

Universidade Federal de Pernambuco

Graduação em Ciência da Computação

Centro de Informática

2016.1

**Lista de Exercício 1**

Aprendizagem de máquina

**Aluno:** Rubens Lopes de Farias Silva (rlfs@cin.ufpe.br)

**Professor:** George Darmiton da Cunha Cavalcanti (gdcc@cin.ufpe.br)

Recife, Abril de 2016.

# Introdução

K-NN é um dos mais métodos de reconhecimento de padrão mais antigos, porém ele ainda consegui produzir bons resultados para base de dados atuais. Para exemplificar sua capacidade de reconhecimento, foram realizados alguns experimentos com base de dados do UCI [1] e os resultados estão descritos nesse relatório. No total foram usados três métodos de calculo de distância: euclidiana[2], vdm[3] e hvdm[4]. Para cada distância usada, foram escilhidas duas bases de dados, calculado o tempo excecução do algoritmo com cada base e a acurácia para os valores de K = {1, 2, 3, 5, 7, 9, 11, 13, 15}.

# Escolha dos Dados

As bases de dados foram escolhidas dependendo da necessidade para realizar o teste de cada distância. Nesse caso, foram escolhidas duas com apenas dados núméricos para testar a distância euclidiana, duas com apenas dados categóricos para testar a distância vdm e duas com dados mistos (numéricos e categóricos) para testar aa distância hvdm.

Bases apenas numéricas:

1. [Letter Recognition](https://archive.ics.uci.edu/ml/datasets/Letter+Recognition)[5]: O objetivo dessa base é reconhecer qual letra uma imagem representa. A base possui 17 variáveis e 20.000 instâncias, onde a 1ª é a letra no qual devesse reconhecer.
2. [Banknote Authentication](https://archive.ics.uci.edu/ml/datasets/banknote+authentication)[6]: O objetivo dessa base é reconhecer qual classe cada nota (de banco) pertence. A base possui 5 variáveis e 1372 instâncias, onde a 5ª é o número que representa a class.

Bases apenas categóricas:

1. [Balance Scale](https://archive.ics.uci.edu/ml/datasets/Balance+Scale)[7]: O objetivo dessa base é identificar se a balança vai tender para algum lado ou ficar parada. A base possui 5 variáveis e 625 instâncias, onde a 1ª é a letra que representa a classe.
2. [Fertility](https://archive.ics.uci.edu/ml/datasets/Fertility)[8]: O objetivo dessa base é identificar se o candidato produz um esperma normal ou alterado. A base possui 10 variáveis e 100 instâncias, onde a 10ª é a class a ser avaliada.

Bases mista (numéricos e categóricos):

1. [Contraceptive Method Choice](https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice)[9]: O objetivo dessa base é identificar qual o método contraceptivo usado. A base possui 10 variáveis e 1473 instâncias, onde 1ª é a classe a ser avaliada e [2,3,5,6,7,8,9] são categóricos.
2. [Teaching Assistant Evaluation](https://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation)[10]: O objetivo dessa base é identificar qual o desempenho de um monitor. A base possui 6 variáveis e 151 instâncias, onde 6ª é a classe a ser avaliada e [1,2,3,4] são categóricos.

# Pré-Processamento e Carregamento dos Dados

Apenas uma base escolhida preisou passar por uma etapa de pré-processamento (Banknote Authentication). Nessa base foi necessário realizar a normalização dos dados para que todos os valores ficassem no intervalo [0, 1].

A separação de instâncias entre treinamento e teste foi de 80% para treinamento e 20% para teste de forma aleatória. Essa separação é realizada automaticamente, caso não foi realizada previamente, criando uma basta com o mesmo nome do arquivo da base de dados com dois arquivos internos *Traning.data* e *Testing.data*. Assim se for necessário realizar a divisão novamente, basta excluir a pasta e rodar o *script* novamente.

# Fluxo de Execução

## Divisão:

O projeto está separado em três pastas de acordo a questão (Q1,Q2 e Q3). Dentro de cada pasta há três arquivos bases: init.py, kNN.py e util.py. Para realizar a execução do script devesse executar o arquivo init.py. Para facilitar a execução com bases diferentes sem a necessidade de reconfiguração, foi adicionado os arquivos init1.py e init2.py que rodam a primeira e a segunda base de acordo com a função de distância da questão respectivamente.

## Configuração:

Para realizar a configuração de uma nova base de dados é preciso prencher em *hardcode* os seguintes variáveis:

* Database //representa o caminho da base de dados
* Classpos //representaa a posição da vairável a ser inferida

Para o caso de das questão em hvdm há uma vairável adicional:

* isCategore // Que é uma lista com todas as posições das variáveis que são categorias (iniciándo do 0).

Caso precise normalizar os dados, devesse modificar a variável *normal* para *True* e realizar as configurações de forma semelhante a anterior. Obs. a normalização está apenas disonível na questão 1 (distância eucliana).

## Resultado:

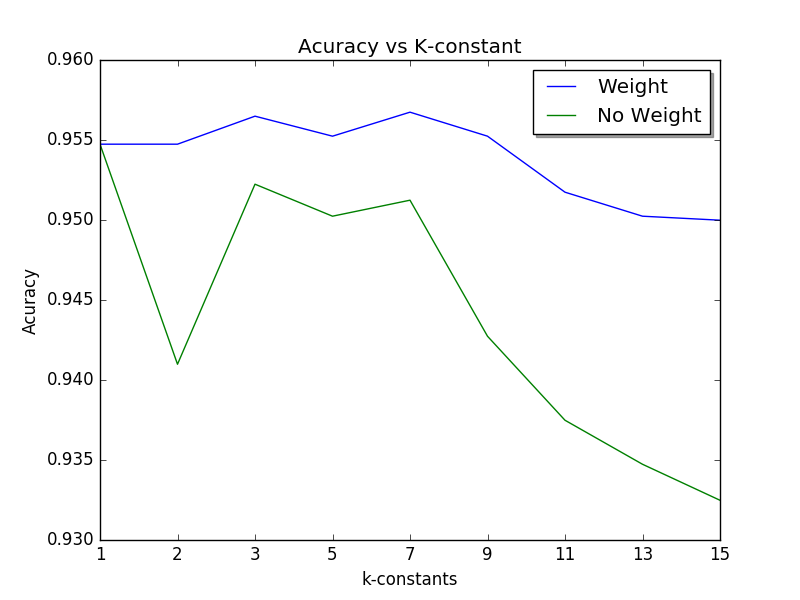
O resultado é impresso como a acurácia de cada valor de k usando os pesos (1/1+distância) e sem usar-los. Caso o usuário possua a class *matplotli*[11] intalada em seu computador, um gráfico será plotado para facilitar a comparação dos resultados.

# Resultados Finais

## Questão 1

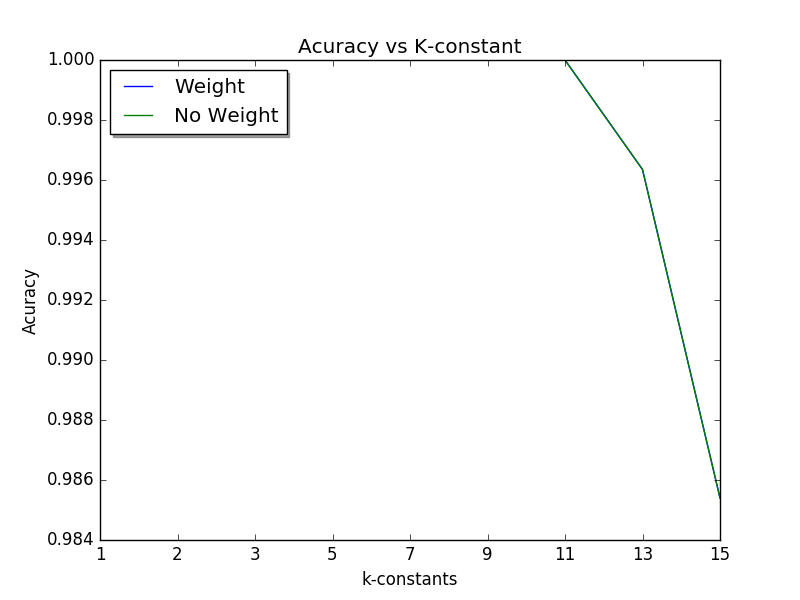
Letter Recognition:

|  |  |  |
| --- | --- | --- |
| K/Peso | Não | Sim |
| 1 | 0.9547386846711677 | 0.9547386846711677 |
| 2 | 0.9409852463115779 | 0.9547386846711677 |
| 3 | 0.95223805951487 | 0.9564891222805701 |
| 5 | 0.9502375593898474 | 0.9552388097024256 |
| 7 | 0.9512378094523631 | 0.956739184796199 |
| 9 | 0.9427356839209803 | 0.9552388097024256 |
| 11 | 0.9374843710927732 | 0.9517379344836209 |
| 13 | 0.9347336834208552 | 0.9502375593898474 |
| 15 | 0.932483120780195 | 0.9499874968742186 |



Banknote Authentication:

|  |  |  |
| --- | --- | --- |
| K/Peso | Não | Sim |
| 1 | 1.0 | 1.0 |
| 2 | 1.0 | 1.0 |
| 3 | 1.0 | 1.0 |
| 5 | 1.0 | 1.0 |
| 7 | 1.0 | 1.0 |
| 9 | 1.0 | 1.0 |
| 11 | 1.0 | 1.0 |
| 13 | 0.9963503649635036 | 0.9963503649635036 |
| 15 | 0.9854014598540146 | 0.9854014598540146 |

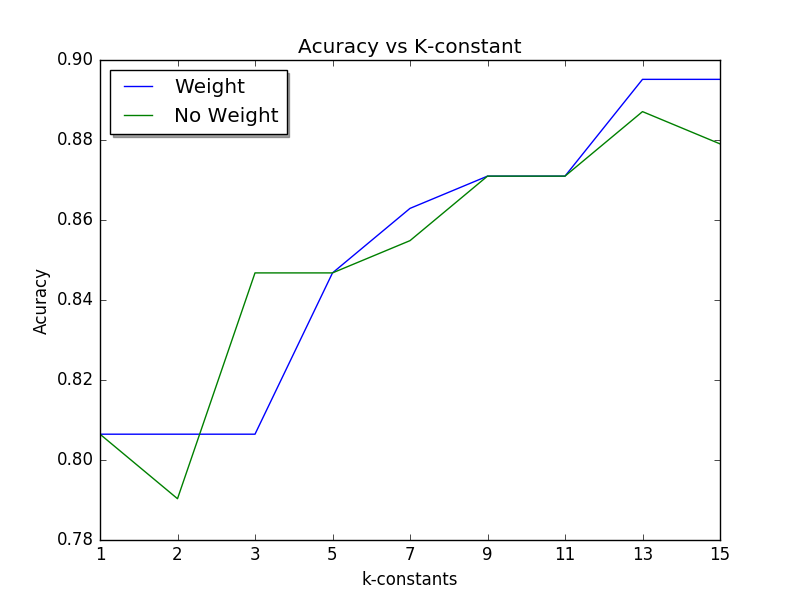


\*Os dois casos (com peso e sem peso) obtiveram os mesmos resultados.

## Questão 2

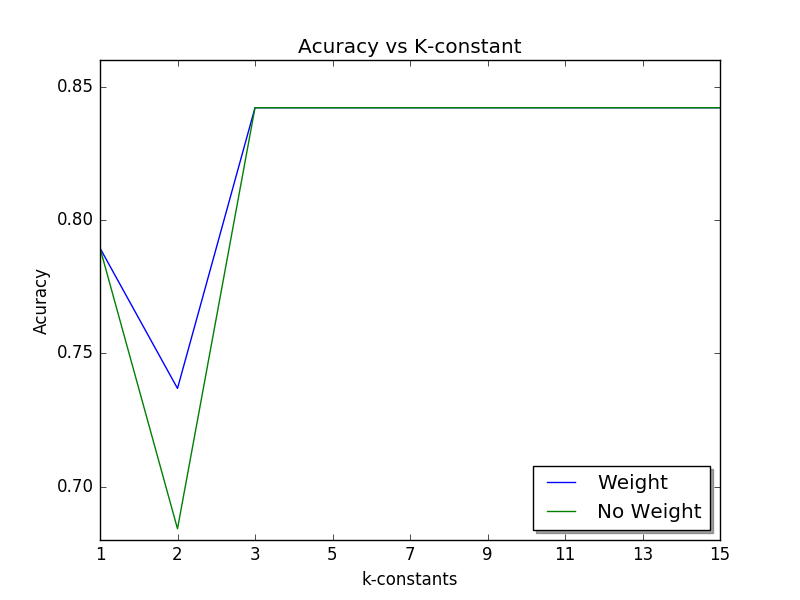
Balance Scale:

|  |  |  |
| --- | --- | --- |
| K/Peso | Não | Sim |
| 1 | 0.8064516129032258 | 0.8064516129032258 |
| 2 | 0.7903225806451613 | 0.8064516129032258 |
| 3 | 0.84677419354838 | 0.8064516129032258 |
| 5 | 0.8467741935483871 | 0.8467741935483871 |
| 7 | 0.8548387096774194 | 0.8629032258064516 |
| 9 | 0.8709677419354839 | 0.8709677419354839 |
| 11 | 0.8709677419354839 | 0.8709677419354839 |
| 13 | 0.8870967741935484 | 0.8951612903225806 |
| 15 | 0.8790322580645161,73684 | 0.8951612903225806 |



Fertility:

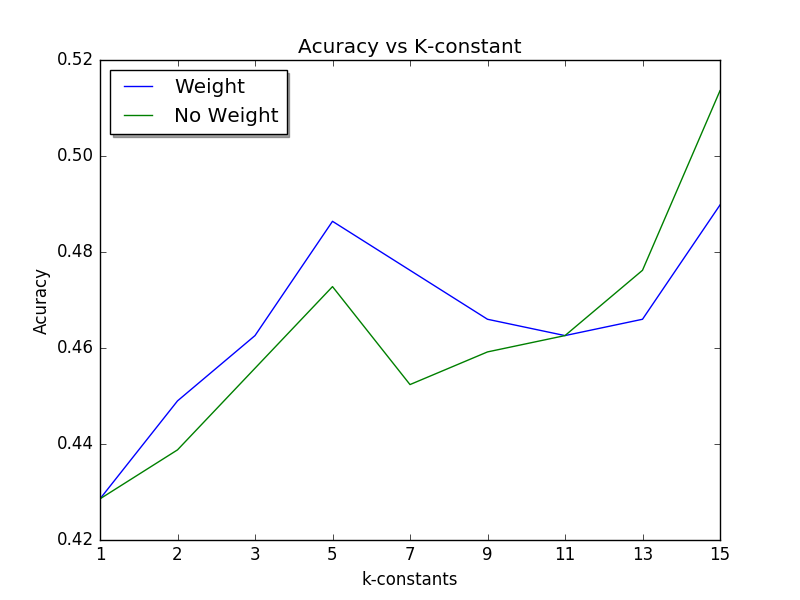
|  |  |  |
| --- | --- | --- |
| K/Peso | Não | Sim |
| 1 | 0.7894736842105263 | 0.7894736842105263 |
| 2 | 0.6842105263157895 | 0.7368421052631579 |
| 3 | 0.8421052631578947 | 0.8421052631578947 |
| 5 | 0.8421052631578947 | 0.8421052631578947 |
| 7 | 0.8421052631578947 | 0.8421052631578947 |
| 9 | 0.8421052631578947 | 0.8421052631578947 |
| 11 | 0.8421052631578947 | 0.8421052631578947 |
| 13 | 0.8421052631578947 | 0.8421052631578947 |
| 15 | 0.8421052631578947 | 0.8421052631578947 |



## Questão 3

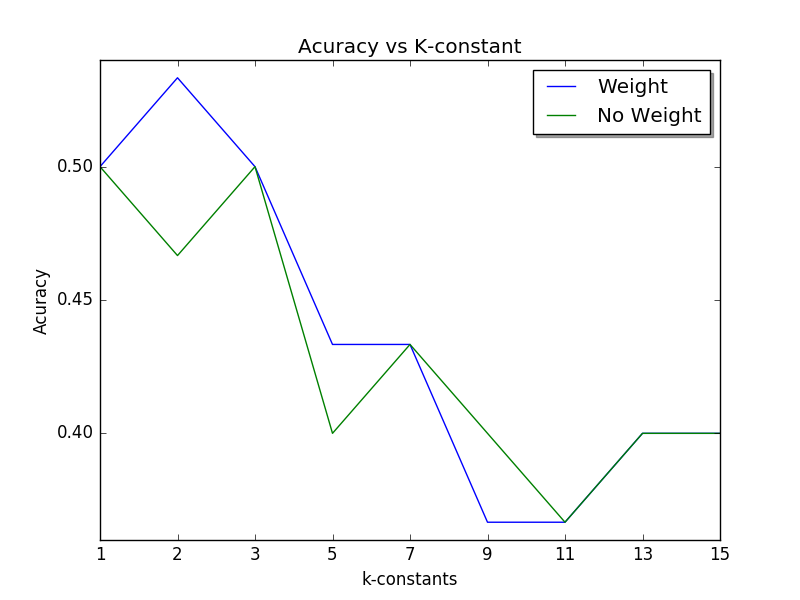
Contraceptive Method Choice:

|  |  |  |
| --- | --- | --- |
| K/Peso | Não | Sim |
| 1 | 0.42857142857142855 | 0.42857142857142855 |
| 2 | 0.4387755102040816 | 0.4489795918367347 |
| 3 | 0.4557823129251701 | 0.46258503401360546 |
| 5 | 0.47278911564625853 | 0.48639455782312924 |
| 7 | 0.4523809523809524 | 0.47619047619047616 |
| 9 | 0.45918367346938777 | 0.46598639455782315 |
| 11 | 0.46258503401360546 | 0.46258503401360546 |
| 13 | 0.47619047619047616 | 0.46598639455782315 |
| 15 | 0.5136054421768708 | 0.4897959183673469 |



Teaching Assistant Evaluation:

|  |  |  |
| --- | --- | --- |
| K/Peso | Não | Sim |
| 1 | 0.5 | 0.5 |
| 2 | 0.4666666666666667 | 0.5333333333333333 |
| 3 | 0.5 | 0.5 |
| 5 | 0.4 | 0.43333333333333335 |
| 7 | 0.43333333333333335 | 0.43333333333333335 |
| 9 | 0.4 | 0.36666666666666664 |
| 11 | 0.36666666666666664 | 0.36666666666666664 |
| 13 | 0.4 | 0.4 |
| 15 | 0.4 | 0.4 |



# Questão 4

Na maioria dos resultados foi comprovado que nem sempre o aumento do K repercurte na melhoria dos resultados, dessa forma foi de fundamental importância variar o K para buscar o melhor resultado. Como a avaliação KNN não usa recursos de aleatoriedade então só foi necessário rodar uma vez para cada conjunto de treinamento e teste. Se deseja adicionar a aleatoriedade, devesse deletar a pasta que tem o teste e treinamento para que ele seja recriado.

## Tempo

Por motivo de não ser prioridade, o tempo de excecução foi bastante prejudicado devido ao uso das funcionalidades que facilitaram a programação oferecidas pela linguagem python. Como o tempo de treinamento é despresível no algoritmo de KNN, o tempo total de excecução foi baseado apenas no tempo de teste e preparação dos resultado. Os resulados serão descritos no formato QxBy onde representa a questão x e a base y.

* Q1B1: 477.616000175 segundos. (há muitas instâncias)
* Q1B2: 1.04999995232 segundos.
* Q2B1: 1.09299993515 segundos.
* Q2B2: 0.0590000152588 segundos. (há poucas instâncias)
* Q3B1: 0.113999843597 segundos.
* Q3B2: 13.7669999599 segundos.

## Desempenho

Com esses resultado plotados em um gráfico, fica melhor analizar os resultados de K e o uso do peso que tiveram melhor resultado em cada base. Os resultados serão descritonos no mesmo formato já apresentados na seção Tempo:

* Q1B1: k = 7, usando o peso.
* Q1B2: k < 13, unsando ou não o peso.
* Q2B1: k = 13, usando o peso.
* Q2B2: k > 3, usando ou não o peso.
* Q3B1: k = 2, usando o peso.
* Q3B2: k =15, sem usar o peso.

# Referências

* 1 - *UC Irvine Machine Learning Repository* (website: <https://archive.ics.uci.edu/ml/datasets.html> acessado em 04/08/2016)
* 2 – Euclidean Distance (website: <https://en.wikipedia.org/wiki/Euclidean_distance> acessado em 04/08/2016)
* 3 – VDM distance (website: <http://axon.cs.byu.edu/~randy/jair/wilson2.html> acessado em 04/08/2016)
* 4 – Wilson, D. Randall; Martinez, Tony R. *Improved Heterogeneous Distance Functions* . Journal of Artificial Intelligence Research 6 (1997) 1-34. (Disponível em: <https://www.jair.org/media/346/live-346-1610-jair.pdf> acessado em 04/08/2016)
* 5 – Letter Recognition Data Set (website: <https://archive.ics.uci.edu/ml/datasets/Letter+Recognition> acessado em 04/08/2016)
* 6 – Banknote Authentication Data Set (website: https://archive.ics.uci.edu/ml/datasets/banknote+authentication acessado em 04/08/2016)
* 7 – Balance Scale Data Set (website: <https://archive.ics.uci.edu/ml/datasets/Balance+Scale> acessado em 04/08/2016)
* 8 - Fertility Data Set (website: <https://archive.ics.uci.edu/ml/datasets/Fertility> acessado em 04/08/2016)
* 9 – Contraceptive Method Choice Data Set (website: <https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice> acessado em 04/08/2016)
* 10 – Teaching Assistant Evaluation Data Set (website: <https://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation> acessado em 04/08/2016)
* 11 – matplotli (Disponível em: <http://matplotlib.org> acessado em 04/08/2016)