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

Aprendizado de máquina é atualmente um dos principais campos da computação, sendo um sub-campo da Inteligência Artificial, o qual pretende dar habilidade às máquinas de obter conhecimento e se aperfeiçoarem em determinada tarefa, sendo ela muitas vezes pouco trivial ou até mesmo impossível para um humano realizar devido à complexidade ou ao volume dos dados.

Nesse projeto, essa tarefa consiste em comparar o desempenho dos principais algoritmos de classificação estudados na disciplina de Aprendizado de Máquina: o KNN (K-nearest neighbours), a Regressão logística, as Redes Neurais Artificiais (RNA), as Máquinas de Vetores de Suporte (SVM) e o Naive-Bayes na classificação de padrões de renda. [1]

A predição de padrões de renda, é uma necessidade crescente de instituições fincanceiras, como bancos, seguradoras, factories, casas de câmbio, cooperativas de crédito, entre outras. Equipadas com ferramentas e dados, as instituições, tem a possibilidade de fornecer serviços personalizados para seus clientes como, por exemplo, taxas diferenciadas para clientes baseados em suas rendas anuais. Ou adequar seu modelo de negócios a determinado tipo de consumidor sabendo-se que, por exemplo, este tem um perfil inadimplente. [2]

Em análises econômicas, prever e classificar padrões de renda é parte fundamental, dada a necessidade de estimar o desenvolvimento econômico de um país, e traçar perfis dos cidadãos, como por exemplo, qual setor da economia tem os melhores salários, qual idade tem a parcela da população que possuí maior renda anual. Além de auxiliar o planejamento econômico, controle de inflação, definição de taxas de juros. [3]

A comparação se baseia na classificação de renda dos cidadãos norte-americanos, em 2 classes, sendo elas: os que possuem renda menor ou renda maior ou igual à 50 mil dólares anuais, com base em 14 atributos. A base utilizada para comparação pode ser consultada em [4] e [5].

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 capi-

tal;

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 os 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.

Para se permitir a visualizção dos dados, implementou-se a Análise de componentes principais (PCA), para redução dos 105 atributos para 2, resultando na imagem disposta na Figura 1, na qual a classe positiva, renda igual ou superior a 50 mil dólares anuais, é representada por + e a classe negativa, renda inferior a 50 mil dólares anuais, é representada por o.

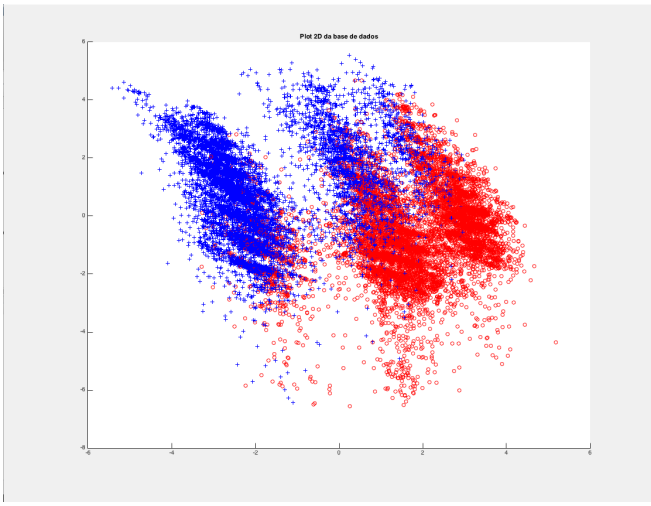

Figura 1. Plot 2D dos atributos

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, contabilizando também o tempo de treinamento e teste de cada partição.

A fim de verificar a possibilidade de superajustamento ou subajustamento, gerou-se também os gráficos das curvas de aprendizando, realizando os treinamento com partições incrementais, iniciando com 1 partição e finalizando com 9.

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

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 o qual testou-se com os valores: 1, 3, 5, 7, 11, 21, 51.

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 θ.

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 seleção dos parâmetros testou-se, através de busca em grid, as 3 hipóteses, com parâmetro λ = 0, ou seja, sem regularização, e com a regularização variando de 10^0 a 10^3

C. Redes Neurais Artificiais

As Redes Neurais Artificias, utilizadas foram os Perceptrons Multi-camadas, que consistituem uma série de camadas massivamente conectadas de regressores logísticos, portanto, o método consiste em ajustar matrizes de coeficientes θ a fim de minimizar o erro de classificação das amostras.

Entre os parâmetros a serem ajustados, existe a taxa de aprendizagem α, o número de camadas o número de neurônios de cada camada.

D. SVM - Máquinas de vetores de suporte

O SVM foi implementado utilizando-se a biblioteca LIBSVM [6]

Os parâmetros incluem a seleção do kernel, dos coeficientes *C*, que representa o parâmetro de custo (para os kernel linear, radial e polinomial) e γ (para os kernel radial e polinomial).

Testou-se o SVM com kernel linear, com *C* com valores de 10^{-4} a 10^2, com passo incremental 1 na potência. Para o kernel radial e polinomial, testou-se através de busca em grid, com *C* variando de 10^{-4} a 10^2, e γ variando de 10^{-2} a 10^2 ambos com passo incremental 1 na potência.

E. Naive Bayes

O método Naive-Bayes se baseia nas probabilidades de ocorrência de cada classe, e de cada atributo individualmente sabendo a classe em que o mesmo se encontra. O métodos Naive-Bayes se baseia apenas nas probabilidades, portanto não possuí parametros a serem ajustados.

IV. Resultados

Para a 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 único parâmetro do KNN é o valor K, para os resultados selecionou-se o valor K = 51, obtendo-se os resultados apresentados na Tabela I:

Tabela I
RESULTADOS PARA O KNN SENDO K = 51

Partição	Acurácia	F-medida	Precisão	Revocação	Tempo
1	0.74607	0.74607	0.74607	0.74607	37.549
2	0.75515	0.75515	0.75515	0.75515	37.233
3	0.74925	0.74925	0.74925	0.74925	37.347
4	0.75603	0.75603	0.75603	0.75603	38.548
5	0.76179	0.76179	0.76179	0.76179	39.026
6	0.74662	0.74662	0.74662	0.74662	38.081
7	0.75293	0.75293	0.75293	0.75293	37.536
8	0.75183	0.75183	0.75183	0.75183	36.817
9	0.75803	0.75803	0.75803	0.75803	36.604
10	0.74341	0.74341	0.74341	0.74341	36.824
Média	0.75211	0.75211	0.75211	0.75211	37.557

B. Regressão logística

Visando melhor desempenho de tempo, selecionou-se a hipótese linear com um fator de regularização λ = 1, obtendo-se os resultados apresentados na Tabela II.

Tabela II
RESULTADOS PARA A REGRESSÃO LOGÍSTICA SENDO A HIPÓTESE LINEAR E λ = 1

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

C. Redes Neurais Artificiais

D. Máquinas de vetores de suporte

Tabela III
RESULTADOS PARA SVM

Partição	Acurácia	F-medida	Precisão	Revocação	Tempo
1	0.84725	0.84635	0.84835	0.84553	387.15
2	0.85753	0.85665	0.85894	0.85575	371.86
3	0.85985	0.85886	0.86134	0.85784	398.54
4	0.85467	0.85379	0.85627	0.85291	396.04
5	0.86017	0.85941	0.86129	0.85862	370.89
6	0.85363	0.85271	0.85463	0.85186	371.2
7	0.86295	0.86198	0.86457	0.86094	371.87
8	0.86137	0.86044	0.86265	0.85948	373.39
9	0.85854	0.8577	0.85991	0.85684	371.56
10	0.85725	0.85617	0.8588	0.85508	371.22
Média	0.85732	0.85641	0.85867	0.85548	378.37

E. Naive Bayes

V. Conclusão

The conclusion goes here. this is more of the conclusion

AGRADECIMENTOS

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

REFERÊNCIAS

[1] T. A. Almeida. (2015) Predição de renda anual. [Online]. Available: http://www.moodle.ufscar.br/file.php/4639/projeto/grupos/P1/P1.pdf

[2] P. Chetty. (2011) Importance of prediction of income of customers to banks. [Online]. Available: http://www.projectguru.in/publications/importance-of-prediction-of-income-of-customers-to-banks/

[3] A. Dahir. (2014) The importance of estimating the national income of a country and the difficulties economist encounter while carrying such estimation especially in the developing countries. [Online]. Available: http://pt.slideshare.net/delmujahid/question-2-40726032

[4] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[5] R. Kohavi and B. Becker. (1996) Adult data set. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Adult

[6] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] D. F. Gordon and M. Desjardins, "Evaluation and selection of biases in machine learning," 1995.