Predição de Renda Anual

Leandro Luciani Tavares, Luiz Benedito Aidar Gavioli, Victor Narcizo de Oliveira Neto
Departamento de Computação (DComp)
Universidade Federal de São Carlos (UFSCar)
18052-780, Sorocaba, São Paulo, Brasil
leandro.ltavares@gmail.com, luizbag@gmail.com, vnarcizo@gmail.com

***Resumo*—Resumo, deixar para o final.**

***Keywords*-component; formatting; style; styling;**

I. Introdução

O projeto consiste em comparar o desempenho de algoritmos de classificação estudados em aula, entre eles o KNN (K - nearest neighbours), Regressão logística, Redes Neurais Artificiais (RNA) e Máquina de vetores de suporte (SVM). A base utilizada para comparação é a base disponibilizada em http://archive.ics.uci.edu/ml/datasets/Adult, a qual possui 48842 amostras. A comparação se baseia na classificação de renda dos cidadãos norte-americanos, em 2 classes, sendo elas renda menor ou renda maior ou igual à 50 mil dólares anuais, com base em 14 atributos.

II. Base de dados

A base de dados fornecida estava inicialmente separada em 2 arquivos, adult_test e adult_data, aos quais adicionou-se uma linha de cabeçalho para importação no Matlab, unificou-se ambos arquivos para o pré-processamento. A base de dados é composta por 14 atributos e 1 atributo-alvo, que representa se a renda é inferior a 50 mil dólares anuais ou igual ou superior a 50 mil dólares anuais, sendo eles:

Age
: Atributo contínuo que representa idade;

Workclass
: Atributo categórico que representa uma das 9 classes de trabalho;

Fnlwgt
: Atributo contínuo;

Education
: Atributo categórico que representa um dos 16 graus de escolaridade;

Education-num
: Atributo continuo relacionado ao grau de escolaridade;

Marital-status
: Atributo categórico que representa um dos 7 estados civis;

Occupation
: Atributo categórico que representa uma das 14 áreas de trabalho;

Relationship
: Atributo categórico que representa um dos 6 parentescos;

Race
: Atributo categórico que representa uma das 5 etnias;

Sex
: Atributo categórico que representa um dos 2 sexos possíveis;

Capital-gain
: Atributo contínuo que representa o ganho de capital;

Capital-loss
: Atributo contínuo que representa a perda de capital;

Hours-per-week
: Atributo continuo que representa as horas trabalhadas por semana;

Native-country
: Atributo categórico que representa um das 41 nacionalidades.

Após o carregamento removeu-se as amostras duplicadas, resultado em um total de 48813 amostras únicas, removeu-se também amostras com atributos idênticos porém com atributo-alvo distinto, resultando em 48785 amostras.

Apresentaram-se 3615 amostras com informações ausentes para os atributos: **work-class, occupation, native-country**. Essas amostras representavam cerca de 8% do total, portanto, optou-se por removê-las da base dados. Resultando em 45170 amostras.

Os atributos contínuos não sofreram modificações para os métodos do KNN, Regressão Logística, Redes Neurais e SVM, entretanto para o método Naive Bayes discretizou-se os valores em 10 cestas e aplicou-se a suavização de Laplace a fim de tratar cestas que não contenham valores.

Os atributos categóricos foram convertidos em colunas, sendo que cada coluna representa um dos valores possíveis para o atributo original e, o valor de cada uma das colunas passa a ser binário, indicando se a categoria do atributo original é a representada pela coluna.

Devido a expansão dos atributos categóricos, 3 colunas representavam atributos ausentes para 3 atributos originais, devido a remoção das amostras com atributos ausentes,

tornou-se irrelevante manter essas colunas, portanto, as mesma foram removidas, após essas transformações os 14 atributos originais tornaram-se 105. Para o método Naive-Bayes os 14 atributos originais tornaram-se 160.

O atributo-alvo foi convertido em um atributo binário, 1 para representar a classe positiva (renda igual ou superior a 50 mil dolares anuais) e, 0 para representar a classe negativa (renda inferior a 50 mil dolares anuais). 11197 amostras (24,79%) representam a classe positiva e 33973 amostras representam a classe negativa (75,21%).

Implementou-se 2 tipos de normalização para todos atributos exceto o atributo-alvo:

Normalização por reescala
: Restringe o intervalo de valores entre 0 e 1 para um atributo, mais sensível a outliers;

Normalização por padronização
: Garante que os valores tenham média igual a 0 e desvio-padrão igual a 1;

III. Metodologia experimental

Particionou-se a base de dados utilizando-se a metodologia de validação cruzada *(k-fold cross-validation)*, visto que os dados não são sensíveis ao tempo. Utilizou-se 10 partições, sendo 9 delas para o treinamento e 1 para a validação, dessa forma os conjuntos de treinamento contém 40653 amostras e os conjuntos de teste 4517 escolhidas aleatóriamente.

Para avaliação do poder de classificação de cada método aplicou-se as medidas mais utilizadas, como acurácia, F-medida, precisão e revocação.

Apresenta-se aqui os parâmetros selecionados, a fim de possibilitar a reprodução dos resultados obtidos em cada método:

Para normalização obteve-se resultados ligeiramente superiores (cerca de 1%) utilizando-se normalização por padronização, portanto, esta foi a opção utlizada em todos os testes.

A. KNN

O KNN *(K-vizinhos mais próximos)* é um método baseado em distâncias que consiste em selecionar os K vizinhos do conjunto de treinamento menos distante da amostra de teste, e por distante entende-se, que apresente a menor diferença entre os atributos.

O único parâmetro do KNN é o valor K, para este trabalho selecionou-se um valor 51, obtendo-se os resultados apresentados na Tabela I:

Tabela I
Resultados para o KNN sendo K = 51

Partição	Acurácia	F-medida	Precisão	Revocação
1	0.74607	0.74607	0.74607	0.74607
2	0.75515	0.75515	0.75515	0.75515
3	0.74925	0.74925	0.74925	0.74925
4	0.75603	0.75603	0.75603	0.75603
5	0.76179	0.76179	0.76179	0.76179
6	0.74662	0.74662	0.74662	0.74662
7	0.75293	0.75293	0.75293	0.75293
8	0.75183	0.75183	0.75183	0.75183
9	0.75803	0.75803	0.75803	0.75803
10	0.74341	0.74341	0.74341	0.74341
Média	0.75211	0.75211	0.75211	0.75211

B. Regressão logística

O método da regressão logística consiste em encontrar uma função *(hipótese)* que classifique os atributos, minimizando o erro entre as amostras, através do ajuste dos coeficientes do polinômio

Implementou-se 3 variações das hipóteses:

Hipótese Linear
: Atributos elevados a primeira potência;

Hipótese Quadrática
: Atributos elevados a primeira e segunda potência;

Hipótese Cúbica
: Atributos elevados a primeira, segunda e terceira potência;

A regressão logística ainda pode utilizar um parâmetro de regularização a fim de evitar os super ajustamento ao conjunto de treinamento, balançeando a complexidade da hipótese.

Para este trabalho selecionou-se a hipótese linear com um fator de regularizaçao lambda igual a 1, obtendo-se os resultados apresentados na Tabela II

Tabela II
Resultados para a Regressão logística sendo a hipótese linear e lambda = 1

Partição	Acurácia	F-medida	Precisão	Revocação
1	0.85707	0.85605	0.85852	0.85503
2	0.8628	0.86187	0.8644	0.86086
3	0.85621	0.85517	0.85814	0.85411
4	0.85834	0.85739	0.8603	0.85639
5	0.86338	0.86256	0.86481	0.86167
6	0.8591	0.85808	0.86049	0.85705
7	0.86925	0.8682	0.8712	0.867
8	0.86938	0.86837	0.87098	0.86722
9	0.86246	0.86157	0.8641	0.86059
10	0.85825	0.85708	0.86031	0.85587
Média	0.86162	0.86063	0.86333	0.85958

C. Redes Neurais Artificiais

D. Máquinas de vetores de suporte

E. Naive Bayes

IV. Conclusão

The conclusion goes here. this is more of the conclusion

Agradecimentos

The authors would like to thank... more thanks here

REFERÊNCIAS

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.