Universidade Federal de Pernambuco - CIN

Aprendizagem de Máquina

Relatório da Lista de Exercícios 1

Feito por :
Arnaldo Rafael Morais Andrade
arma

Conteúdo

1 Objetivos

Implementar classificadores k-NN, com e sem peso, e o k-NN adaptativo, utilizando a distância euclidiana. Avaliar as três máquinas em duas bases de dados distintas do repositório Promise. Estas bases devem contem apenas atributos numéricos. Construir um gráfico que mostre o comportamento da taxa de acerto à medida que o valor de k, referente ao conjunto {1,2,3,5,7,9,11,13,15}, muda para os três classificadores. Analisar os resultados em relação ao tempo de processamento e à taxa de acerto. Argumentar sobre as melhores escolhas para as bases de dados escolhidas.

2 Metodologia dos Experimentos

Para a implementação do k-NN e suas variações foi utilizada a linguagem Python, bem como a distância euclidiana.

2.1 Bases de Dados

As duas bases escolhidas foram retiradas do repositório Promise, sendo a primeira datatrieve.arff e a segunda cm1.arff. A base DATATRIEVE possui 9 atributos e 130 instâncias que podem ser classificadas como "0"(91.54%) ou "1"(8.46%). Já a base CM1 dispõe de 22 *features* e 498 registros, dos quais podem ter os valores "false"(90.16%) ou "true"(9.83%).

Em ambos conjuntos de dados, cada atributo foi normalizado de acordo com o método *min-max*, descrito pela equação (1).

$$x_{norm} = a * \frac{x - x_{min}}{x_{max} - x_{min}} + b \tag{1}$$

2.2 Treino

Para a fase de treino, a divisão dos conjuntos foi dada pelo método de validação cruzada *k-fold* estratificado, visto o desbalanceamento entre as classes das bases de dados. O número de *folds* escolhido foi 10.

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

Para cada base de dados 3 figuras foram construídas. As figuras 1, 2 e 3 pertencem ao conjunto de dados DATATRIEVE, já as outras se referem à base de dados CM1.

A figura 1 mostra separadamente a taxa de acerto de cada variação do k-NN em relação ao número de vizinhos k. O eixo y de cada gráfico foi limitado à [0.7,1.0] para uma melhor visualização.

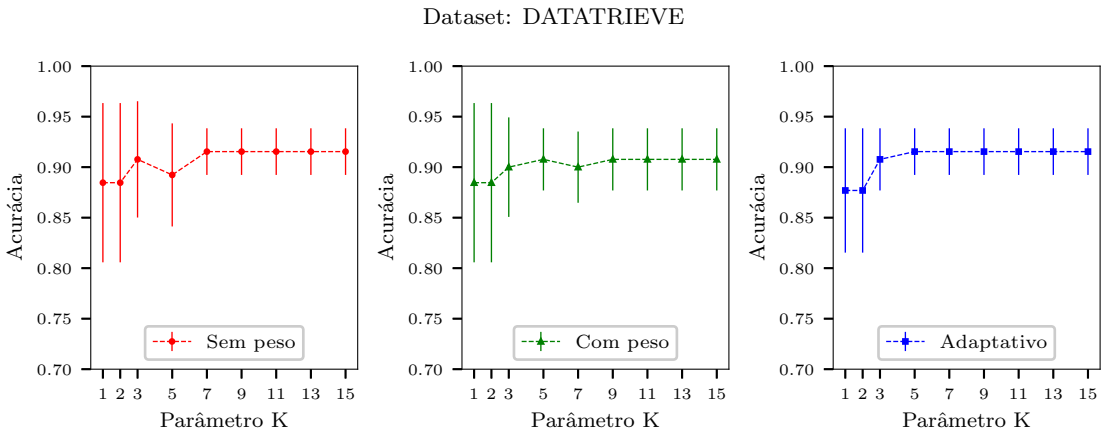

Figura 1: Média e desvio padrão dos 3 classificadores k-NN

É possível afirmar que a medida que o número de vizinhos aumenta, a média da acurácia dos classificadores também aumenta e o desvio padrão diminui.

A abordagem da figura 2 é diferente. Ela agrupa os diferentes classificadores em dois gráficos. No gráfico da esquerda, o eixo y, tempo de processamento em segundos, é dado pela soma do tempo de treinamento e o tempo de teste.

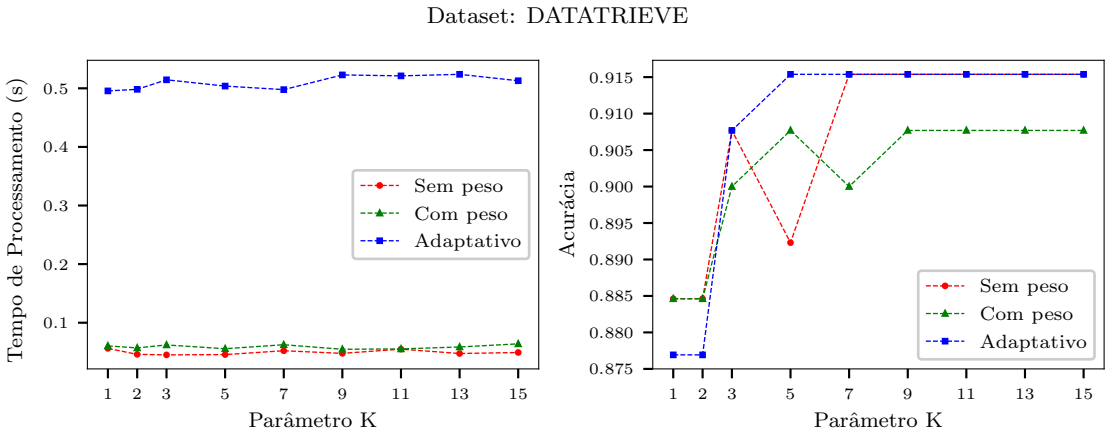

Figura 2: Tempo de processamento e acurácia em relação à k vizinhos

Nota-se que a variação adaptativa tem um tempo de processamento superior às outras variações. Isto se deve à fase de treinamento, quem além de armazenar os exemplos de treinamento, também armazena a maior esfera que exclui todos os padrões das outras classes para cada instância de treino. Este procedimento é proporcional ao quadrado da quantidade de dados de treinamento. Não há diferenças relevantes na fase de testes das implementações.

Outra observação é de que a variação adaptativa atinge a maior média de acurácia utilizando 5 vizinhos. A implementação ponderada se mostra inferior às outras duas em relação à taxa de acerto.

Na figura a seguir é mostrado a matriz de confusão não normalizada para cada um dos classificadores. A escolha do k foi 5 por ser um número padrão utilizado em diversas abordagens. Como a base de dados DATATRIEVE classifica os padrões de forma binária, a matriz compõe os 4 tipos de valores: falsos positivos (FP), falsos negativos (FN), verdadeiros positivos (TP) e verdadeiros negativos (TN). O detalhe é que a diagonal das matrizes estão invertidas, diferente do usual, resultando assim no elemento $M_{0,0}$ como o TN e o elemento $M_{1,1}$ como o TP.

Para a construção das matrizes, foi utilizado um conjunto de testes estratificado contendo 25% das instancias.

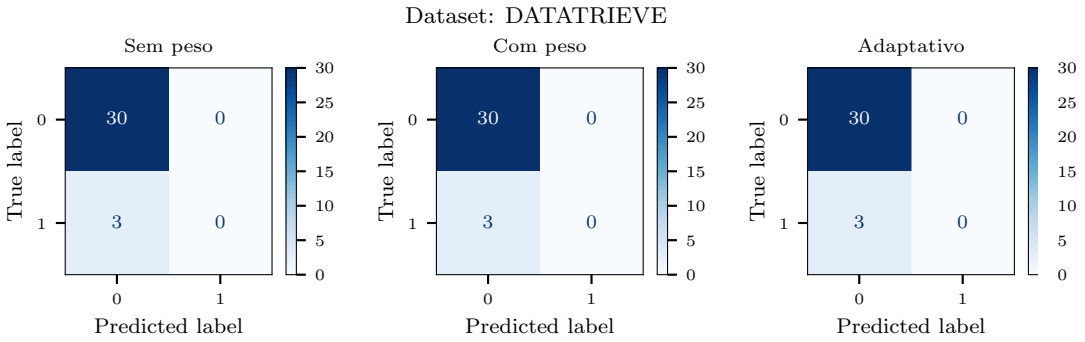

Figura 3: Matriz de confusão dos 3 classificadores k-NN para k = 5

Baseado nas matrizes, os 3 modelos se mostram equivalentes. Para todos eles, todos os casos foram previstos como negativos, pois TP + FP = 0.

Assim como a análise da figura 1, a figura a seguir mostra o mesmo comportamento. K aumenta e, portanto, a média da acurácia também aumenta, enquanto o desvio padrão diminui.

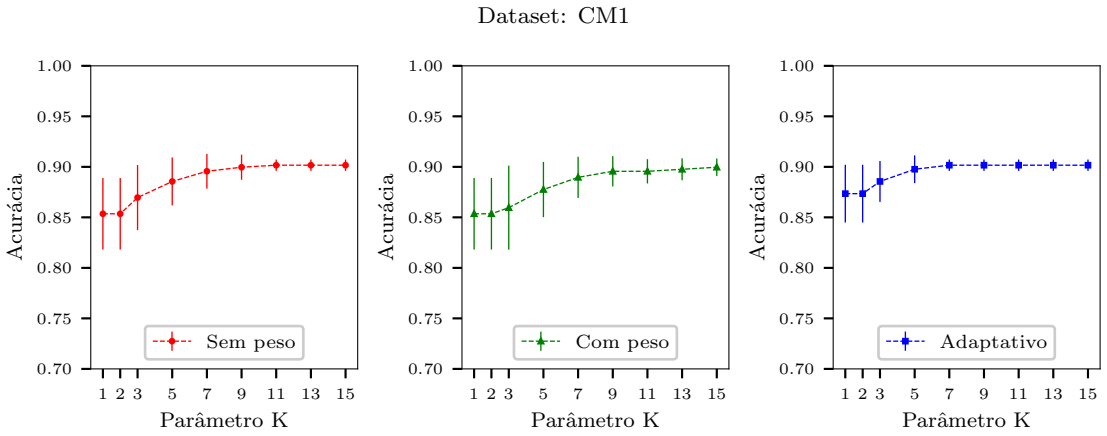

Figura 4: Média e desvio padrão dos 3 classificadores k-NN

Nos agrupamentos da figura 5 percebe-se que por se tratar de uma base de dados maior, tanto em número de atributos quanto em número de instancias, o tempo de processamento se elevou, se comparado ao *dataset* anterior. Entretanto, o mesmo comportamento é observado.

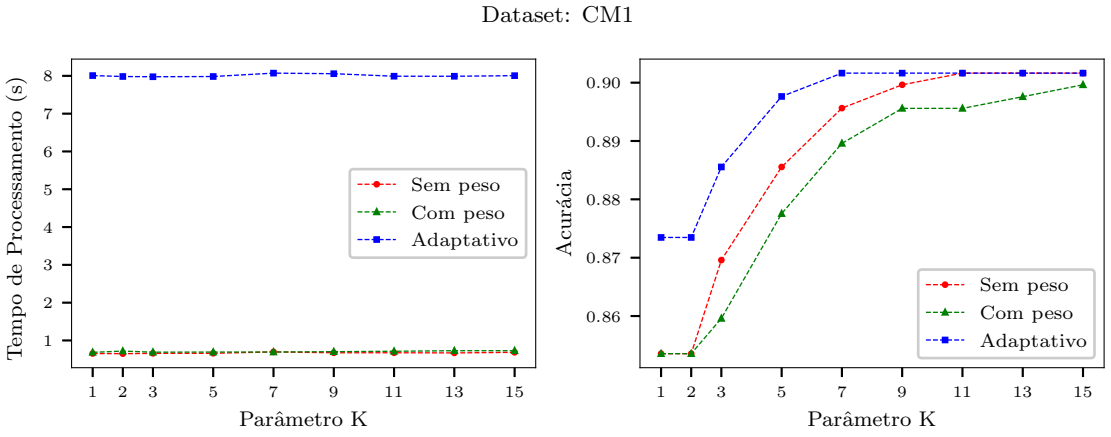

Figura 5: Tempo de processamento e acurácia em relação à k vizinhos

Por fim, na figura abaixo, temos as matrizes de confusão não normalizada. A diagonal delas também está invertida, como no exemplo passado. O processo de construção delas também se repetiu.

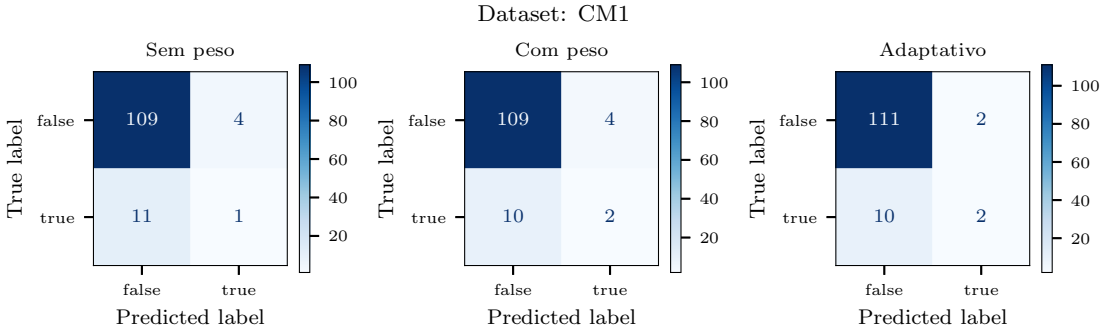

Figura 6: Matriz de confusão dos 3 classificadores k-NN para k = 5

Embora as matrizes sejam semelhantes em valores, cada uma possui sua particularidade. O modelo adaptativo apresenta a maior taxa de acerto (113/125) e também a maior precisão (2/4). Logo em seguida vem o k-NN ponderado, com uma taxa de acerto de 111/125 e uma precisão de 2/6.

4 Conclusões

Para a base DATATRIEVE, baseado nos resultados e discussões anteriores, como se trata de um conjunto de dados relativamente pequeno e como as 3 máquinas apresentaram eficácias semelhantes, basta escolher, portanto, qualquer um dos modelos.

Já para a base CM1, se tratando de uma base de dados mais robusta, o tempo de processamento deve ser levado em consideração. Por isso, o modelo ponderado se mostra uma melhor escolha. Caso o tempo de processamento seja desconsiderado (computador mais potente), o k-NN adaptativo se torna a melhor opção.