X^1 = [x_1, x_2, x_3, x_4]$
Valores de saída desejado: y = 2, para $y \in \{0,1,2\} \rightarrow y = [y_1\ y_2\ y_3] \rightarrow y = [0\ 0\ 1]$
Valores obtidos na RNA: $\hat{y} = [0{,}3\ 0{,}1\ 0{,}7]$
Então:

$$\hat{y}_n = \frac{e^n}{\sum_{i=0}^{3} e^{\hat{y}_i}} \qquad \hat{y}_1 = \frac{e^{0,3}}{\left(e^{0,3}+e^{0,1}+e^{0,7}\right)} = 0{,}30 \qquad \hat{y}_2 = \frac{e^{0,1}}{\left(e^{0,3}+e^{0,1}+e^{0,7}\right)} = 0{,}25 \qquad \hat{y}_1 = \frac{e^{0,7}}{\left(e^{0,3}+e^{0,1}+e^{0,7}\right)} = 0{,}45$$

$$\hat{y} = \text{argmax}\ \hat{y} = [0\ 0\ 1]$$

Assim:
Se desejado (y) = obtido ($\hat{y}$), então acerto.
Obtivemos um acerto, pois y = [0 0 1] é igual a $\hat{y}$ = [0 0 1].

Vizinhos mais próximos (k-NN)

1. Com o conjunto de treino já preprocessado:
   i. Execute o algoritmo k-NN com valores *k* diferentes.
   ii. Anote todos os resultados obtidos.
2. Agora, junte o conjunto de treino e o conjunto de testes preprocessados.
   i. Modifique o tamanho do conjunto treinamento.
   ii. Execute novamente o k-NN, alterando seu valor k.
   iii. Anote todos os resultados obtidos.
3. Agora, adicione e/ou remova dimensões do seu problema (use os autovetores salvos na etapa de preprocessamento dos dados). Fique à vontade e use sua intuição para escolher isso.
   i. Execute novamente o k-NN com essas novas dimensões;
   ii. Faça testes e escolha o melhor valor para *k* e também a melhor quantidade de dados do conjunto de treinamento.
   iii. Anote todos os testes.
4. Adicione ao relatório as informações:
   i. gráfico com os resultados do passo 1;
   ii. uma comparação do passo 1 com o melhor resultado da RNA;
   iii. gráfico com os resultados do passo 2;
   iv. descrição do que foi feito no passo 3;
   v. gráfico com os resultados do passo 3.

Finalização

Faça uma conclusão dizendo:

i. quais foram os resultados encontrados;
ii. uma tabela com os melhores resultados de cada método;
iii. dificuldades encontradas;
iv. aprendizado obtido;
v. ideias de como melhorar a acurácia desse conjunto de dados.

Detalhes de entrega

Data final: 28/12/2017 (qualquer trabalho recebido após essa data, receberá nota zero)
O que deve ser enviado ao e-mail do professor:

1. os códigos utilizados/implementados para fazer todas as etapas do trabalho.
2. o relatório seguindo as normas ABNT.

Vocês podem conversar entre si e discutir sobre o desenvolvimento do trabalho, mas evitem qualquer plágio.