Universidade Federal de Pernambuco - CIN

Aprendizagem de Máquina

Relatório da Lista de Exercícios 4

Feito por :
Arnaldo Rafael Morais Andrade
arma

Conteúdo

1 Objetivos

Implementar um classificador que faça uso do k-Means para cada uma das classes, obtendo o melhor k e em seguida aplicar o Naive Bayes. Comparar esse classificador gerado com 1-NN e com o Naive Bayes sem realizar o agrupamento. A avaliação dos algoritmos deve ser feita usando o *k-fold cross-validation* em dois bancos de dados distintos. O relatório deve conter os resultados e suas análises, além de informações suficientes para a replicação dos experimentos.

2 Metodologia dos Experimentos

As implementações foram feitas utilizando a linguagem Python.

2.1 Bases de Dados

As duas bases retiradas do repositório Promise são relacionadas à detecção de erros em software. São elas: datatrieve.arff e cm1.arff. A base DATATRIEVE possui 9 atributos e 130 instâncias que podem ser classificadas como "0"(91.54%) ou "1"(8.46%). Enquanto que a outra, CM1, dispõe de 22 *features* e 498 registros, dos quais podem ter os valores "false"(90.16%) ou "true"(9.83%).

Os dados de cada base foram padronizados de acordo com a equação (1):

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

2.2 Treino

Para a fase de treino a divisão dos conjuntos foi dada pelo *k-fold cross validation* estratificado, visto o desbalanceamento entre as classes. O número de *folds* escolhido foi 10.

Em cada iteração do *k-fold* o melhor k de cada classe, k*, foi definido através do método da silhueta [1].

2.3 Experimentos

Os experimentos foram realizados em uma máquina virtual Ubuntu 18.04.5 LTS, com 7.8 GB de memória e 3 núcleos de CPU virtuais.

3 Resultados e Discussões

Nesta seção, há 3 figuras para cada base de dados. As figuras 1, 2 e 3 pertencem ao conjunto de dados DATATRIEVE e as outras referem-se ao CM1.

A primeira figura mostra a escolha do melhor k, para cada classe, em cada iteração do *k-fold*.

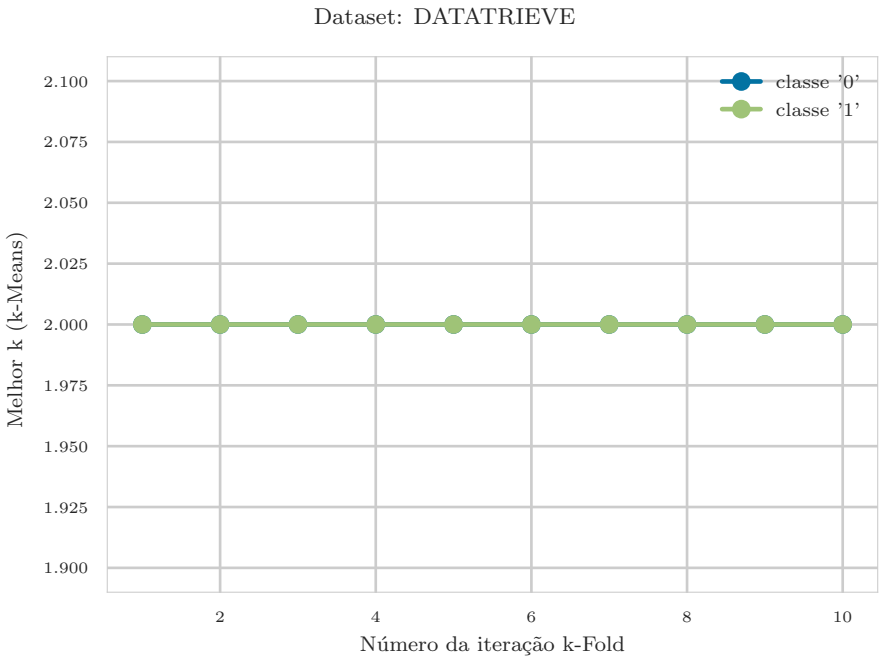

Figura 1: Melhor k (k-Means) para cada iteração do k-fold

Apesar de não claro, as duas retas estão sobrepostas, ou seja, o valor de k = 2 foi escolhido como o melhor para todas as variantes.

A partir do k* definido, foi possível comparar esta implementação com as outras 2: 1-NN e o Naive Bayes aplicado diretamente às bases, resultando na figura 2.

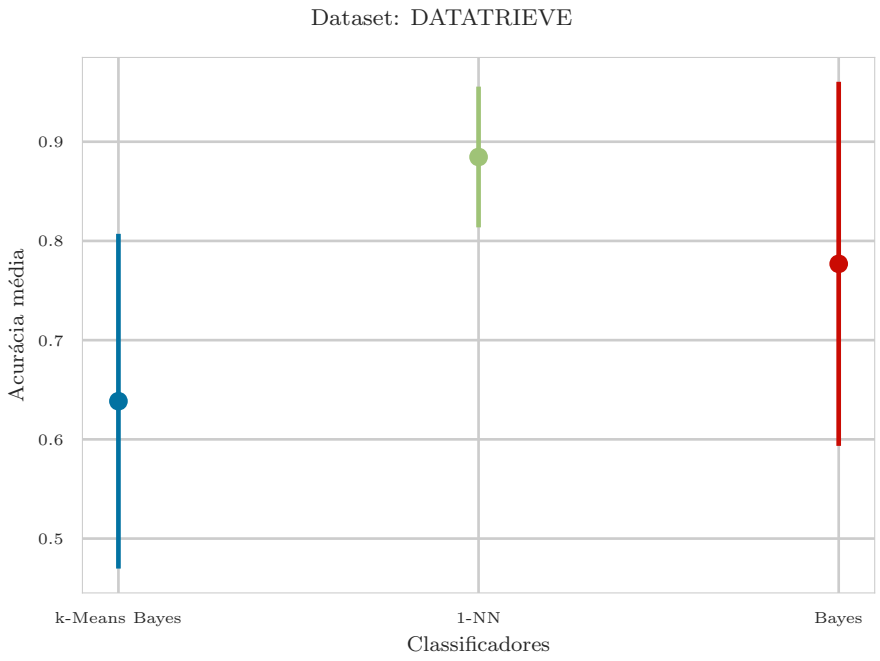

Figura 2: Acurácia média e desvio dos classificadores

Nota-se que para este problema, o agrupamento inicial prejudicou a acurácia média. O que pode indicar a não existência de subgrupos relevantes dentro de cada classe do banco de dados. Destaca-se os valores obtidos pela implementação do 1-NN, que mostrou-se o melhor.

A fim de uma comparação mais aprofundada, uma matriz de confusão foi construída para cada máquina. A construção delas foi a partir de um conjunto de testes estratificado contendo 25% instâncias, gerando assim, a figura 3

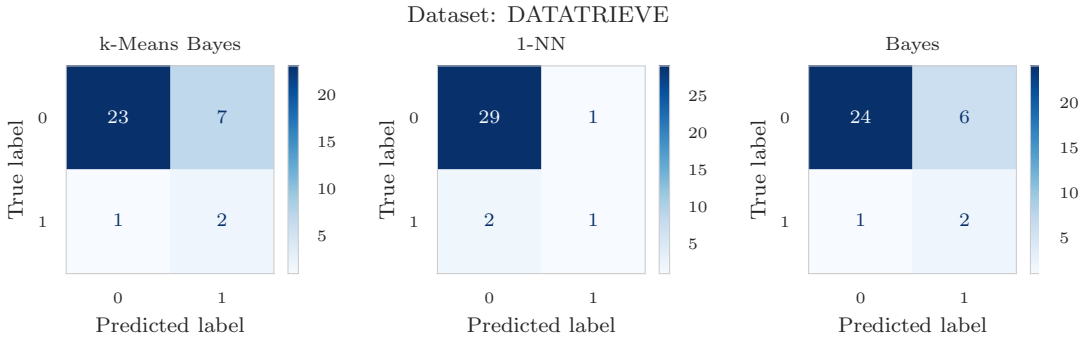

Figura 3: Matriz de confusão dos classificadores

Assim como em outros relatórios, vale destacar que a diagonal das matrizes estão invertidas, resultando no elemento $M_{0,0}$ como o TN e o elemento $M_{1,1}$ como o TP.

Dessa maneira é possível enfatizar os valores True Positive, False Positive e F1-measure de cada matriz como: k-Means Bayes(7, 2, 0.333), 1-NN(1, 1, 0.400) e Bayes(6, 2, 0.364). Nota-se que para esta divisão do conjunto de dados, o agrupamento não degradou significativamente os resultados, diferente do que pôde ser visto na imagem 2.

A próxima figura é semelhante à primeira mostrada. É relatado a escolha dos k* na fase de treinamento para a base CM1.

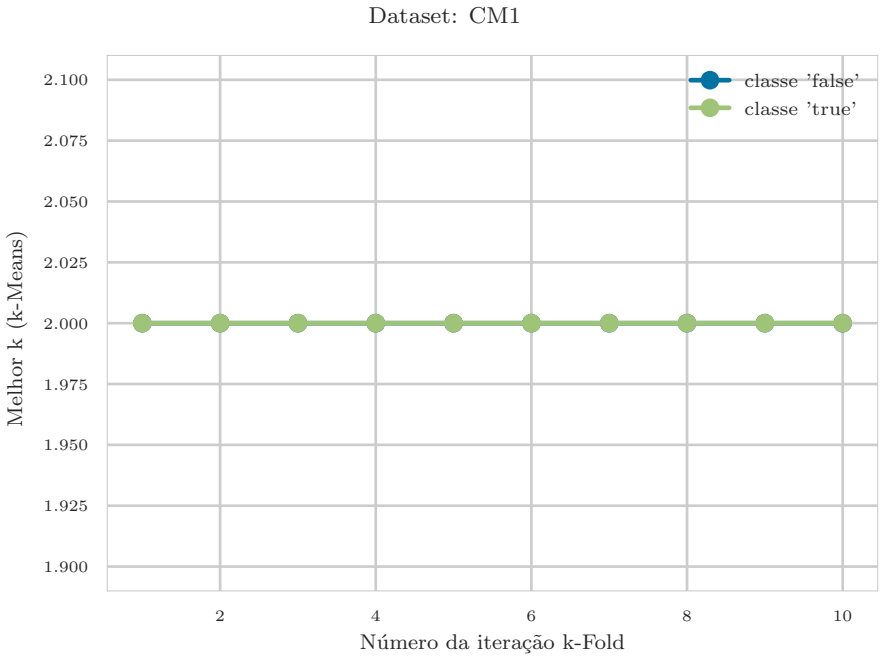

Figura 4: Melhor k (k-Means) para cada iteração do k-fold

Ainda que se trate de um conjunto de dados diferente, o valor de k manteve-se o mesmo em cada iteração do *k-fold*.

Entretanto, a acurácia média do classificador customizado teve um comportamento distinto, que é apresentado na figura 5.

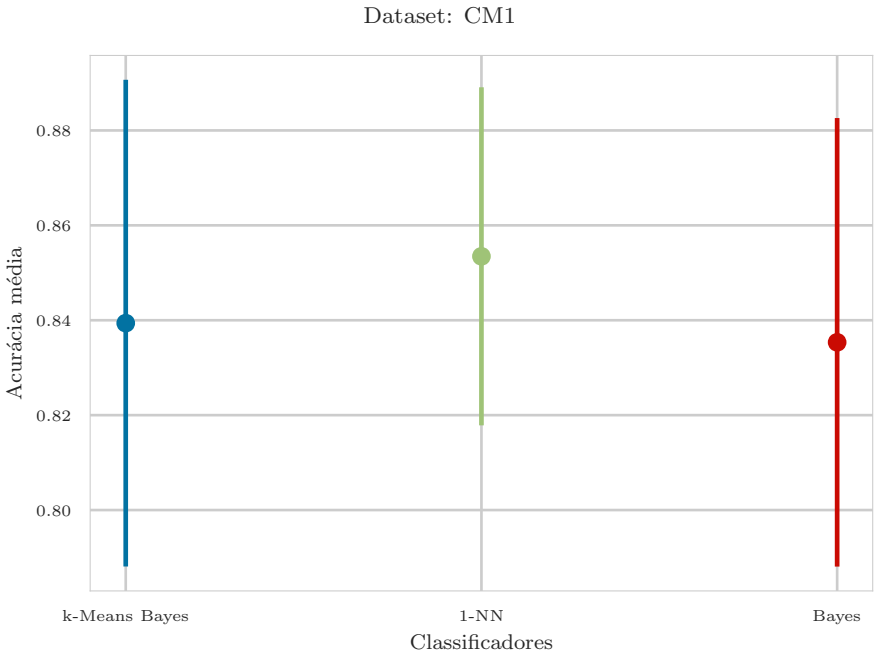

Figura 5: Acurácia média e desvio dos classificadores

Ela foi superior se comparada ao classificador Bayesiano sem o agrupamento inicial, apontando que as instâncias podem ainda ser agrupadas dentro da mesma classe. O Classificador 1-NN ainda manteve-se com a melhor acurácia média.

Por fim, a figura 6 exibe os erros e acertos por classe através das matrizes de confusão relacionadas ao *dataset* CM1.

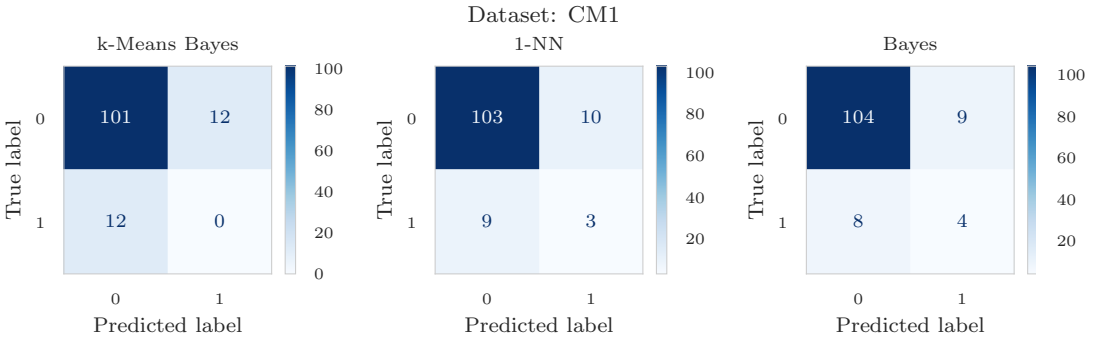

Figura 6: Matriz de confusão dos classificadores

A construção delas foi similar à realizada anteriormente na imagem 3. Para esta divisão do dados, as métricas True Positive, False Positive e F1-measure foram: k-Means Bayes(0, 12, 0.000), 1-NN(3, 10, 0.240) e Bayes(4, 9, 0.320). Desta vez, o Naive Bayes se mostrou o mais balanceado em relação à *precision* e *recall*.

4 Conclusões

Baseado nos resultados anteriores, a clusterização inicial se mostrou altamente dependente da natureza dos dados, deteriorando os resultados para a base DATATRIEVE e melhorando-os na outra base. Além disso, este método se atém ao conjunto pré-estabelecido dos valores de k e ao variá-lo as avaliações também mudam.

Bibliografia

[1] Peter J.Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, Volume 20, November 1987, Pages 53-65.