Segunda Lista de Exercícios - Machine Learning 2018.2

Raissa Camelo Salhab

[1]Departamento de Computação – Universidade Federal Rural Pernambucana (UFRPE)
Bacharelado em Ciência da Computação

srtacamelo@gmail.com

Resumo. *Este documento se trata de uma apresentação formal da(s) atividade(s) solicitada(s) em sala de aula. O conteúdo contido abaixo refere-se a resolução da lista de questões n^a 2 da disciplina.*

1. Projeto

O exercício proposto se trata da aplicação dos algoritmos: Árvore de Decisão, Naïve Bayes,K-Vizinhos Mais Próximos (K-NN), K-Vizinhos Mais Próximos Ponderado (W-K-NN) e Máquina de Vetores de Suporte (SVM), utilizando como base de dados sets da UCI machine Learning.

2. Base de Dados

2.1. Iris

A base de dados da Iris consiste em um arquivo *txt* contendo 150 casos de flores do tipo Iris. Cada caso é representado por 4 atributos reais: largura de sépala, comprimento de sépala, largura de pétala e comprimento de pétala. Existem 3 classificações possíveis para cada flor: Setosa, Versicolour Virginica.

2.2. Statlog(Shuttle)

A base Statlog (Shuttle) contém 58000 casos em arquivo *txt*, cada um contendo 9 atributos numéricos (inteiros), de acordo com a descrição da UCI, porém nos arquivos existem 10 valores. O primeiro atributo consiste em uma medição de tempo e o ultimo refere as classes dos objetos. Não há informações claras sobre cada um dos atributos, além destes dois.
Classes: 1 - Rad Flow, 2 - Fpv Close, 3 - Fpv Open, 4- High, 5 - Bypass, 6- Bpv Close, 7- Bpv Open.

2.3. Yeast

Esta base integra informações em *txt* sobre aglomerados celulares e onde, nestes aglomerados, existe a maior probabilidade de localizar proteínas. O Yeast UCI dispõe de 1484 instancias de aglomerados, cada uma possuindo um total de 8 atributos numéricos + sequencial (valores reais): Número de sequencia, mcg, gvh, alm, mit, erl, pox, vac, nuc. As siglas referem-se a índices de presença de determinadas substancias e outros tipos de análises biológicas. As classes referem-se as posições nas quais é mais provável encontrar proteínas em cada instancia.

3. Pré-processamento

Para o pré-processamento dos dados foram aplicadas técnicas comuns em todas as 3 bases, uma vez que as 3 foram fornecidas em formato *txt*. Os arquivos foram lidos e convertidos para o padrão de planilhas *csv*, para então serem convertidos para o formato pandas.DataFrame, utilizando a biblioteca *Pandas*. Para os DataSets da Iris e da Yeast UCI foram discretizados os valores referentes as classes dos dados, transformando os valores nominais em números na base decimal. Para a base Shuttle todos os valores foram normalizados para uma faixa numérica entre 1 e 0, com exceção da coluna-classe, que já estava discretizada. As bases Yeast e Iris também tiveram seus atributos normalizados, afim de embutir os valores na mesma faixa (1 à 0). O objetivo da normalização foi padronizar os valores de atributos distintos, dado que alguns atributos possuíam valores em faixas mais elevadas que os outros.

4. Experimentos

Para os experimentos os dados normalizados foram lidos novamente, reordenados de maneira aleatória, afim de realizar o procedimento 10-folds, como solicitado. Para cada um dos algoritmos de classificação implementados o método 10-folds foi chamado 5 vezes, cada uma das vezes reordenando os dados novamente. Ao fim do procedimento foram obtidos 50 valores de acurácia para cada algoritmo, dos quais foram calculadas a média, mediana e o desvio padrão.

5. Resultados

A seguir os resultados obtidos para cada algoritmo, para cada base. Para a realização da analise dos dados os valores foram reduzidos para apenas 4 casas decimais, fazendo com que algumas medidas de classificadores distintos ficassem iguais. Afim de verificar se existiu alguma diferença estatisticamente significante entre os classificadores a hipótese de Gosset (t-studant) foi aplicada. Para fins de melhor visualização optamos por aplicar o teste de Gosset entre as três variações do KNN e do WKNN, isoladamente, para determinar se houve alguma melhora significativa com a mudança dos parâmetros. E também um teste individual entre os classificadores SVM-Linear e SVM-RBF. As variações cuja média de acurácia foram melhores foram escolhidas para serem comparadas com os demais algoritmos (Árvore de decisão e Naive Bayes), através da média de acurácia. A função t-studant da biblioteca scipy foi utilizada para a medição dos valores críticos e da probabilidade de hipótese.

5.1. Iris Database

O dataSet Iris foi um dos mais simples e rápidos de processar, por conter uma base de dados pequena e cada um de seus ”cases”possuírem apenas 4 atributos. A baixo encontra-se uma tabela com os resultados obtidos para cada classificador, a média, a mediana e o desvio padrão foram calculados utilizando os 50 valores de acurácia adquiridos durante os experimentos.

Figura 1. Resultados - Iris DataSet

Iris DataSet			
Classificador	Média	Desvio Padrão	Mediana
Árvore	0.9364	0.0713	0.9333
Naive Bayes	0.9393	0.0628	0.9333
KNN (n=1)	0.9164	0.0779	0.9333
KNN (n=3)	0.9366	0.0672	0.9333
KNN (n=5)	0.9475	0.0579	0.9333
WKNN (n=1)	0.9164	0.0779	0.9333
WKNN (n=3)	0.9393	0.0670	0.9333
WKNN (n=5)	0.9420	0.0651	0.9333
SVM-LINEAR	0.9489	0.0568	0.9333
SVM-RBF	0.9462	0.0590	0.9333

Observando a tabela percebe-se que a taxa de acerto (acurácia) fica em torno de 0,91 à 0,94 para todos os classificadores. As demais taxas também aparentam variar moderadamente. A baixo encontra-se a tabela com os dados obtidos através da hipótese t-studant. Os valores foram reduzidos para apenas compreenderem as primeiras 4 casas decimais após a virgula, para fins de visualização.

Figura 2. T-Stutant(Gosset) - Iris DataSet

Isis DataSet t-studant Hipotesis		
Comparação	t-value	p-value
KNN-1 e KNN-3	-1.5351	0.1279
KNN-1 e KNN-5	-2.3996	0.0183
KNN-3 e KNN-5	-0.7793	0.4376
WKNN-1 e WKNN-3	-1.6466	0.1028
WKNN-1 e WKNN-5	-1.9286	0.0566
WKNN-3 e WKNN-5	-0.2251	0.8223
SVM-LIN e SVM-RBF	0.3485	0.7281

Como pode ser observado os valores críticos obtidos são menores do que as probabilidades de hipótese para todas as variações de classificadores. Logo constata-se que para essa base de dados não houve alteração significativa entre resultados das KNNs alternando o valor k, nem entre os KNNs ponderados e as duas SVMs. Olhando novamente para a tabela de acurácia média entretanto, nota-se que os classificadores KNN ponderados e não ponderados com k=5 obtiveram uma taxa maior em relação aos demais KNNs. A árvore de decisão e o Naive Bays ficam praticamente empatados em termos de acurácia média, assim como as duas SVMs. Para a base Isis observa-se também que o desvio padrão das acurácias obtidas para cada classificador são menores que 0,1, demonstrando que para cada conjunto de testes a acurácia não variou muito. De maneira geral, todos os classificadores tiveram um desempenho semelhante para a base Iris.

5.2. Shuttle DataSet

A base de dados Shuttle também obteve bons índices de acerto para cada um dos classificadores, com baixa variação, como no data Set anterior. A analise da hipótese de t-studant entre as variações de classificadores mostrou que existem diferenças entre os métodos. Abaixo encontra-se a tabela de médias, medianas e desvio padrão do conjunto de acurácias de cada classificador.

Figura 3. Resultados - Shuttle DataSet

Shuttle DataSet			
Classificador	Média	Desvio Padrão	Mediana
Árvore	0.9968	0.0015	0.9965
Naive Bayes	0.8505	0.0185	0.8493
KNN (n=1)	0.9984	0.0009	0.9986
KNN (n=3)	0.9979	0.0012	0.9979
KNN (n=5)	0.9970	0.0014	0.9972
WKNN (n=1)	0.9984	0.0009	0.9986
WKNN (n=3)	0.9984	0.0009	0.9986
WKNN (n=5)	0.9979	0.0011	0.9979
SVM-LINEAR	0.9579	0.0041	0.9579
SVM-RBF	0.9265	0.0050	0.9255

As médias de acurácia são maiores que 0.9 para todos os classificadores, com excessão do Naive Bayes, que obteve um resultado de aproximadamente 0.85, indicando que a abordagem de probabilidade não se adequou tão bem ao problema como as demais. A seguir encontra-se a tabela de comparação entre as variações dos classificadores KNN, WKK e SVM (Hipótese de Gosset).

Podemos observar que os valores críticos obtidos pela hipótese de t-studant são maiores que a taxa de probabilidade da hipótese, indicando que existem variações significativas entre os classificadores comparados. É importante observar que as médias obtidas pelos pelos mesmos classificadores não variam muito (figura 3), logo, o teste estatístico encontrou variações na distribuição da acurácia que não seriam possíveis de visualizar apenas através da média.
Os classificadores SVM-Linear e SVM-RBF obtiveram um valor crítico muito maior que o limiar apontado pela probabilidade de hipótese, o que indica uma diferença significativa entre as duas abordagens. Na tabela de acurácia, podemos observar que o classificador SVM-Linear obteve uma média maior que o SVM-RBF em 0,03, podendo ser um reflexo dessa diferença.
Os classificadores que obtiveram as melhores médias foram os KNNs ponderados k=1 e k=3 e o não ponderado k=1. Na tabela apenas os 4 primeiros valores após a virgula foram computados, o que fez com que estes os classificadores tenham obtido a mesma média.

Figura 4. T-Stutant(Gosset) - Shuttle DataSet

Shuttle DataSet t-studant Hipotesis		
Comparação	t-value	p-value
KNN-1 e KNN-3	2.8989	0.0046
KNN-1 e KNN-5	5.5972	1.9834
KNN-3 e KNN-5	2.8321	0.0056
WKNN-1 e WKNN-3	0.6510	0.5165
WKNN-1 e WKNN-5	2.5412	0.0126
WKNN-3 e WKNN-5	1.8840	0.0625
SVM-LIN e SVM-RBF	30.3189	1.3585

5.3. Yaest DataSet

O dataSet Yeast obteve bons resultados através das técnicas de vizinhos próximos (KNNs) porem os resultados não foram replicados para os demais classificadores, que obtiveram médias insatisfatórias. A seguir a tabela com os valores de média, desvio padrão e mediana do conjunto de testes.

Figura 5. Resultados - Yeast DataSet

Yeast DataSet			
Classificador	Média	Desvio Padrão	Mediana
Árvore	0.3743	0.0155	0.3792
Naive Bayes	0.0337	0.0061	0.0327
KNN (n=1)	0.9972	0.0014	0.9971
KNN (n=3)	0.9969	0.0017	0.9971
KNN (n=5)	0.9958	0.0024	0.9961
WKNN (n=1)	0.9972	0.0014	0.9971
WKNN (n=3)	0.9970	0.0016	0.9971
WKNN (n=5)	0.9967	0.0016	0.9971
SVM-LINEAR	0.3636	0.0133	0.3610
SVM-RBF	0.3572	0.0144	0.3567

O classificador Naive Bayes obteve a menor taxa de acertos entre todos os outros algoritmos, quase não acertando, a árvore de decisão também obteve um índice baixo, com uma média de acertos de apenas 30%.
Este comportamento pode ser explicado pela dependência de ambos os algoritmos de eventos encadeados. Ex: Uma alta frequência de uma classe, dada a existência de um atributo especifico, ou dependência entre atributos, que com frequência possuem determinado valor quando um terceiro atributo possui um terceiro valor x. A inexistência desses casos em uma base de Dados pode levar a árvore de decisão e o Naive bayes à resultados inferiores.
Ambos os classificadores de support vector machine falharam na classificação, obtendo uma taxa inferior ao da árvore de decisão. O grande volume de dados do dataSet Yeast e a respectiva grande quantidade de atributos por caso influenciam nesse resultado, mostrando empiricamente que as SVMs não são adequadas a esse tipo de problema.
Abaixo encontra-se a tabela de comparação entre as variações dos classificadores.

O teste de hipóteses indica que não houve grandes variações entre os métodos KNN k=1 e k-3, porém houve uma mudança significativa entre o KNN k =5 e os demais. Os resultados se repetem para o KNN ponderado, não havendo grande variação utilizando k = 1 ou k =3. Observa-se na tabela de acurácias (figura 5) que os KNNs ponderados e não ponderados com k = 5 obtiveram um resultado inferior aos k = 1 e k =3.
Para esta base de dados ficou constatado que o uso de algoritmos probabilísticos como o

Figura 6. T-Stutant(Gosset) - Shuttle DataSet

Yeast DataSet t-studant Hipotesis		
Comparação	t-value	p-value
KNN-1 e KNN-3	0.4635	0.6440
KNN-1 e KNN-5	3.1613	0.0020
KNN-3 e KNN-5	2.7880	0.0063
WKNN-1 e WKNN-3	0.1201	0.9045
WKNN-1 e WKNN-5	0.7014	0.4847
WKNN-3 e WKNN-5	0.5958	0.5526
SVM-LIN e SVM-RBF	2.0375	0.0442

Naive Bayes e o SVM que usa uma função limiar, não são adequados para dataSets como o Yeast. O fato da SVM ter falhado em classificar este dataSet indica que os dados são bastante dispersos quanto a sua classificação e relação com os atributos, se montássemos um gráfico para esse dataSet, possivelmente verificaríamos que a função de separação entre as classes não linear.

6. Conclusão

Através dessa atividade proposta pode-se constatar principalmente que as métricas de avaliação nem sempre são suficientes para determinar qual o melhor classificador. Detalhes da distribuição de qualquer tipo amostra são ignorados ao apenas se considerar a média geral de um espaço amostral. Os testes de hipótese ajudam a visualizar melhor as pequenas variações que podem ser significantes para a avaliação final.
Também constata-se a importância de selecionar o classificador mais adequado para cada tipo de dado e problema de classificação. A performance de determinados classificadores podem não ser satisfatórias em alguns casos, fazendo valer o esforço de analisar primeiramente o tipo de dado a ser classificado, verificar as possibilidades de pre-processamento, que podem impactar no desempenho e no tipo de classificador que melhor se adequara ao problema.