~~~
---
title: "10.2 Machine Learning"
author: "Moshe Burnstein"
date: "`r Sys.Date()`"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)

```
# knn
## Binary Classifier

binary_df <- read.csv("binary-classifier-data.csv")
```{r}
binary_df <- read.csv("binary-classifier-data.csv")
head(binary_df)
plot(binary_df)
library(ggplot2)
ggplot(binary_df, aes(x, y)) + geom_point()
str(binary_df)
sqrt(1498)
k = 39
```
```
 The plot shows points strewn over the grid. There does not appear to be a
meaningful correlation. That is why the glm produced only 58% accuracy.
Logistic regression is a linear model. Hence it is a poor fit. There appears
to
be definable clusters, neighbors. That is why knn produces accuracy upwards of
96% on each model. In fact, k=3 produced the greatest accuracy. The scatter-
plot for the trinary classifier data likewise shows a non-linear relationship.
Although there is greater variability in the knn models, they still far
surpass
any linear model. In fact, k=3 still boasts the greatest accuracy.
```

```{r}
library(class)
library(caTools)
split <- sample.split(binary_df, SplitRatio = .8)
train_binary <- subset(binary_df, split == "TRUE")
test_binary <- subset(binary_df, split == "FALSE")
knn_binary3 <- knn(train = train_binary,
                  test = test_binary,
                  cl = train_binary$label,
                  k = 3)
knn_binary3
str(knn_binary3)
plot(knn_binary3)
cm3 <- table(test_binary$label, knn_binary3)
cm3
accuracy_bi3 <- mean(knn_binary3 == test_binary$label)
~~~

~~~
accuracy_bi3
```


```{r}
library(class)
library(caTools)
split <- sample.split(binary_df, SplitRatio = .8)
train_binary <- subset(binary_df, split == "TRUE")
test_binary <- subset(binary_df, split == "FALSE")
knn_binary5 <- knn(train = train_binary,
                  test = test_binary,
                  cl = train_binary$label,
                  k = 5)
knn_binary5
cm5 <- table(test_binary$label, knn_binary5)
cm5
accuracy_bi5 <- mean(knn_binary5 == test_binary$label)
accuracy_bi5


```


```{r}
library(class)
library(caTools)
split <- sample.split(binary_df, SplitRatio = .8)
train_binary <- subset(binary_df, split == "TRUE")
test_binary <- subset(binary_df, split == "FALSE")
knn_binary10 <- knn(train = train_binary,
                  test = test_binary,
                  cl = train_binary$label,
                  k = 10)
knn_binary10
cm10 <- table(test_binary$label, knn_binary10)
cm10
accuracy_bi10 <- mean(knn_binary10 == test_binary$label)
accuracy_bi10
```

```{r}
library(class)
library(caTools)
split <- sample.split(binary_df, SplitRatio = .8)
train_binary <- subset(binary_df, split == "TRUE")
test_binary <- subset(binary_df, split == "FALSE")
knn_binary15 <- knn(train = train_binary,
                  test = test_binary,
                  cl = train_binary$label,
                  k = 15)
knn_binary15
cm15 <- table(test_binary$label, knn_binary15)
cm15
~~~

~~~
accuracy_bi15 <- mean(knn_binary15 == test_binary$label)
accuracy_bi15
```

```{r}
library(class)
library(caTools)
split <- sample.split(binary_df, SplitRatio = .8)
train_binary <- subset(binary_df, split == "TRUE")
test_binary <- subset(binary_df, split == "FALSE")
knn_binary20 <- knn(train = train_binary,
                  test = test_binary,
                  cl = train_binary$label,
                  k = 20)
knn_binary20
cm20 <- table(test_binary$label, knn_binary20)
cm20
accuracy_bi20 <- mean(knn_binary20 == test_binary$label)
accuracy_bi20
```


```{r}
library(class)
library(caTools)
split <- sample.split(binary_df, SplitRatio = .8)
train_binary <- subset(binary_df, split == "TRUE")
test_binary <- subset(binary_df, split == "FALSE")
knn_binary25 <- knn(train = train_binary,
                  test = test_binary,
                  cl = train_binary$label,
                  k = 25)
knn_binary25
cm25 <- table(test_binary$label, knn_binary25)
cm25
accuracy_bi25 <- mean(knn_binary25 == test_binary$label)
accuracy_bi25
```

```{r}
x <- c(3, 5, 10, 15, 20, 25)
y <- c(accuracy_bi3, accuracy_bi5, accuracy_bi10, accuracy_bi15,
accuracy_bi20, accuracy_bi25)
accuracy_bidf <- data.frame(x, y)
library(ggplot2)
accuracy_biplot_binary <- ggplot(accuracy_bidf, aes(x, y)) + geom_point() +
                 geom_text(aes(label = x), hjust = 1, vjust = 2) +
                 labs( x ="k clusters", y = "Accuracy", title = "Accuracy vs.
k Binary")
accuracy_biplot_binary
```
~~~

~~~
## Trinary Classifier
trinary_df <- read.csv("trinary-classifier-data.csv")

```{r}
trinary_df <- read.csv("trinary-classifier-data.csv")
ggplot(trinary_df, aes(x, y)) + geom_point()
str(trinary_df)
trinary_df$label
sqrt(1568)
k = 39
```

```{r}
library(class)
library(caTools)
split <- sample.split(trinary_df, SplitRatio = .8)
train_trinary <- subset(trinary_df, split == "TRUE")
test_trinary <- subset(trinary_df, split == "FALSE")
knn_trinary5 <- knn(train = train_trinary,
                  test = test_trinary,
                  cl = train_trinary$label,
                  k = 5)
knn_trinary5
str(knn_trinary5)
plot(knn_trinary5)
cm5 <- table(test_trinary$label, knn_trinary5)
cm5
accuracy5 <- mean(knn_trinary5 == test_trinary$label)
accuracy5
```

```{r}
library(class)
library(caTools)
split <- sample.split(trinary_df, SplitRatio = .8)
train_trinary <- subset(trinary_df, split == "TRUE")
test_trinary <- subset(trinary_df, split == "FALSE")
knn_trinary10 <- knn(train = train_trinary,
                  test = test_trinary,
                  cl = train_trinary$label,
                  k = 10)
knn_trinary10
str(knn_trinary10)
plot(knn_trinary10)
cm10 <- table(test_trinary$label, knn_trinary10)
cm10
accuracy10 <- mean(knn_trinary10 == test_trinary$label)
accuracy10
```

```{r}
library(class)
library(caTools)
~~~

~~~
split <- sample.split(trinary_df, SplitRatio = .8)
train_trinary <- subset(trinary_df, split == "TRUE")
test_trinary <- subset(trinary_df, split == "FALSE")
knn_trinary15 <- knn(train = train_trinary,
                  test = test_trinary,
                  cl = train_trinary$label,
                  k = 15)
knn_trinary15
str(knn_trinary15)
plot(knn_trinary15)
cm15 <- table(test_trinary$label, knn_trinary15)
cm15
accuracy15 <- mean(knn_trinary15 == test_trinary$label)
accuracy15
```

```{r}
library(class)
library(caTools)
split <- sample.split(trinary_df, SplitRatio = .8)
train_trinary <- subset(trinary_df, split == "TRUE")
test_trinary <- subset(trinary_df, split == "FALSE")
knn_trinary20 <- knn(train = train_trinary,
                  test = test_trinary,
                  cl = train_trinary$label,
                  k = 20)
knn_trinary20
str(knn_trinary20)
plot(knn_trinary20)
cm20 <- table(test_trinary$label, knn_trinary20)
cm20
accuracy20 <- mean(knn_trinary20 == test_trinary$label)
accuracy20
```

```{r}
library(class)
library(caTools)
split <- sample.split(trinary_df, SplitRatio = .8)
train_trinary <- subset(trinary_df, split == "TRUE")
test_trinary <- subset(trinary_df, split == "FALSE")
knn_trinary3 <- knn(train = train_trinary,
                  test = test_trinary,
                  cl = train_trinary$label,
                  k = 3)
knn_trinary3
str(knn_trinary3)
plot(knn_trinary3)
cm3 <- table(test_trinary$label, knn_trinary3)
cm3
accuracy3 <- mean(knn_trinary3 == test_trinary$label)
accuracy3
```
~~~

~~~
```{r}
library(class)
library(caTools)
split <- sample.split(trinary_df, SplitRatio = .8)
train_trinary <- subset(trinary_df, split == "TRUE")
test_trinary <- subset(trinary_df, split == "FALSE")
knn_trinary25 <- knn(train = train_trinary,
                  test = test_trinary,
                  cl = train_trinary$label,
                  k = 25)
knn_trinary25
str(knn_trinary25)
plot(knn_trinary25)
cm25 <- table(test_trinary$label, knn_trinary25)
cm25
accuracy25 <- mean(knn_trinary25 == test_trinary$label)
accuracy25
```



```{r}
x <- c(3, 5, 10, 15, 20, 25)
y <- c(accuracy3, accuracy5, accuracy10, accuracy15, accuracy20, accuracy25)
accuracy_df <- data.frame(x, y)
library(ggplot2)
accuracy_plot_trinary <- ggplot(accuracy_df, aes(x, y)) + geom_point() +
                 geom_text(aes(label = x), hjust = 1, vjust = 2) +
                 labs( x ="k clusters", y = "Accuracy", title = "Accuracy vs.
k Trinary")
accuracy_plot_trinary
```




# k-Means

k_means_cluster <- read.csv("clustering-data.csv")
```{r}
k_means_cluster <- read.csv("clustering-data.csv")
head(k_means_cluster)
dist(k_means_cluster, method = "euclidean")
library(factoextra)
library(NbClust)
fviz_nbclust(k_means_cluster, kmeans, method = "wss") +
             geom_vline(xintercept = 2, linetype = 2) +
             labs(subtitle = "Elbow method")
fviz_nbclust(k_means_cluster, kmeans, method = "wss") +
             geom_vline(xintercept = 6, linetype = 2) +
             labs(subtitle = "Elbow method")

```

```
~~~

~~~
2 clusters appears to be at the elbow...unless one favors the 6 clusters.
There appears to be some gain up until 6. I would try both and use 2 unless
there is significant improvement at 6.

```


```{r}
library(dplyr)
library(ggplot2)
k_means_cluster%>%ggplot(aes(x,y)) + geom_point()
```

```{r}
set.seed(278613)
clustersk2 <- kmeans(x = k_means_cluster, centers = 2)
clustersk2
clustersk2$betweenss
clustersk2$centers
clustersk2$size
clustersk2$totss
clustersk2$iter
clustersk2$ifault
```

```{r}
library(useful)
plot(clustersk2, data = k_means_cluster)
```

```{r}
set.seed(278613)
clustersk3 <- kmeans(x = k_means_cluster, centers = 3)
clustersk3
```

```{r}
library(useful)
plot(clustersk3, data = k_means_cluster)
```

```{r}
set.seed(278613)
clustersk4 <- kmeans(x = k_means_cluster, centers = 4)
clustersk4
library(useful)
plot(clustersk4, data = k_means_cluster)
```

```{r}
set.seed(278613)
clustersk5 <- kmeans(x = k_means_cluster, centers = 5)
clustersk5
library(useful)
plot(clustersk5, data = k_means_cluster)
~~~

~~~
```

```{r}
set.seed(278613)
clustersk6 <- kmeans(x = k_means_cluster, centers = 6)
clustersk6
library(useful)
plot(clustersk6, data = k_means_cluster)
```

```{r}
set.seed(278613)
clustersk7 <- kmeans(x = k_means_cluster, centers = 7)
clustersk7
library(useful)
plot(clustersk7, data = k_means_cluster)
```

```{r}
set.seed(278613)
clustersk8 <- kmeans(x = k_means_cluster, centers = 8)
clustersk8
library(useful)
plot(clustersk8, data = k_means_cluster)
```

```{r}
set.seed(278613)
clustersk9 <- kmeans(x = k_means_cluster, centers = 9)
clustersk9
library(useful)
plot(clustersk9, data = k_means_cluster)
```

```{r}
set.seed(278613)
clustersk10 <- kmeans(x = k_means_cluster, centers = 10)
clustersk10
library(useful)
plot(clustersk10, data = k_means_cluster)
```

```{r}
set.seed(278613)
clustersk11 <- kmeans(x = k_means_cluster, centers = 11)
clustersk11
library(useful)
plot(clustersk11, data = k_means_cluster)
```

```{r}
set.seed(278613)
clustersk12 <- kmeans(x = k_means_cluster, centers = 12)
clustersk12
~~~

~~~
library(useful)
plot(clustersk12, data = k_means_cluster)
```
~~~