Centro de Informática UFPE
Aprendizagem de Máquina – 2015.1
Aluno: Danilo Neves Ribeiro (dnr2)

# Relatório - Lista de Exercício #1

## 1 – Metodologia dos Experimentos

Os experimentos foram realizados utilizando as seguintes bases do UCI, que seguiam o padrão de base de dados que foram pedidos em cada uma das questões. Foram priorizados bases de dados que continham um número razoável de instâncias e classes (não muito pequenos para ter um resultado significativo, mas não tão grande para não denegrir a performance). Segue a lista das bases com seus links:

PROBLEMA 1 -
Base Iris: https://archive.ics.uci.edu/ml/datasets/Iris
Base Transfusion: https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

PROBLEM 2 -
Base Tic-Tac-Toe Endgame: https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame
Base Congressional Voting: https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

PROBLEM 3 -
Base Acute Inflammations: https://archive.ics.uci.edu/ml/datasets/Credit+Approval
Base Credit Approval: https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations

Para implementação dos algoritmos foi utilizado a lingaugem de programação em python, juntamente com algumas bibliotecas que facilitaram a leitura dos dados, manipulação de matrizes e operações matemáticas.

As bases de dados foram aleatoriamente embaralhadas utilizando a função shuffle da biblioteca numpy. A razão entre o número de instâncias de treinamento em relação ao total foi de 70%, já o número de instâncias para teste foi de 30%. Não foi necessário separar dados para validação.

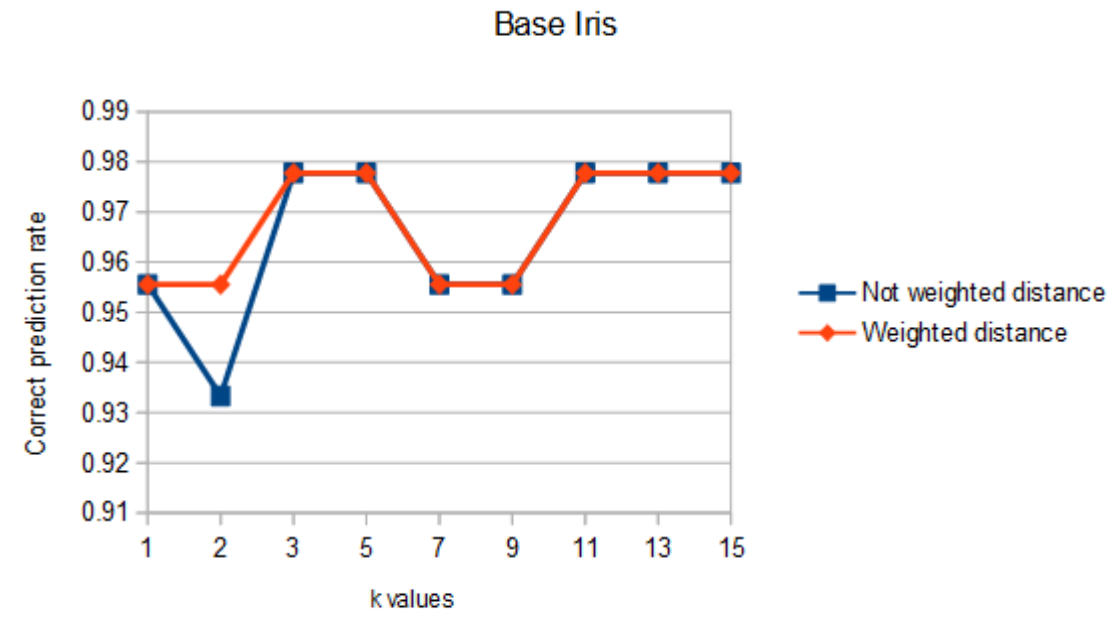
## 2 – Resultados

Todas as questões trataram do algoritmo k-NN, variando apenas a base de dados, tipo de distância, classificador com e sem peso, e variação dos valores de k = {1,2,3,5,7,9,11,13,15}. Assim será exposto de forma agrupada a taxa de acerto de acordo com a variação de valores k, e classificador com e sem peso. Diferentes distãncias e bases de dados serão mostradas em diferentes gráficos e tabelas para facilitar a visualização.

**OBS:** As primeiras linhas das tabelas correspondem as distâncias sem preso, já a segunda linha a distâncias com peso. Tais *labels* foram omitidas para melhor visualização dos valores.

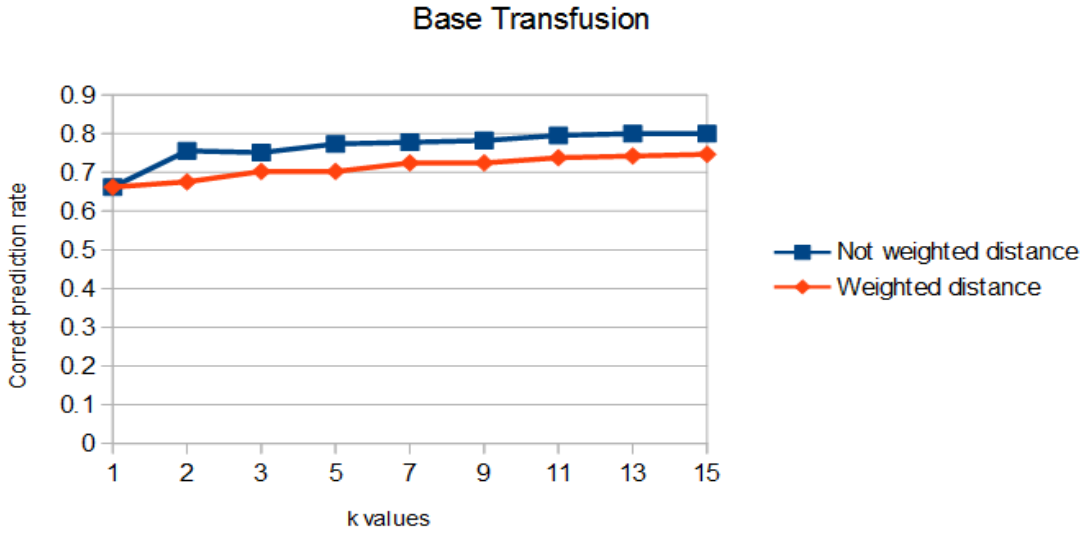
## PROBLEMA 1 – Base Iris

| 1 | 2 | 3 | 5 | 7 | 9 | 11 | 13 | 15 |
|---|---|---|---|---|---|---|---|---|
| 0.9555555556 | 0.9333333333 | 0.9777777778 | 0.9777777778 | 0.9555555556 | 0.9555555556 | 0.9777777778 | 0.9777777778 | 0.9777777778 |
| 0.9555555556 | 0.9555555556 | 0.9777777778 | 0.9777777778 | 0.9555555556 | 0.9555555556 | 0.9777777778 | 0.9777777778 | 0.9777777778 |



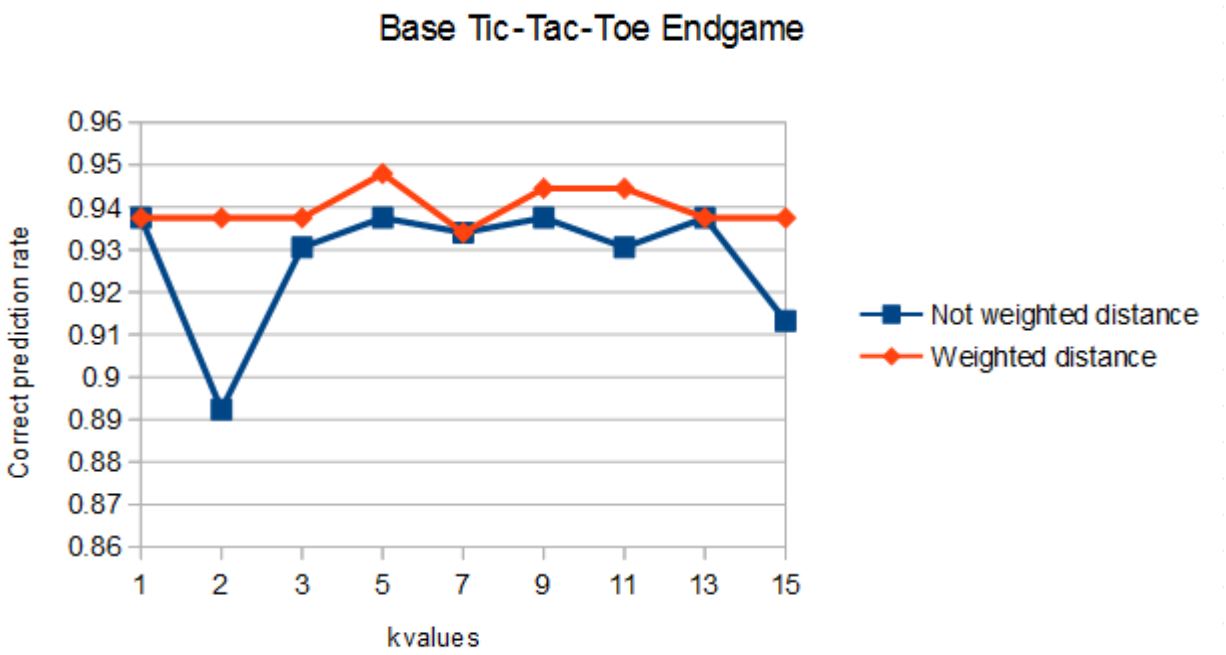

## PROBLEMA 1 – Base Transfusion

| 1 | 2 | 3 | 5 | 7 | 9 | 11 | 13 | 15 |
|---|---|---|---|---|---|---|---|---|
| 0.6622222222 | 0.7555555556 | 0.7511111111 | 0.7733333333 | 0.7777777778 | 0.7822222222 | 0.7955555556 | 0.800000001 | 0.800000001 |
| 0.6622222222 | 0.6755555556 | 0.7022222222 | 0.7022222222 | 0.7244444444 | 0.7244444444 | 0.7377777778 | 0.7422222222 | 0.7466666667 |



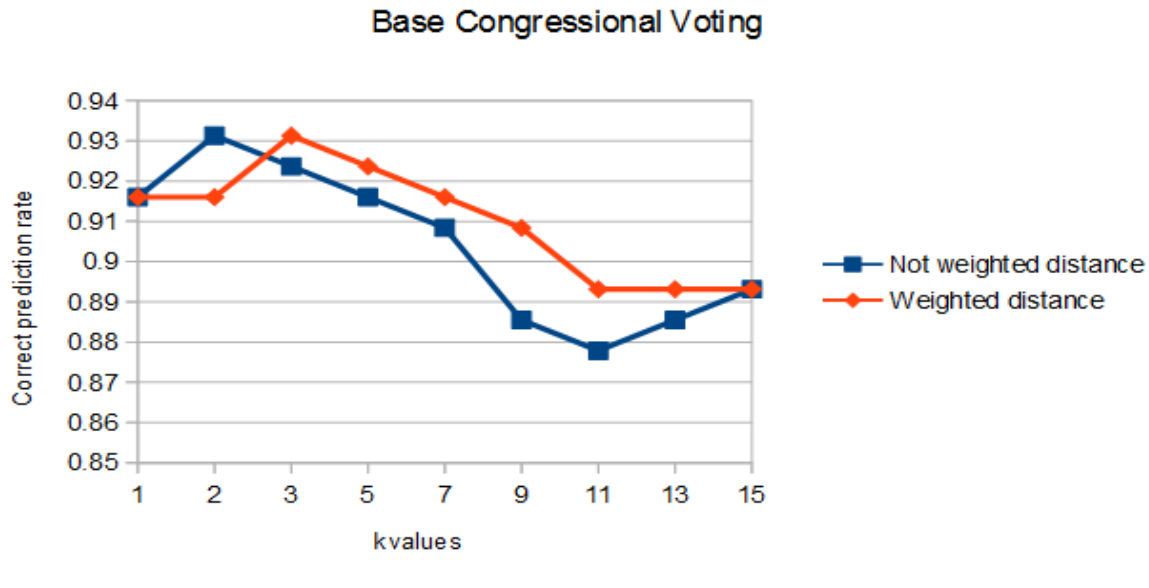
## PROBLEMA 2 - Base Tic-Tac-Toe Endgame

| 1 | 2 | 3 | 5 | 7 | 9 | 11 | 13 | 15 |
|---|---|---|---|---|---|---|---|---|
| 0.9375 | 0.8923611111 | 0.9305555556 | 0.9375 | 0.9340277778 | 0.9375 | 0.9305555556 | 0.9375 | 0.9131944444 |
| 0.9375 | 0.9375 | 0.9375 | 0.9479166667 | 0.9340277778 | 0.9444444444 | 0.9444444444 | 0.9375 | 0.9375 |



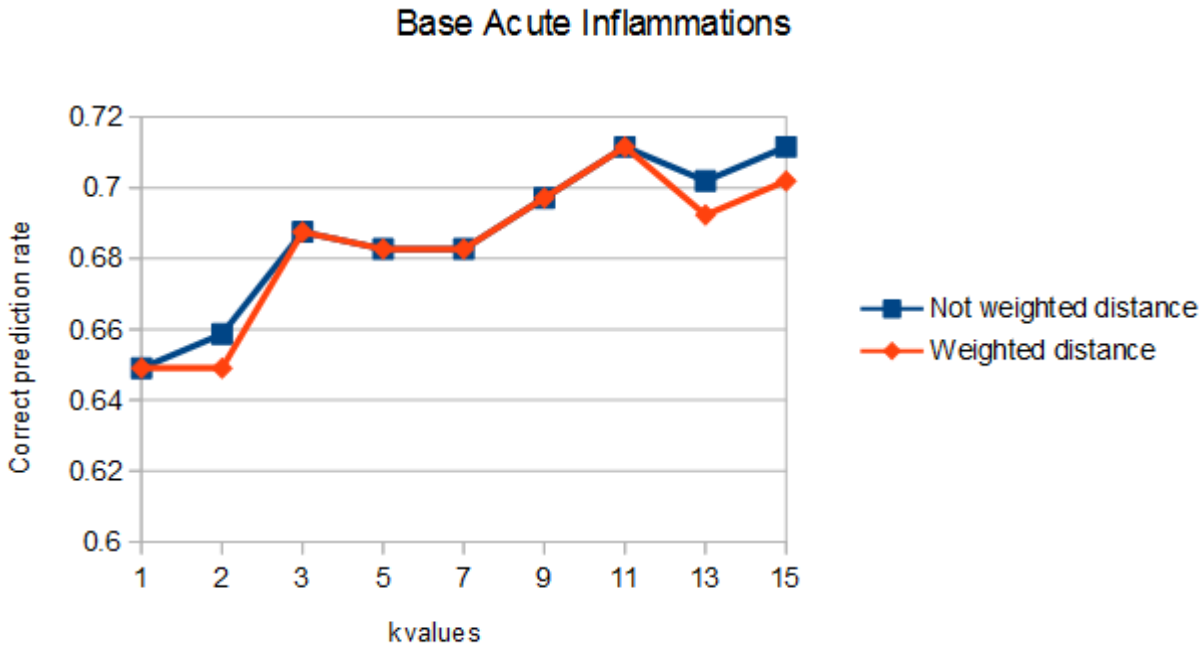

## PROBLEMA 2 - Base Congressional Voting

| 1 | 2 | 3 | 5 | 7 | 9 | 11 | 13 | 15 |
|---|---|---|---|---|---|---|---|---|
| 0.9160305344 | 0.9312977099 | 0.9236641221 | 0.9160305344 | 0.9083969466 | 0.8854961832 | 0.8778625954 | 0.8854961832 | 0.893129771 |
| 0.9160305344 | 0.9160305344 | 0.9312977099 | 0.9236641221 | 0.9160305344 | 0.9083969466 | 0.893129771 | 0.893129771 | 0.893129771 |



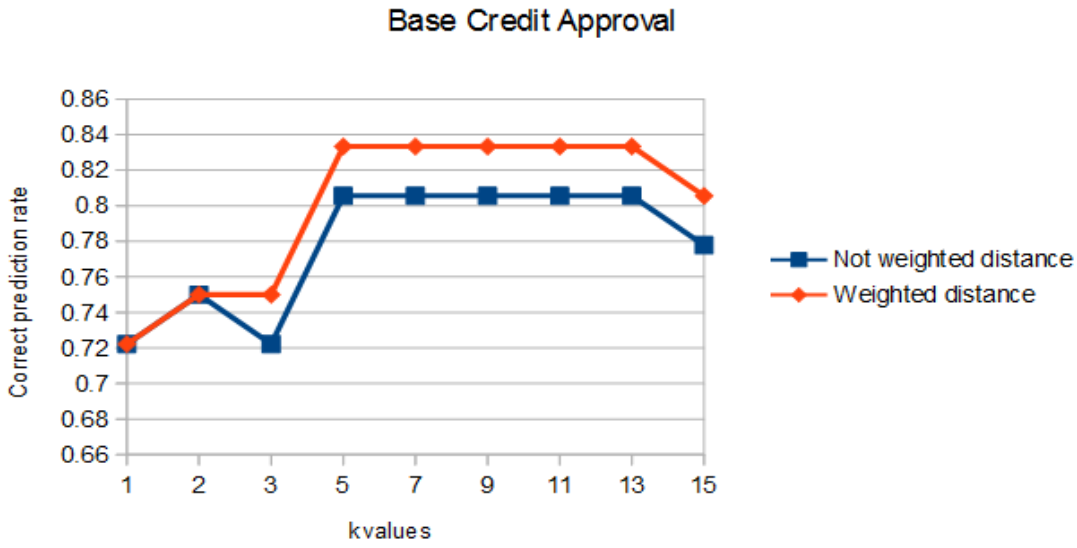
## PROBLEMA 3 - Base Acute Inflammations

| 1 | 2 | 3 | 5 | 7 | 9 | 11 | 13 | 15 |
|---|---|---|---|---|---|---|---|---|
| 0.6490384615 | 0.6586538462 | 0.6875 | 0.6826923077 | 0.6826923077 | 0.6971153846 | 0.7115384615 | 0.7019230769 | 0.7115384615 |
| 0.6490384615 | 0.6490384615 | 0.6875 | 0.6826923077 | 0.6826923077 | 0.6971153846 | 0.7115384615 | 0.6923076923 | 0.7019230769 |




## PROBLEMA 3 - Base Credit Approval

| 1 | 2 | 3 | 5 | 7 | 9 | 11 | 13 | 15 |
|---|---|---|---|---|---|---|---|---|
| 0.7222222222 | 0.75 | 0.7222222222 | 0.8055555556 | 0.8055555556 | 0.8055555556 | 0.8055555556 | 0.8055555556 | 0.7777777778 |
| 0.7222222222 | 0.75 | 0.75 | 0.8333333333 | 0.8333333333 | 0.8333333333 | 0.8333333333 | 0.8333333333 | 0.8055555556 |



## 3 – Análise dos Resultados

É possível observar a partir dos resultados que a tendência é que a taxa de acerto aumente com o aumento do valor k (algo que era esperado), apenas a base *Congressional Voting* não teve esse comportamento.

Quanto a variação do uso de peso para os classificadores, não ficou claro qual é melhor, apesar que é possível notar que para bases de dados com apenas valores numéricos (problema 1) os classificadores sem uso do peso tem resultados tão bons ou melhores quanto os com peso, e o oposto ocorre para dados com valores categóricos (problema 2), onde os cassificadores com peso normalmente tem melhores resultados que os sem peso. Esse padrão se extende também para bases mistas já que a base *Acute Inflammations* tem mais dados numéricos que a base *Credit Approval* e o comportamento foi que *Acute Inflammations* foi melhor sem pesos, e *Credit Approval* foi melhor com pesos.

A variação do tempo de execução se mostrou indiferente a variação do parâmetro k e ao uso do peso, contudo foi bastante sensível ao número de instâncias das bases. Foi possível notar também que para problemas com o cálculo da distância utilizando o Value Difference Metric o tempo de execução aumentou bastante, já que é necessário fazer um processamento das probabilidades envolvidas na fórmula.