Author: Bin Dong

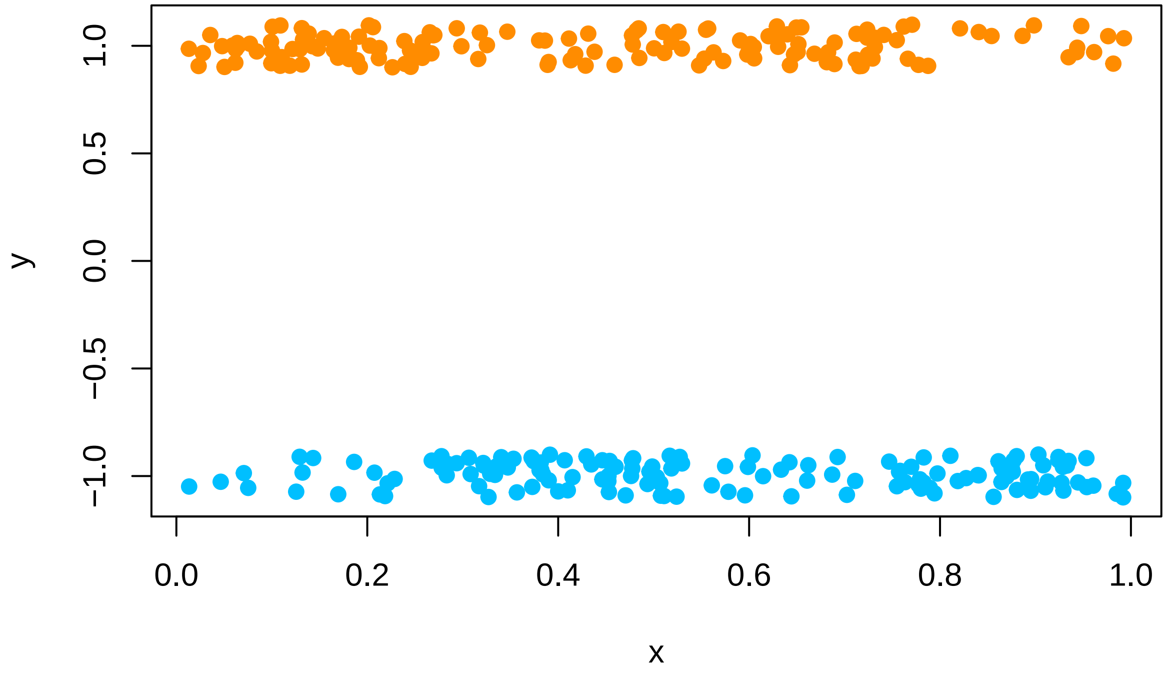
Email: bindong2@illinois.edu

Question 1

```
rm(list=ls())

set.seed(1)

n = 300
x = runif(n)
py <- function(x) sin(4*pi*x)/3 + 0.5
y = (rbinom(n, 1, py(x))-0.5)*2
plot(x, y + 0.1*runif(n, -1, 1), ylim = c(-1.1, 1.1), pch = 19,
col = ifelse(y == 1, "darkorange", "deepskyblue"), ylab = "y")
```



```
testx = seq(0, 1, length.out = 1000)
testy = (rbinom(1000, 1, py(testx))-0.5)*2
```

```
get_p_hat <- function(x, y, weight)
{
  return <- sum(weight * (y+1)/2)/sum(weight)
}

get_gini <- function(p)
{
  return <- p*(1-p)
}

split_data <- function(x, y, weight, value)
{
  dataframe <- data.frame("x"=x, "y"=y, "weight"=weight)
  data.left <- subset(dataframe, x<value)
  data.right <- subset(dataframe, x>=value)
  gini.left <- get_gini(get_p_hat(data.left$x, data.left$y, data.left$weight))
  gini.right <- get_gini(get_p_hat(data.right$x, data.right$y, data.right$weight))
  score <- -sum(data.left$weight)*gini.left-sum(data.right$weight)*gini.right
  score <- score/sum(weight)

  left <- if(sum(data.left$y)>0) 1 else -1
  right <- if(sum(data.right$y)>0) 1 else -1

  return <- list(score=score,left=left, right=right)
}
```

A test of Stump model

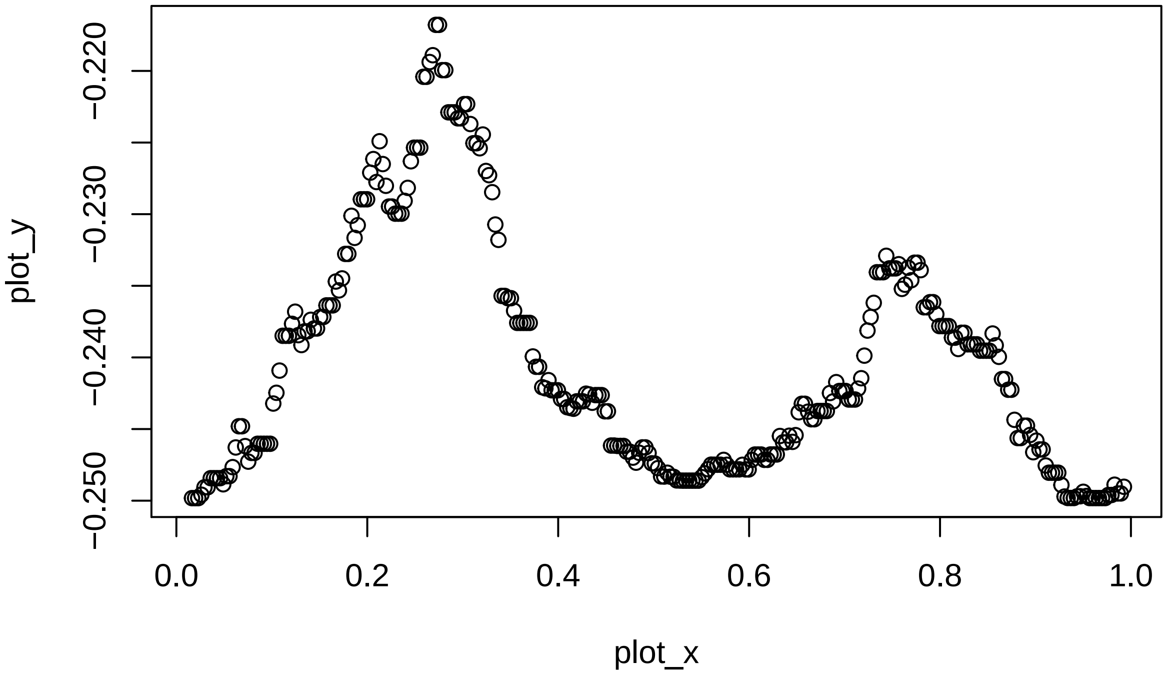
```
max_score <- (split_data(x, y, rep(1/n,n), 0.5))$score
split_criteria <- 0

plot_x <- seq(min(x), max(x), length=n)

plot_y <- rep(0, n)

for(i in c(1:n))
{
  r <- split_data(x, y, rep(1/n,n), plot_x[i])
  plot_y[i] <- r$score
  if(!is.nan(plot_y[i]) && plot_y[i] >= max_score)
  {
    split_criteria <- plot_x[i]
    max_score <- plot_y[i]
  }
}

plot(plot_x , plot_y)
```



```
result <- split_data(x, y, rep(1/n,n), split_criteria)
print(split_criteria)
```

```
## [1] 0.2751796
```

```
print(result)
```

```
## $score
## [1] -0.2167849
##
## $left
## [1] 1
##
## $right
## [1] -1
```