UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
COLEGIADO DE CIÊNCIA DA COMPUTAÇÃO

Relatório Trabalho 1 – Aprendizagem de Máquina

Título: KNNJS
Acadêmicos: Gilberto Antunes & Henrique Tomé Data: 21/03/2019

---

1. DESCRIÇÃO DO QUE É O TRABALHO

O objetivo do trabalho consiste em implementar um classificador baseado em vizinhanças (KNN). Para isso, foi implementada a entrada de dados com base em arquivo em formato .csv, o qual é utilizado nos experimentos.

Para que o processo tenha base para análise, deverão ser executadas 32 repetições. Os valores a serem comparados deverão ser os valores médios das 32 execuções.

2. DESCRIÇÃO DO CONJUNTO DE DADOS

O conjunto de dados utilizado durante os experimentos consiste em uma base composta por 690 instâncias, as quais são divididas em apenas duas classes, 1 e 2. Dentre os 690 exemplares do conjunto, 383 pertencem à classe 1 e 307 à classe 2. Cada instância é composta por 14 atributos, todos do tipo float.

3. DESCRIÇÃO PASSO-A-PASSO DO EXPERIMENTO

O primeiro passo consiste no carregamento do conjunto de dados. Para tanto, implementou-se uma rota que carrega o arquivo desejado em formato .csv e o armazena em uma lista encadeada de objetos, em que cada objeto corresponde a uma tupla contendo os atributos de cada instância. O último atibuto de cada objeto contém a classe daquele objeto.

O segundo passo consiste na divisão da base original em três subconjuntos mutuamente exclusivos: treino, teste e validação (Conforme apresentado na Figura 1). A instância que for designada para um conjunto não aparece nos outros.O conjunto de treino possui 50% do tamanho do arquivo original. Já as bases de validação e teste, tem 25% da dimensão, seguindo a estratégia de avaliação Hold-out.

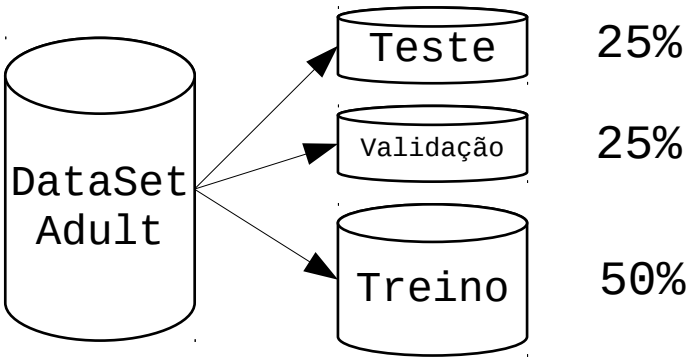

Figura 1: Divisão estratificada do conjunto de entrada.

No momento de separar a base original nos três conjuntos (treino, teste e validação), manteve-se as proporções originais das classes. Por exemplo, se um conjunto possui 200 instâncias da classe 1 e 100 da classe 2, o conjunto de treino terá 100 instâncias da classe 1 e 50 da classe 2.

As definições dos votos do classificador se deu por escolha do Tomé, o voto majoritário e voto ponderado por meio do método Inverso da distância Euclidiana.

A escolha das instâncias que formarão cada um dos conjuntos é totalmente aleatória. Assim, em cada execução do experimento, os conjuntos formados são diferentes. Depois de formados os conjuntos, o passo seguinte foi o treinamento do modelo de classificação. Como o KNN possui treinamento preguiçoso, esta etapa consiste apenas em armazenar o conjunto de teste. Para se determinar quais os melhores parâmetros dos métodos de classificação, adotou-se o conjunto de validação.

4. AVALIAÇÃO DOS EXPERIMENTOS

Depois de implementados os métodos de entrada, treinamento e escolha do melhor k, este sendo igual a 7 (k=7), foram realizados experimentos para avaliar alguns fatores dentro do classificador. Nestes experimentos foi comparado o desempenho do método ao se usar o Voto Majoritário Simples e o Voto Ponderado. Para o voto ponderado, foi implementada a estratégia do inverso da distância euclidiana.

Os resultados obtidos após as dez execuções são apresentados na Tabela 1 a seguir. Cada coluna apresenta a acurácia de uma das abordagens ao longo das 10 execuções. A última linha contém a Acurácia Média da estratégia e o seu devido Desvio Padrão.

5. ANÁLISE RESULTADOS

Observando-se os valores apresentados na Tabela 1 nota-se que odesempenho das abordagens implementadas é bastante similar. Todavia, é possível perceber que a abordagem do voto majoritário obteve a maior taxa de acerto, porém vale ressaltar que este resultado é situacional e em teste anteriores, houve casos em que o voto ponderado se sobressaiu em relação ao voto majoritário.

Repetição	Voto Majoritário	Voto Ponderado
0	83,23699	82,65896
1	83,23699	82,65896
2	87,86127	87,86127
3	84,97110	83,23699
4	84,39306	84,39306
5	87,86127	86,70520
6	85,54913	84,97110
7	87,28324	86,12717
8	84,97110	84,97110
9	90,75145	90,17341
10	85,54913	83,81503
11	87,28324	86,70520
12	85,54913	85,54913
13	86,12717	86,12717
14	83,23699	83,23699
15	86,12717	84,97110
16	85,54913	84,97110
17	84,39306	82,65896
18	88,43931	87,28324
19	88,43931	87,28324
20	86,12717	84,97110
21	87,28324	87,28324
22	80,92486	80,34682
23	85,54913	84,97110
24	87,28324	87,28324
25	84,97110	84,39306
26	91,32948	89,01734
27	79,76879	79,19075
28	84,97110	83,81503
29	83,23699	82,08092
30	82,65896	83,23699
31	84,39306	83,81503
Média	**85,6033**	**84,8988**
DP	**2,4802**	**2,3754**

Tabela 1: Resultados de ambas estratégias