06/05/2022

Primeiro Trabalho de Inteligência Artificial e Sistemas Inteligentes

Prof. Flávio Miguel Varejão

1. ***Descrição***

Este trabalho consiste em realizar uma comparação experimental entre um conjunto pré-definido de técnicas de aprendizado e classificação automática aplicadas a um problema de classificação. As técnicas escolhidas são: ZeroR (ZR), Naive Bayes (NB), Decision Tree (DT), K Vizinhos Mais Próximo (KNN), Multi-layer Perceptron (MLP), Random Forest (RF) e Heterogeneous Pooling (HP). O procedimento experimental será dividido em duas etapas.

A primeira etapa consiste no treino e teste com 3 rodadas de validação cruzada estratificada de 10 folds do classificador que não possui hiperparâmetros, isto é, os classificadores ZR e NB .

A segunda etapa consiste no treino, validação e teste dos classificadores que precisam de ajuste de hiperparâmetros, isto é, os classificadores DT, KNN, MLP, RF e HP. Neste caso o procedimento de treinamento, validação e teste será realizado através de 3 rodadas de ciclos aninhados de validação e teste, com o ciclo interno de validação contendo 4 folds e o externo de teste com 10 folds. A busca em grade (grid search) do ciclo interno deve considerar os seguintes valores de hiperparâmetros de cada técnica de aprendizado:

Decision Tree: *{'criterion': ['gini','entropy'],*
'max_depth': range(1,10)}

K Vizinhos Mais Próximo: {‘n_neighbors’:[1,3,5,7,9]}

Support Vector Machine: {'hidden_layer_sizes': [(100,),(10,)],
'activation': ['relu'],
'solver': ['adam'],
'alpha': [0.0001, 0.05],
'learning_rate': ['constant','adaptive']}

Random Forest: {'n_estimators': [50, 100, 150],
'max_depth': [5, 10, 15, None],
'max_features': ['sqrt', 'log2', None]}

HeterogeneousPooling: {‘n_samples’: [1, 3, 5, 7]}

Os resultados de cada classificador devem ser apresentados numa tabela contendo a média das acurácias obtidas em cada fold, o desvio padrão e o intervalo de confiança a 95% de significância dos resultados, e também através do boxplot dos resultados de cada classificador em cada fold.

Um exemplo de uma tabela para uma base de dados hipotética é mostrado a seguir.

Método	Média	Desvio Padrão	Limite Inferior	Limite Superior
ZR	0.95	0.04	0.94	0.96
NB	0.91	0.04	0.87	0.95
DT	0.95	0.02	0.94	0.96
KNN	0.90	0.05	0.89	0.95
MLP	0.96	0.07	0.88	0.99
RF	0.97	0.02	0.95	0.98
HP	0.93	0.01	0.92	0.95

O método HP deve ser implementado. Os métodos ZR, NB, DT, KNN, MLP e RF estão disponíveis no scikit-learn.

Além das tabelas e dos gráficos bloxplot, será necessário apresentar também a tabela pareada dos resultados (p-values) dos testes de hipótese entre os pares de métodos. Na matriz triangular superior devem ser apresentados os resultados do **teste t pareado corrigido** (amostras dependentes) e na matriz triangular inferior devem ser apresentados os resultados do **teste não paramétrico de wilcoxon**. Os valores da célula da tabela que rejeitarem a hipótese nula para um nível de significância de 95% devem ser escritos em negrito.

Um exemplo de uma tabela pareada para uma base de dados hipotética é mostrado a seguir.

ZeroR	0.085	**0.045**	0.065	0.089
0.045	BA	0.105	0.105	0.076
0.096	**0.036**	AB	0.085	0.096
0.105	0.105	0.096	RF	0.105
0.024	0.094	0.105	0.084	HP

2. **HP**

O classificador Heterogeneous Pooling é um combinado de classificadores heterogêneos que usa como classificadores base: Árvore de Decisão (DT), Naive Bayes

Gaussiano (NB) e K Vizinhos Mais Proximo (KNN), sempre com valores default do sklearn para seus hiperparâmetros. O único parâmetro do método Heterogeneous Pooling é o *n_samples*, que indica o número de vezes que os classificadores base serão usados para gerar o combinado. Por exemplo, se *n_samples* é igual a 3, o combinado será composto por 9 classificadores: 3 árvores de decisão, 3 naive bayes e 3 vizinhos mais próximos. Para diferenciar os classificadores de mesmo tipo em um combinado, o primeiro deles será treinado com a base de treino original e os demais serão treinados com uma base de treino diferente, obtida a partir da base de treino original através de um método para seleção de características. O método consiste em selecionar aleatoriamente um número de características, variando de 2 até o (número de características - 1) e, a partir desse número, devem ser selecionadas as N características através de uma roleta onde a chance de uma característica ser escolhida é baseada no seu ANOVA F-value. **Atenção, uma característica não deve ser escolhida mais de uma vez.** Assim, toda vez que uma característica é selecionada a roleta deve redistribuir as chances das demais características.

As descrições dos métodos KNN, NB e DT usados no HP e implementados no sklearn podem ser acessadas respectivamente em:
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

O critério de decisão para classificar uma instância é a votação majoritária, ou seja, deve-se escolher a classe mais escolhida dentre os classificadores que compõem o combinado. Em caso de empate, a classe escolhida deve ser a mais frequente na base de dados de treino original dentre as que empataram na votação.

O pseudo código a seguir mostra como o HP é obtido a partir de uma base de dados de treino:

- Obter e armazenar a ordenação das classes de acordo com a ocorrência nos exemplos na base de treino (ordenar decrescentemente da mais frequente para a menos frequente)
- Para cada um dos *n_samples* faça
    - Se for a primeira iteração então
        - Usar a base original para treino dos classificadores
    - Senão
        - Escolher o número de características aleatoriamente (2 a *n_caracteristicas*-1)
        - Montar a base selecionando as características utilizando o método da roleta.
    - Fim-se
    - Treinar os classificadores KNN, NB, DT na base de treino corrente e incluí-los no combinado de classificadores
- Fim-para

O pseudo código seguinte mostra como o combinado HP é usado para classificar um exemplo da base de dados de teste:

- Para cada um dos classificadores individuais do combinado faça
    - Obter a classificação do exemplo usando o classificador individual e armazenar a classe selecionada
- Fim-para
- Contar quantas vezes cada classe foi selecionada e obter a(s) mais votada(s)
- Se mais de uma classe for a mais votada então
    - Retornar a classe mais votada mais frequente na base de treino dentre as que empataram
- Senão
    - Retornar a classe mais votada
- Fim-se

---

3. Base de Dados

A base de dados usada no trabalho foi obtida de um projeto de pesquisa que visa utilizar informações de pacientes para classificação de leucoplasia oral e carcinoma. A base completa contém 237 exemplos.

O conjunto de dados contém imagens histopatológicas de carcinoma de células escamosas oral e leucoplasia (representadas por amostras com e sem displasia epitelial). Também contém dados sociodemográficos (gênero, idade e cor da pele), bem como dados clínicos (uso de tabaco, consumo de álcool, exposição ao sol, lesão fundamental, tipo de biópsia, cor da lesão, superfície da lesão e diagnóstico da lesão). Os dados foram coletados entre 2010 e 2021 em pacientes tratados no projeto de Diagnóstico Oral (NDB) da Universidade Federal do Espírito Santo (UFES), Brasil.

Existem três tarefas distintas que foram delineadas com base nos dados coletados no projeto de Diagnóstico Oral (NDB) da Universidade Federal do Espírito Santo (UFES). Cada uma dessas tarefas visa abordar diferentes aspectos da classificação e análise das amostras histopatológicas e dados demográficos/clínicos associados:

- Tarefa I - Leucoplasia (L)/OSCC (NDB-UFES): Esta tarefa tem como objetivo a classificação das amostras em Leucoplasia ou Carcinoma de Células Escamosas Bucais (CCEB), **utilizando apenas os dados demográficos e clínicos**.
- Tarefa II - Ausência (S/D)/Presença de displasia (C/D) (NDB-UFES): Nesta tarefa, o foco está na distinção entre amostras que apresentam ou não displasia, **utilizando apenas os dados demográficos e clínicos.**
- Tarefa III - OSCC/Leucoplasia com displasia (LC/D)/Leucoplasia sem displasia (LC/S) (NDB-UFES): Nesta tarefa, o objetivo é classificar as amostras em Carcinoma de Células Escamosas Bucais (OSCC), Leucoplasia com displasia ou Leucoplasia sem displasia, **utilizando apenas os dados demográficos e**

clínicos.

Os alunos serão designados para implementar uma das três tarefas com base no último dígito de sua matrícula. Alunos cujos números de matrícula terminam de 0 a 3 serão responsáveis pela Tarefa I, aqueles com números finais de 4 a 6 pela Tarefa II, e os alunos com números finais de 7 a 9 pela Tarefa III.

Cada tarefa possui classes diferentes, as Tabelas 1, 2 e 3, apresentam respectivamente, a distribuição de classe de cada tarefa.

Tabela 1 – Distribuição das exemplos por classe da Tarefa I

Classe	Quantidade	(%)
Leukoplakia	146	61,6%
OSCC	91	38,4%
Total	237	100,0%

Tabela 2 – Distribuição das exemplos por classe da Tarefa II

Classe	Quantidade	(%)
Presence	180	75,9%
Absence	57	24,1%
Total	237	100,0%

Tabela 3 – Distribuição das exemplos por classe da Tarefa III

Classe	Quantidade	(%)
OSCC	91	38,4%
Leukoplakia with dysplasia	89	37,5%
Leukoplakia without dysplasia	57	24,1%
Total	237	100,0%

Um notebook contendo as informações de como utilizar os dados está disponível e em anexo a esta atividade. **Faz parte do trabalho fazer o pré-processamento dos dados, como limpeza dos dados, separação das características e codificação de variáveis categóricas.**

4. **Informações Complementares**

a. Use o valor 36851234 para o parâmetro random_state (`random_state=36851234`) nas chamadas a RepeatedStratifiedKFold para que os resultados sejam reproduzíveis.

b. Use o valor 11 para o parâmetro random_state (`random_state=11`) na inicialização de todos os classificadores que tenham não determinismo (isto é, que tenham um parâmetro random_state) para que chamadas sucessivas não retornem valor diferente e, portanto, tornar os resultados reproduzíveis.

c. Para que os resultados do método HeterogeneousPooling sejam reproduzíveis use o valor 0 para o parâmetro random_state (`random_state=0`) na primeira chamada a resample. A

partir daí, use o valor corrente incrementado de 1 nas chamadas sucessivas de resample (random_state = 1, random_state = 2, ...).

d. Use as seguintes funções do scipy.stats para obter os resultados dos testes de hipóteses:

from scipy import stats
stat, p = stats.ttest_rel(scores1, scores2)
...
stat, p = stats.wilcoxon(scores1, scores2)

d. Os gráficos bloxplot requeridos no treino e no teste devem ser gerados usando função específica do pacote seaborn.

e. O apêndice deste enunciado apresenta instruções de instalação e uso do overleaf para a escrita do artigo.

5. ***Artigo***

Após a realização dos experimentos, um artigo descrevendo todo o processo experimental realizado deverá ser escrito em latex usando o software overleaf. O artigo deve ter um máximo de 5 páginas e ser estruturado da seguinte forma:

1. Título
2. Resumo
3. Seção 1. Introdução
4. Seção 2. Base de Dados
    a. Descrição do Domínio
    b. Definição das Classes e das Características
    c. Número de Instâncias
5. Seção 3. O Método Heterogeneous Pooling
6. Seção 4. Descrição dos Experimentos Realizados e seus Resultados
7. Seção 5. Conclusões
    a. Análise geral dos resultados
    b. Contribuições do Trabalho
    c. Melhorias e trabalhos futuros
8. Referências Bibliográficas

Na subseção de análise geral dos resultados é importante discutir, dentre outras coisas, se houve diferença estatística significativa entre quais métodos e responder se teve um método que foi superior.

6. Condições de Entrega

O trabalho deve ser feito individualmente e submetido pelo sistema da sala virtual até a data limite (10 de junho de 2024).

O trabalho deve ser submetido em dois arquivos: um arquivo pdf com o artigo produzido no trabalho e um arquivo ipynb com o notebook jupyter com o código do trabalho. Tanto o arquivo pdf quanto o arquivo ipynb devem possuir o mesmo nome Trab1_Nome_Sobrenome.

Note que a data limite já leva em conta um dia adicional de tolerância para o caso de problemas de submissão via rede. Isso significa que o aluno deve submeter seu trabalho até no máximo um dia antes da data limite. Se o aluno resolver submeter o trabalho na data limite, estará fazendo isso assumindo o risco do trabalho ser cadastrado no sistema após o prazo. Em caso de recebimento do trabalho após a data limite, o trabalho não será avaliado e a nota será ZERO. Plágio ou cópia de trabalhos serão verificadas. Trabalhos em que se configure cópia receberão nota zero independente de quem fez ou quem copiou.

7. Requisitos da implementação

a. Modularize seu código adequadamente.

b. Crie códigos claros e organizados. Utilize um estilo de programação consistente, Comente seu código.

c. Os arquivos do programa devem ser lidos e gerados na mesma pasta onde se encontram os arquivos fonte do seu programa.

Observação importante

Caso haja algum erro neste documento, serão publicadas novas versões e divulgadas erratas em sala de aula. É responsabilidade do aluno manter-se informado, freqüentando as aulas ou acompanhando as novidades na página da disciplina na Internet.

Apêndice A. Boxplots usando seaborn

```
def example1():
      mydata=[1,2,3,4,5,6,12]
      sns.boxplot(y=mydata) # Also accepts numpy arrays
      plt.show()

def example2():
      df = sns.load_dataset('iris')
      #returns a DataFrame object. This dataset has 150 examples.
      #print(df)
      # Make boxplot for each group
      sns.boxplot( data=df.loc[:,:] )
      # loc[:,:] means all lines and all columns
      plt.show()

example1()
example2()
```

Apêndice B. Artigo em Latex usando Overleaf

Juntamente com este enunciado foi disponibilizado um arquivo zip com o template de latex para confecção do artigo. O primeiro passo a ser feito é criar uma conta pessoal no Overleaf (https://www.overleaf.com/register). Uma vez criada sua conta, deve-se entrar nela. Para incluir o template no overleaf, basta apenas selecionar "New Project>Upload Project" e selecionar o arquivo zip, como mostrado na figura abaixo. Não é necessário descompactar, faça o upload do zip direto. Lembrar de renomear o artigo após o upload do arquivo.