Applied Data Mining Homework 3

Xun Zhao, xz2827

1. Classifier g_m

```
train_g = function(m, data){
    x.pos.s = as.matrix(data[data[ncol(data)] == 1, -ncol(data)])
    x.neg.s = as.matrix(data[data[ncol(data)] != 1, -ncol(data)])
    pos.index = sample(1:nrow(x.pos.s), m, replace = TRUE)
    neg.index = sample(1:nrow(x.neg.s), m, replace = TRUE)
    x.pos.m = x.pos.s[pos.index,]
    x.neg.m = x.neg.s[neg.index,]
    if(m == 1){
        x.pos.m = t(as.matrix(x.pos.m))
        x.neg.m = t(as.matrix(x.neg.m))
    }
    w.m = (x.pos.m - x.neg.m) / rowSums((x.pos.m - x.neg.m) ^ 2)
    c.w.m = rowSums(w.m * (x.pos.m + x.neg.m) / 2)
    misc =
        rowSums(w.m %*% t(x.pos.s) < c.w.m) +
        rowSums(w.m %*% t(x.neg.s) > c.w.m)
    v = w.m * ifelse(misc > nrow(data) / 2, -1, 1)
    c = rowSums(v * (x.pos.m + x.neg.m) / 2)
    return(list(v, c))
}
```

Classifier Function

`V` and `c` is the result of function g_m.

```
classify = function(x, V, c){
    return(sign(colSums(sign(V %*% t(x) - c))))
}
```

Load Data from uspsdata

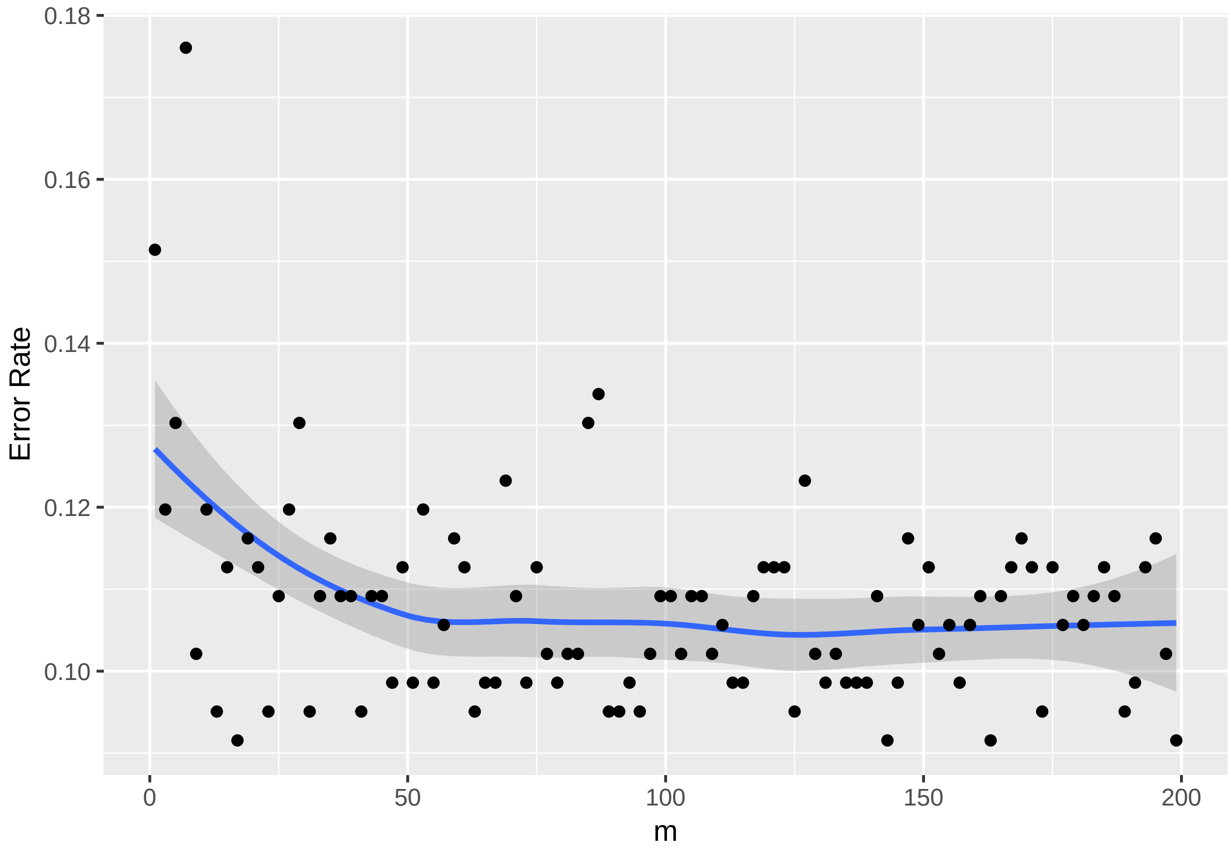
Shuffle the data and separate it into two parts with equal size.

```
data.d = read.csv('../Data/wdbc.data')
data.l = read.csv('../Data/wdbc.labels')
data = data.frame(data.d, data.l)
```

```r
data = data[sample(1:nrow(data)),]
data.tr = data[1:round(nrow(data) / 2),]
data.ts = data[(round(nrow(data) / 2) + 1):nrow(data),]
```

Train with Different m

```r
m.seq = seq(1, 199, 2)
err.rates = sapply(
    m.seq,
    function(m){
        Vc = train_g(m, data.tr)
        V = Vc[[1]]
        c = Vc[[2]]
        esti = classify(data.ts[,-ncol(data.ts)], V, c)
        err.rate = mean(esti != data.ts[,ncol(data.ts)])
        return(err.rate)
    }
)
graph = ggplot(data.frame(cbind(m.seq, err.rates)),
    aes(x = m.seq, y = err.rates)) +
    geom_smooth(formula = y ~ x, method = 'loess') +
    geom_point() +
    xlab('m') +
    ylab('Error Rate')
plot(graph)
```



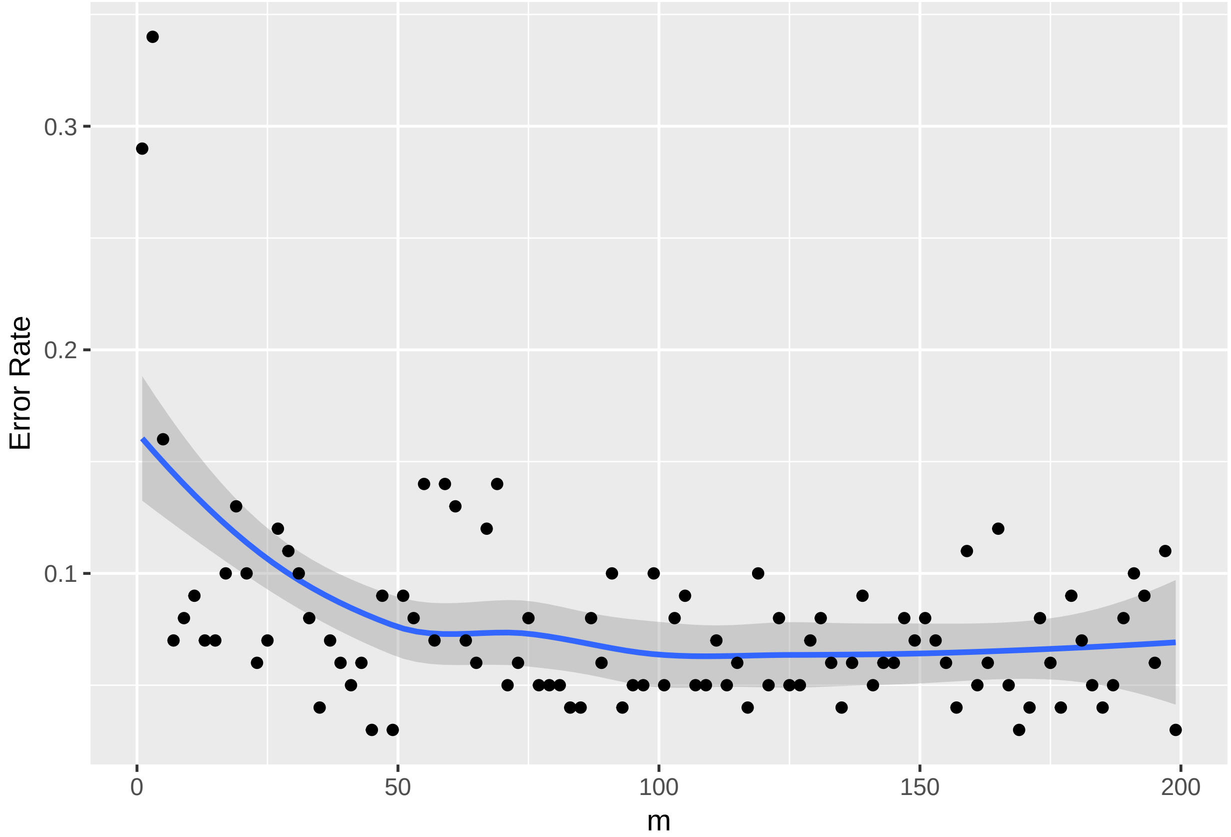
Load Data from uspsdata

Shuffle the data and separate it into two parts with equal size.

```
data.d = read.table('../Data/uspsdata.txt')
data.l = read.table('../Data/uspscl.txt')
data = data.frame(data.d, data.l)
data = data[sample(1:nrow(data)),]
data.tr = data[1:round(nrow(data) / 2),]
data.ts = data[(round(nrow(data) / 2) + 1):nrow(data),]
```

Train with Different m

```
m.seq = seq(1, 199, 2)
err.rates = sapply(
    m.seq,
    function(m){
        Vc = train_g(m, data.tr)
        V = Vc[[1]]
        c = Vc[[2]]
        esti = classify(data.ts[,-ncol(data.ts)], V, c)
        err.rate = mean(esti != data.ts[,ncol(data.ts)])
        return(err.rate)
    }
)
graph = ggplot(data.frame(cbind(m.seq, err.rates)),
    aes(x = m.seq, y = err.rates)) +
    geom_smooth(formula = y ~ x, method = 'loess') +
    geom_point() +
    xlab('m') +
    ylab('Error Rate')
plot(graph)
```


Error Rate
m