31/05/2022

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 Gaussiano (NBG), KMeans Centroides (KMC), K Vizinhos Mais Próximos (KNN) e Árvore de Decisão (AD). 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 dos classificadores que não possuem hiperparâmetros, isto é, os classificadores ZR e NBG.

A segunda etapa consiste no treino, validação e teste dos classificadores que precisam de ajuste de hiperparâmetros, isto é, os classificadores KMC, KNN e AD. 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:

KMC: [k = 1, 3, 5, 7]

KNN: [n_neighbors = 1, 3, 5, 7]

AD: [max_depth = None, 3, 5, 10]

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
NBG	0.91	0.04	0.87	0.95
KMC	0.95	0.02	0.94	0.96
KNN	0.96	0.07	0.88	0.99
AD	0.97	0.02	0.95	0.98

O método KMC deve ser implementado. Os métodos ZR, NBG, KNN e AD estão disponíveis no scikit-learn. As descrições dos métodos implementados no sklearn podem ser acessadas respectivamente em:

https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html
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.htmlr

Os dados utilizados no conjunto de treino em cada rodada de teste devem ser padronizados (normalização com z-score). Os valores de padronização obtidos nos dados de treino devem ser utilizados para padronizar os dados do respectivo conjunto de teste.

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 (amostras dependentes) e na matriz triangular inferior devem ser apresentados os resultado do teste não paramétrico de wilcoxon. Os valores da célula da tabela 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	NB	0.105	0.105	0.076
0.096	**0.036**	KM	0.085	0.096
0.105	0.105	0.096	KNN	0.105
0.024	0.094	0.105	0.084	AD

2. KMC

O classificador KMC utiliza um algoritmo de agrupamento para definir K grupos de exemplos de cada classe na base de treino. Assumindo que uma base de dados possui ncl classes, o algoritmo KMC forma inicialmente K*ncl grupos, sendo K grupos em cada uma das ncl classes. Em seguida, são calculados os centróides de cada um dos grupos e este centróide é associado a classe do grupo a partir do qual foi gerado. O método possui como hiperparâmetro o valor de K.

Para realizar uma classificação, o KMC verifica qual o centróide mais próximo do elemento a ser classificado e retorna a sua classe.

Para se criar o método KMC, o método Kmeans do sklearn (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html) deve ser utilizado com valores default para seus hiperparâmetros.

3. Base de Dados

A escolha da base utilizada neste trabalho é flexível. A base de dados pode ser de algum problema de interesse do estudante ou pode ser escolhida uma das seguintes bases:

digits, wine e breast cancer, todas três disponíveis em https://scikit-learn.org/stable/datasets/toy_dataset.html.

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. Os gráficos bloxplot requeridos no treino e no teste devem ser gerados usando função específica do pacote seaborn (ver instruções de instalação e uso no apêndice A deste enunciado).

c. O apêndice B 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 KMC
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 (27 de junho de 2022).

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 para ser carregado e executado no jupyter. 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.