Classificação de vinhos a partir das componentes físico-químicas utilizando KNN

Lucas Coitinho Brito

UECE - Mestrado Acadêmico em Ciência da Computação
Fundamentos e Análise de Dados

8 de maio de 2019

Sumário

KNN

- KNN(K-Nearest Neighbors) é um dos muitos algoritmos usado no campo de machine learning, ele é um classificador onde o aprendizado é baseado "no quão similar" é um dado (um vetor) do outro. O treinamento é formado por vetores de n dimensões.

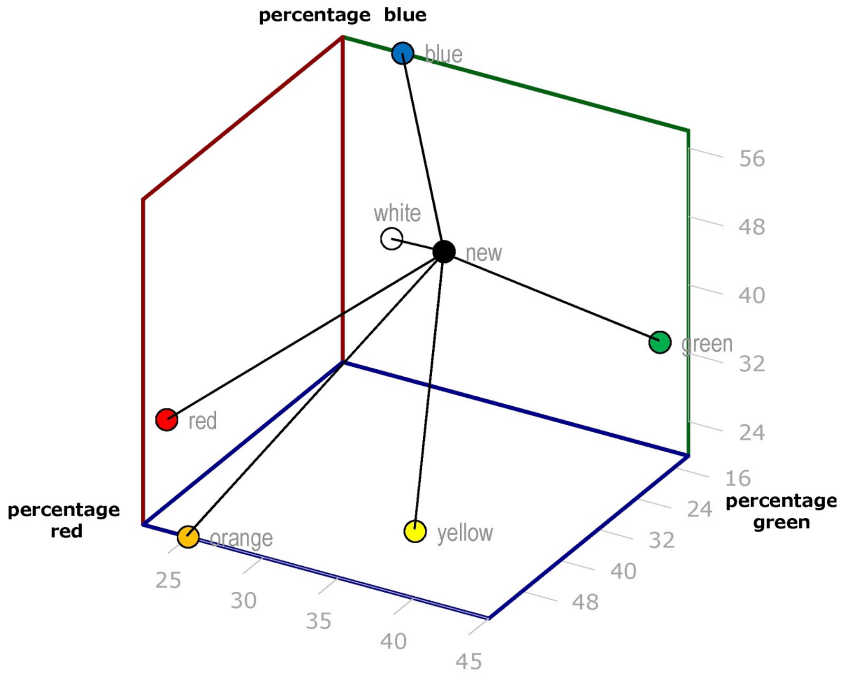

Figura: Exemplo de aplicação do KNN em um espaço de 3 dimensões

Motivação

- O conjunto escolhido está relacionado com variantes **brancas** do vinho "Vinho Verde". As entradas incluem testes objetivos (por exemplo, valores de PH) e a saída é baseada em dados sensoriais (mediana de pelo menos 3 avaliações feitas por especialistas em vinho). Cada perito classificou a qualidade do vinho entre 0 (muito ruim) e 10 (muito bom).

Motivação

- Características do conjunto de dados: Multivariada
- Características do atributo: Real
- Tarefas Associadas: Classificação
- Número de instâncias: 4899
- Número de Atributos: 11, sendo estes, AF: acidez fixa, AV: acidez volátil, AC: ácido cítrico, AR: açúcar residual, Cl: cloretos, SO_2 L: dióxido de enxofre livre SO_2 T: total de dióxido de enxofre, D: densidade, pH: potencial Hidrogeniônico, S: sulfatos, C_2H_6O: álcool, Q: qualidade

Parâmetros

- Distâncias:
  - Manhattan ou L_1:
    - $$dist_{L_1}(X, Y) = |x_1 - x_2| + |y_1 + y_2| \quad (1)$$
  - Euclidiana ou L_2
    - $$dist_{L_2}(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \quad (2)$$
- Classificação:
  - Mais votado
  - Ponderado pelo inverso da distância
- O *dataset* foi dividido de tal forma que 70% das amostras fosse separadas para treinamento e o restante para teste.

Resultados

- Após um teste unitário com $k = 5$, e variando os parâmetros de distância e classificadores, obtemos a seguinte tabela:

weight	distance	accuracy
uniform	L_1	53.4%
uniform	L_2	54%
distance	L_2	63.4%
distance	L_2	62.5%

Tabela: Variações do método KNN com $k = 5$

Resultados

- Utilizando a distância manhattan e o critério de voto ponderado, atingimos uma acurácia de 63.4% de tipos de vinhos classificados indicados corretamente.

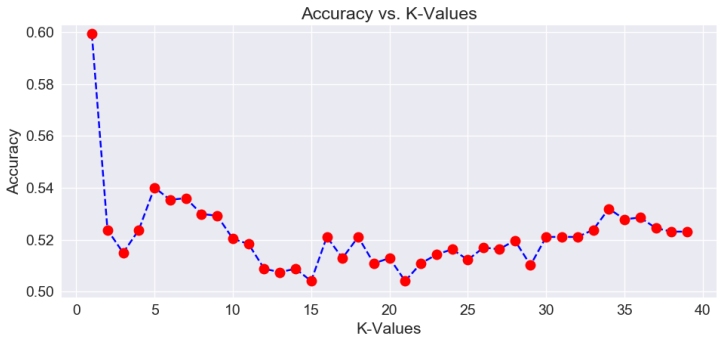

Figura: Aplicação do KNN com k variando 3 a 40 vizinhos

Matriz de Confusão

- o *dataset* possui grande parte dos seus vinhos concentrados entre as notas 2 e 3. Sendo a classe 3 mais comum e o acerto do algoritmo é preciso para esta classe.

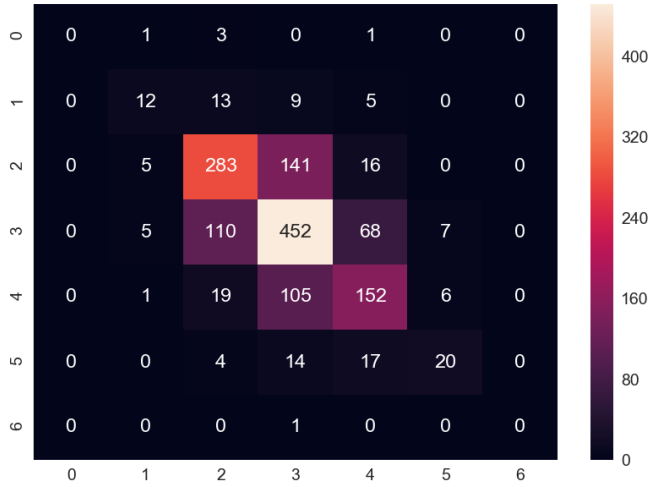

Figura: Matriz de confusão dos *dataset* referente ao vinho branco.

Métricas

- Como parte do relatório, devemos aplicar métricas de avaliação no método afim de qualificar sua análise, seja:

classe	precision	recall	f-score	support
3	0.00	0.00	0.00	5
4	0.50	0.31	0.38	39
5	0.66	0.64	0.6	445
6	0.64	0.71	0.67	642
7	0.61	0.55	0.58	283
8	0.54	0.38	0.45	55
9	0.00	0.00	0.00	1

Tabela: Resultados para cada classe dos parâmetros: `precision`, `recall`, `f-score`, `support`

Métricas

- A tabela 2 mostra as métricas utilizadas para avaliar o algoritmo, portanto, `precision` é a relação $\frac{\texttt{TP}}{\texttt{TP+FP}}$ onde `TP` é o número de positivos verdadeiros e `FP` o número de falsos positivos. A precisão é intuitivamente a capacidade do classificador de não rotular como positiva uma amostra que é negativa.
- O `recall` é a relação $\frac{\texttt{TP}}{\texttt{TP+FN}}$ onde `TP` é o número de positivos verdadeiros e `FN` o número de falsos negativos. O `recall` é intuitivamente a capacidade do classificador de encontrar todas as amostras positivas.
- O `f-score` pode ser interpretado como uma média harmônica ponderada de `precision` e `recall`, em que `f-score` atinge seu melhor valor em 1 e o pior escore em 0.
- E finalmente, `suporte` é o número de ocorrências de cada classe no conjunto de teste.

Conclusão

- Como visto na Seção anterior, o melhor resultado foi obtido quando k=5 e utilizando a distância de manhattan com ponderação, resultando em 64.3% de acerto.
- Deste modo, podemos afirmar que o KNN não é eficiente em sua forma clássica para este problema e necessita de incrementos no seu procedimento para que ele se torne mais "acurado".

Referências Bibliográficas

- P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.
- Raschka, Sebastian (2015). Python Machine Learning, Packt Publishing.