R Notebook

# Problem 1

1. No
2. Yes: Classification
3. Yes: Regression
4. No
5. Yes: Regression
6. Yes: Classification (if we are trying to predict if a given user will click)
7. Yes: Anomaly detection
8. No

# Problem 2

1. Continuous, Quantitative, Ratio (Assuming that 4 is twice as bright as 2)
2. Continuous, Quantitative, Ratio
3. Discrete, Qualitative, Ordinal
4. Continuous, Quantitative, Interval (We can consider the 5 hour interval between 10AM and 3PM)
5. Discrete, Qualitative Ordinal

# Problem 3

library(datasets)  
data(iris)  
sprintf("Number of datapoints: %d", nrow(iris))

## [1] "Number of datapoints: 150"

sprintf("Number of features: %d", ncol(iris))

## [1] "Number of features: 5"

sprintf("Number of classes: %d", length(unique(iris$Species)))

## [1] "Number of classes: 3"

summary(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

sprintf("Sepal.Length Std.Dev.: %f", sd(iris$Sepal.Length))

## [1] "Sepal.Length Std.Dev.: 0.828066"

sprintf("Sepal.Width Std.Dev.: %f", sd(iris$Sepal.Width))

## [1] "Sepal.Width Std.Dev.: 0.435866"

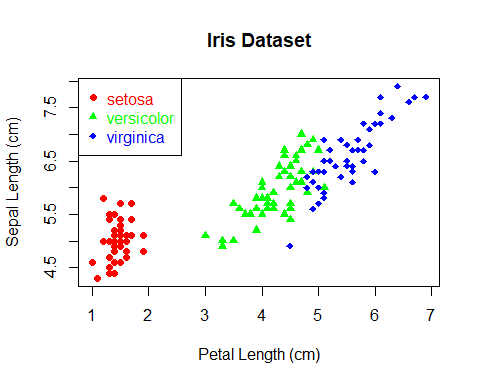
sprintf("Petal.Length Std.Dev.: %f", sd(iris$Petal.Length))

## [1] "Petal.Length Std.Dev.: 1.765298"

sprintf("Petal.Width Std.Dev.: %f", sd(iris$Petal.Width))

## [1] "Petal.Width Std.Dev.: 0.762238"

plot(iris$Sepal.Length ~ iris$Petal.Length,  
 xlab = "Petal Length (cm)",  
 ylab = "Sepal Length (cm)",  
 pch = c(16, 17, 18)[unclass(iris$Species)],   
 main = "Iris Dataset",  
 col = c("red", "green","blue")[unclass(iris$Species)],  
 data = iris)  
  
legend("topleft",   
 legend = c("setosa", "versicolor", "virginica"),   
 text.col = c("red", "green","blue"),  
 col = c("red", "green","blue"),   
 pch = c(16, 17, 18))



# Problem 4

[Here is a graph of d1 and d2](https://www.desmos.com/calculator/zv8fqgvvzt)

Both d1 and d2 take values of 0 when s is 1 and converge to infinity as s approaches 0 (from the right). As such, both can be considered dissimilarity measures on the interval [0, infinity). d2 results in much large dissimilarity values.

# Problem 5

1. If the term only occurs in a single document, then the log term will be maximized, if the term occurs in every document then the log term will be 0
2. Common terms (the, a, because, that, etc.) that are likely to occur in most or all documents will have small values for tfij’. Uncommon words (diabetes, mahalanobis, dimensionality, etc.) will have larger values for tfij’. Words with a high tfij’ might be of greater interest when trying to make predictions about a given document.

# Problem 6

1. Hamming = 4 ; Jaccard = 2/6 = ⅓
2. SMC = 1 - (Hamming Distance) / (Number of Bits)
3. Both Jaccard and Cosine measures ignore 0-0 matches. For our example with binary vectors, the numerator for both measures is the number of 1-1 matches.

# Problem 7

Cosine =

Correlation =

Euclidian =

1. Cosine = 0 ; Correlation = 0 ; Euclidian = 2
2. Cosine = 0 ; Correlation = -1 ; Euclidian = 2 ; Jaccard = 0