Avaliação de Técnicas de Classificação*

Tiago C A Amorim (RA: 100675)[a], Taylon L C Martins (RA: 177379)[b]

[a] *Doutorando no Departamento de Engenharia de Petróleo da Faculdade de Engenharia Mecânica, UNICAMP, Campinas, SP, Brasil*
[b] *Aluno especial, UNICAMP, Campinas, SP, Brasil*

Keywords: Classificação, Regressão Logística, k-Vizinhos mais Próximos, Validação Cruzada

1. Introdução

Este relatório apresenta as principais atividades realizadas no desenvolvimento das atividades propostas na Lista 02 da disciplina IA048: Aprendizado de Máquina, primeiro semestre de 2024. O foco deste exercício é de construir a avaliar o desempenho de algoritmos de classificação usando duas versões de um mesmo estudo.

2. Tarefa Proposta

Nesta atividade, vamos abordar o problema de reconhecimento de atividades humanas (HAR, do inglês *human activity recognition*) a partir de informações capturadas por sensores de smartphones. Em particular, vamos trabalhar com a base de dados UCI HAR [1], que contém registros de sensores inerciais presentes em um smartphone preso à cintura de 30 sujeitos realizando atividades cotidianas. Cada pessoa realizou seis atividades, as quais correspondem aos seguintes rótulos:

Atividade		**Rótulo**
Caminhar	(*Walking*)	1
Subir escadas	(*W. upstairs*)	2
Descer escadas	(*W. downstairs*)	3
Sentado	(*Sitting*)	4
Em pé	(*Standing*)	5
Deitado	(*Laying*)	6

Tabela 1: Rótulos[1].

[1] Foram adicionados os termos originais entre parênteses porque todos os gráficos foram construídos com os termos em inglês

Foram capturadas as amostras dos três eixos (x, y e z) do acelerômetro (ACC, do inglês *accelerometer*) e do giroscópio (GYR, do inglês *gyroscope*) presentes no smartphone, empregando uma taxa de amostragem de 50 Hz. O conjunto completo de amostras foi particionado aleatoriamente em treinamento (70% dos voluntários) e teste (30% dos voluntários).

*Relatório número 02 como parte dos requisitos da disciplina IA048: Aprendizado de Máquina.

2.1. Primeira parte

Primeiramente, será explorada uma versão do conjunto de dados na qual já houve pré-processamento e extração de características. No caso, cada amostra contém 561 atributos derivados de uma mesma janela de 2,56 s dos 6 sinais disponíveis (ACC: x,y,z; GYR: x,y,z), considerando suas representações tanto no domínio do tempo quanto no domínio da frequência.

(a) Construa uma solução para este problema baseada no modelo de regressão logística. Descreva a abordagem escolhida para resolvê-lo (softmax, classificadores binários combinados em um esquema um-contra-um ou um-contra-todos). Obtenha, então, a matriz de confusão para o classificador considerando os dados do conjunto de teste. Além disso, adote uma métrica global para a avaliação do desempenho (médio) deste classificador. Discuta os resultados obtidos.

(b) Considere, agora, a técnica k-nearest neighbors (kNN). Adotando um esquema de validação cruzada, mostre como o desempenho do classificador, computado com a mesma métrica adotada no item (a) varia em função do parâmetro k. Escolhendo, então, o melhor valor para k, apresente a matriz de confusão para os dados de teste e o desempenho medido nesse conjunto. Comente os resultados obtidos, inclusive estabelecendo uma comparação com o desempenho da regressão logística.

2.2. Segunda parte

Agora, vamos utilizar os dados "brutos" combinados de ACC e GYR como entradas dos classificadores. Para isso, devemos recorrer aos registros disponibilizados no diretório 'Inertial Signals', os quais estão separados por eixo e por sensor, sendo que cada amostra individual agora é formada por 128 valores (atributos), que correspondem às amplitudes instantâneas de aceleração (ACC) ou velocidade angular (GYR) dentro de uma janela de 2,56 s.

(a) Monte, então, a nova matriz de entrada concatenando os seis sinais temporais e, então, repita o procedimento experimental detalhado nos itens (a) e (b). Ao final, com base no desempenho obtido, teça uma análise comparativa entre a abordagem do item anterior e

a abordagem baseada nos sinais "brutos" empregada nesta segunda parte.

3. Aplicação

Toda a avaliação foi feita em um único *notebook* Jupyter, em Python. Foi feito o uso da biblioteca *Scikit-learn* [2] para fazer as diferentes manipulações nos dados. O código pode ser encontrado em https://github.com/TiagoCAAmorim/machine_learning.

3.1. Conjuntos de Dados

Os dados foram disponibilizados em formato tabular, já com uma separação entre os dados de treino e de teste (tabela 3). Os dados pré-processados são formados por 561 atributos, enquanto que os dados brutos são formados por 768 atributos[2]. Uma descrição de cada um dos atributos é feita pelos autores no pacote do conjunto de dados [1].

Conjunto	Número de Amostras
Treino	7 352 (71.4%)
Teste	2 947 (28.6%)
Total	10 299 (100%)

Tabela 2: Tamanho da base de dados.

Aparentemente o desbalanço entre as classes não é significativo, mas existe (figura 1). A menor classe tem cerca de 30% menos amostras que a maior classe. De toda forma será utilizada a acurácia balanceada como métrica da qualidade do classificador (média dos *recalls* de cada classe).

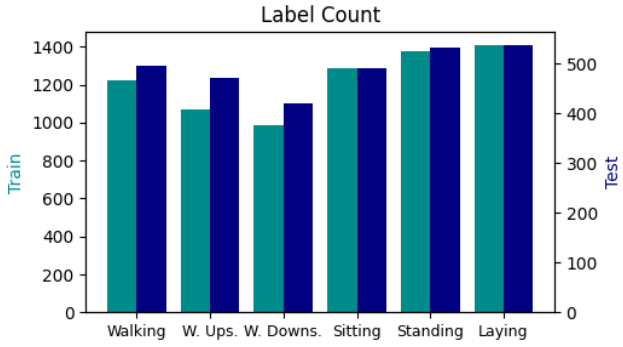

Figura 1: Número de amostras por classe.

3.2. Dados Pré-processados

Os dados pré-processados estão normalizados no intervalo $[-1; 1]$, à exceção de alguns dos atributos (figura 2). Desta forma, em um primeiro momento não existe necessidade de normalizar os dados.

[2]Neste estudo foram ignorados os dados de aceleração total, que são 384 atributos adicionais.

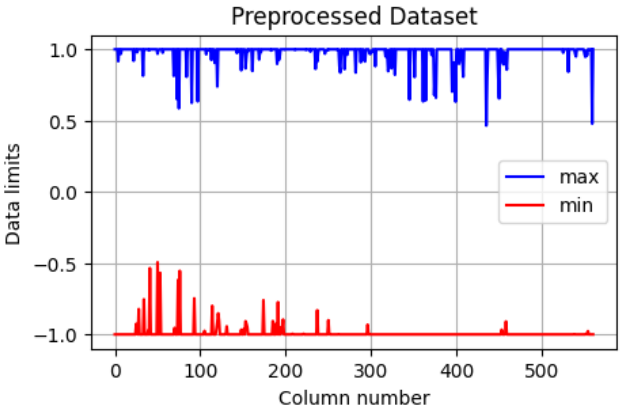

Figura 2: Limites dos atributos do conjunto de treinamento dos dados pré-processados.

3.2.1. Regressão Logística

O classificador de regressão logística foi construído com a classe **LogisticRegressionCV** do *Scikit-Learn*. Esta classe realiza a otimização do parâmetro de regularização junto com validação cruzada. Foram utilizadas as seguintes opções para o ajuste deste modelo:

1. Validação cruzada estratificada em 5 pastas.
2. Normalização do tipo l_2 ($\frac{1}{2}||w||_2^2$), com otimização do seu inverso ($c = \frac{1}{l_2}$)
3. Função objetivo da otimização: acurácia balanceada.
4. Estratégia: multinomial (entropia cruzada[3]).

O modelo ajustado tem $l_2 = 0.3594$. A acurácia balanceada na validação cruzada foi de 0.9932, e com os dados de teste 0.9598.

Classe	Recall	F1 score
Todas	0.9598	0.9606
Caminhar	0.9940	0.9686
Subir escadas	0.9427	0.9569
Descer escadas	0.9690	0.9795
Sentado	0.8717	0.9214
Em pé	0.9812	0.9372
Deitado	1.0000	1.0000

Tabela 3: Resultados do classificador com regressão logística.

[3]Equivalente a utilizar *Softmax*.

Referências

[1] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, et al., A public domain dataset for human activity recognition using smartphones., in: The European Symposium on Artificial Neural Networks, Vol. 3, 2013, p. 3.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.