Problem set 4, ORF525

Bachir El khadir

<2016-03-11 Fri>

1 Q1

1.1)

a)

Some helper functions

```
library(png)
library(kernlab)
library(ggplot2)
library(glmnet)
source("functions.R")

crop <- function(img) crop.r(img, 160, 96)
take.grad <- function(img) grad(img, 128, 64, F)
take.hog <- function(grad.img) hog(grad.img$xgrad, grad.img$ygrad, 4, 4, 6)

plt.grad <- function(grad.img, h=128, w=64, ...) {
    plot(c(),c(), asp=1, xlim=c(0,70), ylim=c(0,130), xlab="X", ylab="Y", ...)
    for (i in 1:h){
        for (j in 1:w){
            arrows(x0=j, y0=h+1-i, x1=j+grad.img$xgrad[i,j]*5, y1=h-i+1+grad.img$ygrad[i,j]*5, length=
        }
    }
}

plt.gray <- function(img.gray, ...) image(t(img.gray)[, nrow(img.gray):1], col  = gray((0:32)/32), ...


load.from.directory <- function(dir) {
    images = list()
    img <- sample(list.files(dir), size=1)
    return(readPNG(file.path(dir, img)))
}
```

Load images, convert to gray, crop if necessary, and then calculate the gradient / hod

```
image.pos <- load.from.directory("pngdata/pos")
image.neg.uncropped <- load.from.directory("pngdata/neg")
image.neg.gray.uncropped <- rgb2gray(image.neg.uncropped)
image.pos.gray <- rgb2gray(image.pos)
image.neg.gray <- crop(image.neg.gray.uncropped)
```

```
grad.pos <- take.grad(image.pos.gray)
grad.neg <- take.grad(image.neg.gray)
hog.pos <- take.hog(grad.pos)
hog.neg <- take.hog(grad.neg)
0
```

And then plot

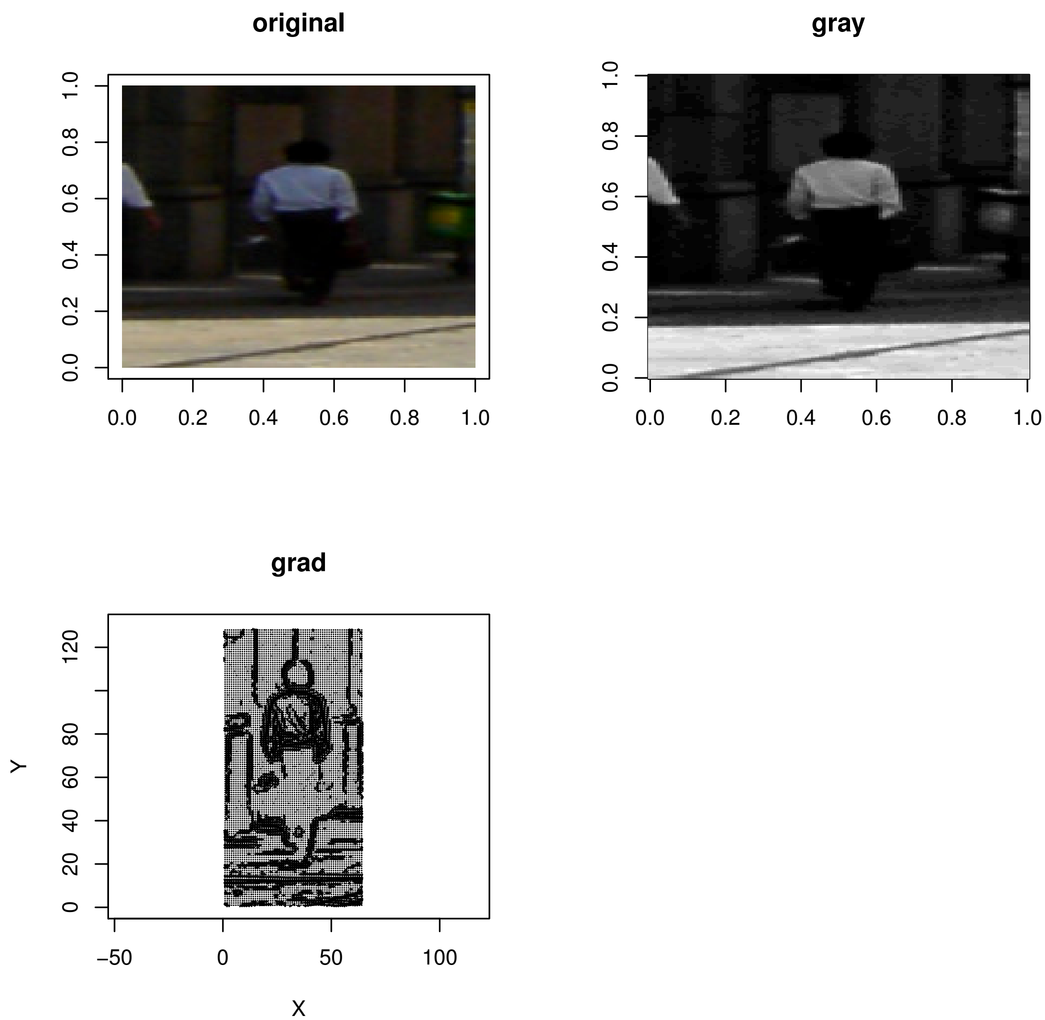

Figure 1: Pos image

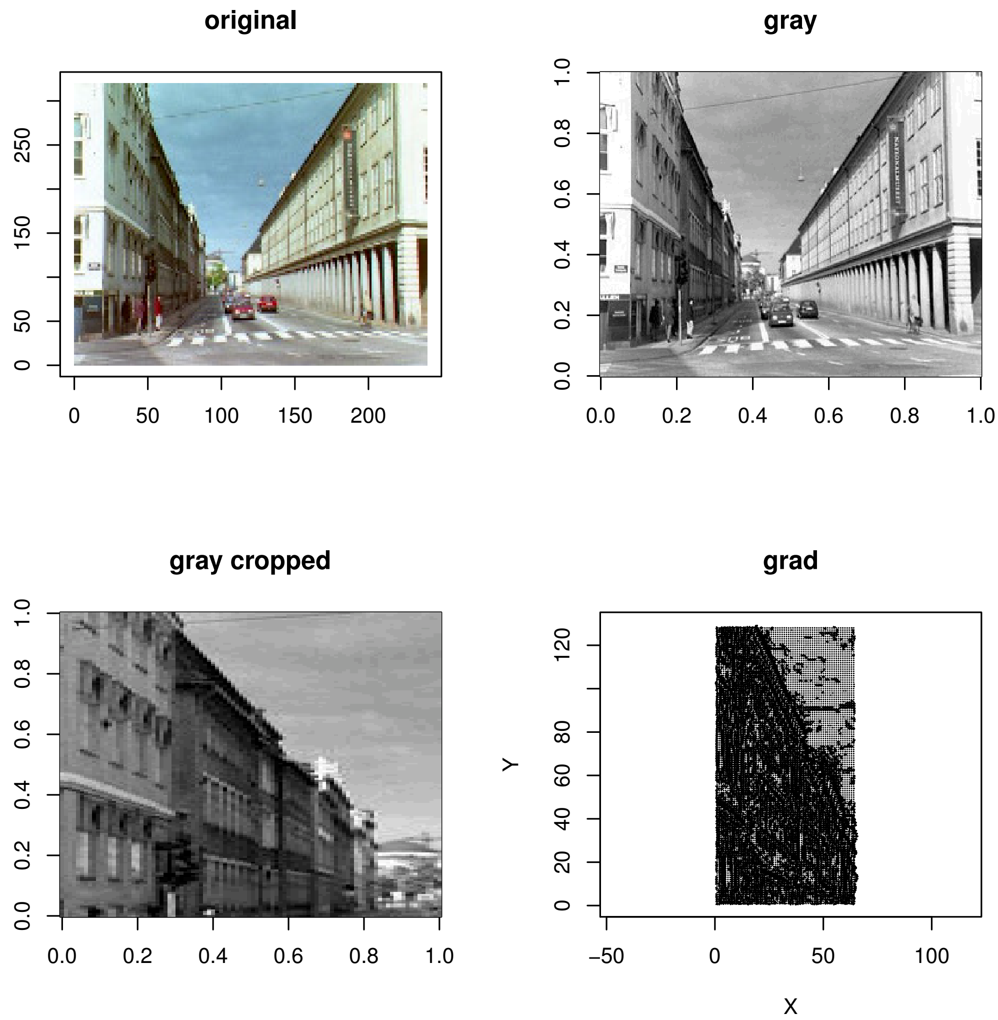

Figure 2: Negative image

b)

Prepare the dataset

```r
# load all images from directory
load.all.directory <- function(dir) {
    images = list()
    for(img in list.files(dir)) {
        images[[img]] <- readPNG(file.path(dir, img))
    }
    return(images)
}

# extract features
feature.pos.img <- function(img) c(1, take.hog(take.grad(rgb2gray(img))))
feature.neg.img <- function(img) c(0, take.hog(take.grad(crop(rgb2gray(img)))))

pos.images <- load.all.directory("pngdata/pos")
neg.images <- load.all.directory("pngdata/neg")
data <- c(
    unname(lapply(pos.images, feature.pos.img)),
```

```
    unname(lapply(neg.images, feature.neg.img))
)
data <- sapply(data, identity)

# construct data frame
df <- data.frame(t(data))
colnames(df) <- c("label", paste("F",1:96, sep='_'))
df[1:3, 1:5]
```

1.2)

1)

$\log(C)$ take values in a uniform grid of 100 points of $[-4, 2]$. For each value, we evaluate the cross validation error of the corresponding SVM and we plot the result.

```
# SVM
logspace <- function(s, e, n=100) 10^((1:n-1) / n * (e-s) + s)
C <- logspace(-4, 2, 10)
formula <- as.formula(paste("label", paste(colnames(df)[-1], collapse='+'), sep='~'))
cross.error <- sapply(C, function(c) {ksvm(formula, df, cross=10, C=c)@cross})
C.best <- C[which.min(cross.error)]
paste("best C", C.best)
```

Best $C \approx 1.58$

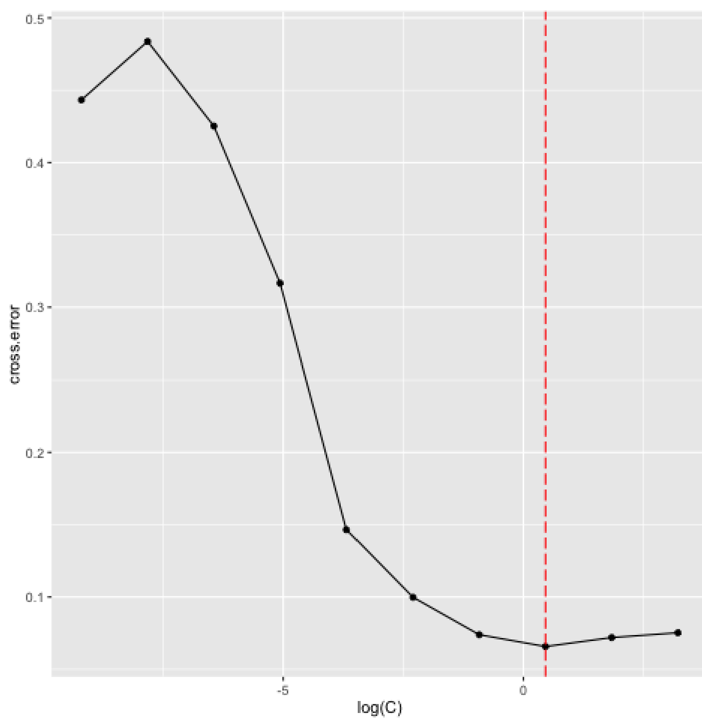

Figure 3: SVM cross validation error

2) Now we use glmnet

```
x <- t(data[2:nrow(data),])
y <- data[1, ]
logit.model <- glmnet(x, y, family="binomial")
cvlogit.model <- cv.glmnet(x, y, family = "binomial", type.measure="class")
```

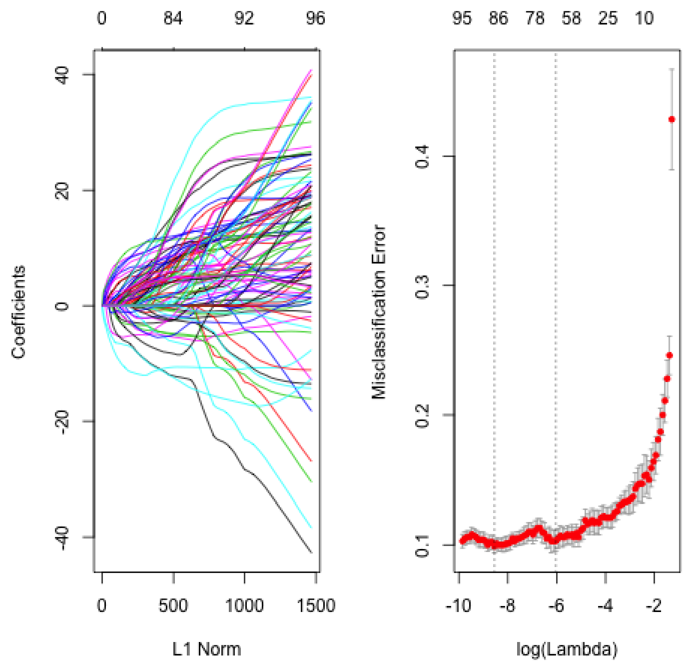

Figure 4: Logit error

3) Compare

Table 1: Cross validation classification error

SVM	Logit 1st Lambda	Logit min Lambda
0.066	0.103	0.099

The errors are of the same order of magnitude.

2 Q2

(a)

$p(x) = p(x|Y=1)p(Y=1) + p(x|Y=-1)p(Y=-1) = \frac{1}{3}\frac{1_{[-5,10]}}{15} + \frac{2}{3}\frac{1_{[-10,5]}}{15}$

$$p(y|x) = \frac{p(x|y)}{p(x)}p(y) \equiv \begin{cases} p(Y=1)p(x|Y=1) & \text{if } y=1 \\ p(Y=-1)p(x|Y=-1) & \text{if } y=-1 \end{cases}$$

The bayes classifier $B(x) := \arg\max_{y\in\{0,1\}} p(y|x)$

$$B(x) = 1 \iff p(Y=1)p(x|Y=1) \geq p(Y=-1)p(x|Y=-1) \iff 1_{[-5,10]}(x) \geq 2 \times 1_{[-10,5]}(x) \iff x \in (5,10)$$

$$B(x) = \begin{cases} 1 & \text{if } x \in (5,10) \\ -1 & \text{o.w} \end{cases}$$

Bayres Risk $R(B) = E[1_{B(X)\neq Y}] = P(Y=1, X\in(-5,5)) = P(X\in(-5,5)|Y=1)P(Y=1) = \frac{2}{3}\times\frac{1}{3} = \frac{2}{9}$

(b)

$R(h) = E[1_{h(X)\neq Y}] = P(sign(\alpha+\beta X^2) < 0|Y=1)P(Y=1) + P(sign(\alpha+\beta X^2) > 0|Y-=1)p(Y=-1) = \frac{1}{3}\left(P_{U\sim\mathcal{U}([-5,10])}(sign(\alpha+\beta U^2) < 0) + 2P_{U\sim\mathcal{U}([-10,5])}(sign(\alpha+\beta U^2) > 0)\right)$

If α and β have the same signs, then $\alpha+\beta X^2$ keeps a constant sign. If not, then $\alpha+\beta X^2$ has two roots $\pm\sqrt{\frac{-\alpha}{\beta}}$, and has the sign of α only between them. Let $r = \sqrt{\frac{-\alpha}{\beta}}$ Cases:

- $\alpha = 0, \beta = 0$??
- $\alpha \geq 0, \beta > 0$ or $\alpha > 0, \beta \geq 0$, $sign(\alpha + \beta X^2) = 1$, $R(h) = \frac{1}{3}$
- $\alpha \leq 0, \beta < 0$ or $\alpha < 0, \beta \leq 0$, $sign(\alpha + \beta X^2) = -1$, $R(h) = \frac{2}{3}$
- $\alpha < 0, \beta > 0$, $sign(\alpha + \beta X^2) = 2 \times 1_{x \in (\pm\sqrt{\frac{-\alpha}{\beta}})} - 1$: $R(h) = \frac{1}{3}\frac{1}{15}\left((10 \wedge r) + (5 \wedge r) + 2((5-r)^+ + (10-r)^+)\right)$

$$R(h) = \frac{1}{45}\begin{cases} 15 & \text{if} \quad r \geq 10 \\ r + 5 + 2(10 - r) = 25 - r & \text{if} \quad 5 < r < 10 \\ 2r + 2(5 - r + 10 - r) = 30 - 2r & \text{if} \quad r \leq 5 \end{cases}$$

- $\alpha > 0, \beta < 0$, can be deduced from the last question because $sign(\alpha + \beta x^2) = -sign(-\alpha - \beta x^2)$

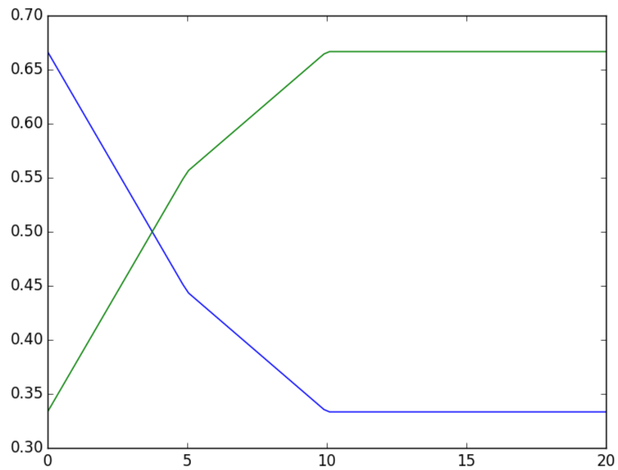

Figure 5: Bayess Error

One possible solution is $\alpha = -1$, $\beta = 0$, and the risk is $R(h) = \frac{1}{3}$
(c)

$$\begin{aligned} R_\Phi(\beta) &= E[(1 - Y\beta X)^+] = E[(1 - \beta U_1)^+]p(Y = 1) + E[(1 + \beta U_2)^+]p(Y = -1) \\ &= \frac{1}{3}\int_0^1 (1 - \beta(15u - 5))^+ + 2(1 + \beta(15u - 10))^+ du \\ &= \frac{1}{3}\int_0^1 (1 - 15\beta(u - \frac{1}{3}))^+ + 2(1 + 15\beta(u - \frac{2}{3}))^+ du \end{aligned}$$

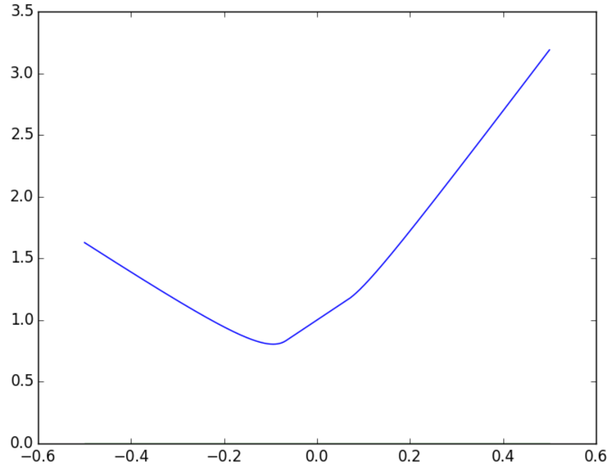

Figure 6: Hinge Error

3 Q3

-

3.1. $f(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{\frac{1}{2}x'\Sigma^{-1}x}$

$$p(y|x) \equiv p(Y=y)p(X=x|Y=x) = \begin{cases} pf(x-\mu_1) & \text{if } y=1 \\ (1-p)f(x-\mu 2) & \text{if } y=-1 \end{cases}$$

bayes estimator:

$$\begin{aligned} B(x)=1 \iff \frac{f(x-\mu_1)}{f(x-\mu_2)} \geq \frac{1-p}{p} &\iff (x-\mu_1)'\Sigma^{-1}(x-\mu_1) - (x-\mu_2)'\Sigma^{-1}(x-\mu_2) \geq \log\frac{1-p}{p} \\ &\iff x\,\underbrace{2\Sigma^{-1}(\mu_2-\mu_1)}_{\omega} \geq \underbrace{\log\frac{1-p}{p} + \mu_2'\Sigma^{-1}\mu_2 - \mu_1'\Sigma^{-1}\mu_1}_{-b} \\ &\iff sign(x.w+b) = 1 \end{aligned}$$

MLE (see ORF524): Write the density:

- MLE for Bernouilli variable: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i=1}$
- MLE for the mean of gaussian: $\hat{\mu}_j = \frac{1}{n_j}\sum_{(Y_i,X_i)\in D_j} X_i$ where $j=1,2$
- Write the density, derive the loglikelihood and take the derivative w.r.t Σ:

$$\begin{aligned} f(x_1,x_2,...,x_n|\mu_1,\mu_2,\Sigma) &= f(D_1|\mu_1,\Sigma)f(D_2|\mu_2,\Sigma) \\ &= \prod_{x_i\in D_1} f(x_i|\mu_1,\Sigma) \prod_{x_i\in D_2} f(x_i|\mu_2,\Sigma) \end{aligned}$$

$$
\begin{aligned}
\hat{\Sigma} &= \frac{1}{n}\left[\sum_{(Y_i,X_i)\in D_1}(x_i-\hat{\mu}_1)(x_i-\hat{\mu}_1)^T + \sum_{(Y_i,X_i)\in D_2}(x_i-\hat{\mu}_2)(x_i-\hat{\mu}_2)^T\right] \\
&= \frac{1}{n}\left[\sum_{(Y_i,X_i)\in D_1} x_i x_i^T - \hat{\mu}_1\hat{\mu}_1^T + \sum_{(Y_i,X_i)\in D_2} x_i x_i^T - \hat{\mu}_2\hat{\mu}_2^T\right] \\
&= \frac{1}{n}\sum_i x_i x_i^T - \frac{n_1}{n}\hat{\mu}_1\hat{\mu}_1^T - \frac{n_2}{n}\hat{\mu}_2\hat{\mu}_2^T
\end{aligned}
$$

Let $\hat{\omega} := 2\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1), \hat{b} = \log\frac{1-\hat{p}}{\hat{p}} + \hat{\mu}_2'\hat{\Sigma}^{-1}\hat{\mu}_2 - \hat{\mu}_1'\hat{\Sigma}^{-1}\hat{\mu}_1$, then by plugging the precedent values we can see that the classifier can be expressed as $sign(\hat{\omega}.x + \hat{b})$.

3.2 The function of (β_0, β) is convex. First order condition gives:

- With respect to β_0: $\sum_i(Y_i - \beta_0 - X_i^T\beta) = 0 \implies \beta_0 = \frac{1}{n}\underbrace{\sum_i Y_i}_{0} - \frac{1}{n}\sum_i X_i^T\beta = -\underbrace{\frac{1}{n}(n_1\mu_1 + n_2\mu_2)'}_{\mu}\beta$

- With respect to β:

$$
\begin{aligned}
0 &= \sum_i (Y_i - \beta_0 - X_i^T\beta)X_i \\
&= \sum_i (Y_i + (\hat{\mu} - X_i)^T\beta)X_i \\
&\implies \sum_i Y_i X_i = \sum_i -X_i(\hat{\mu} - X_i)^T\beta \\
&\implies n(\hat{\mu}_2 - \hat{\mu}_1) = \left(-n\hat{\mu}\hat{\mu}^T + \sum_i X_i X_i^T\right)\beta \\
&\implies n(\hat{\mu}_2 - \hat{\mu}_1) = \underbrace{\left(-n\hat{\mu}\hat{\mu}^T + \sum_i X_i X_i^T\right)}_{n\hat{\Sigma}'}\beta
\end{aligned}
$$

But

$$
\begin{aligned}
\hat{\Sigma}' &= \frac{1}{n}\sum_i X_i X_i^T - \hat{\mu}\hat{\mu}^T \\
&= \hat{\Sigma} + \frac{n_1}{n}\hat{\mu}_1\hat{\mu}_1^T + \frac{n_2}{n}\hat{\mu}_2\hat{\mu}_2^T - \frac{n_1^2}{n}\hat{\mu}_1\hat{\mu}_1^T - \frac{n_2^2}{n}\hat{\mu}_2\hat{\mu}_2^T - \frac{n_1 n_2}{n}(\hat{\mu}_1\hat{\mu}_2^T + \hat{\mu}_2\hat{\mu}_1^T) \\
&= \hat{\Sigma} + \frac{n_1 n_2}{n}(\hat{\mu}_2 - \hat{\mu}_1)u'
\end{aligned}
$$

so that $\Sigma'\beta = \hat{\Sigma}\beta + \frac{n_1 n_2}{n}(\beta' u)(\hat{\mu}_2 - \hat{\mu}_1) = n(\hat{\mu}_2 - \hat{\mu}_1)$, eg $\beta \equiv \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) \equiv \hat{w}$

So $\hat{\beta} \equiv \hat{w}$

3.3 An example where LDA fails but the data is linearly separable:

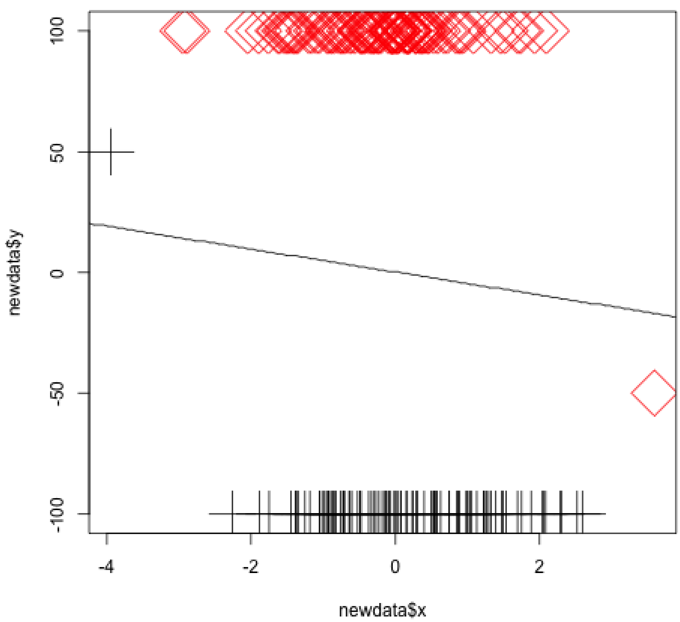

Figure 7: Fail LDA

4 Q4

4.1. Let $y_1, \ldots, y_n$ be any labeling, and let $w = \gamma(y_1, \ldots, y_n)$, then: $y_i(w, e_i) = y_i^2 \gamma = \gamma$
Let $D_0 = \{e_1, \ldots, e_{\frac{d}{2}}\}$, $D_1 = \{e_{\frac{d}{2}+1}, \ldots, e_d\}$ Take s samples from D_0 and s sample from D_1

- Label all points in D_0 by 1 except the one in the sample
- Label all points in D_1 by -1 except the one in the sample

Take w to be the seperating vector. Then w classifies correctly the $2s$ points in the sample but missclassifies the rest of the points $d - 2s \geq \frac{d}{3}$ 4.2. Let j be the first index for which $x_j = 1$ $\sum_{i=1}^{d} \frac{(-1)^{i-1}}{2^{i-1}} x_i = \frac{(-1)^{j-1}}{2^{j-1}}(1 + \underbrace{\sum_{i=1}^{d-j} \frac{(-1)^i}{2^i} x_{i+j}}_{<1})$

So this quantity has the the same sign as $(-1)^{j-1}$, which is what we want.

4.2 Suppose we could have another linear seperator with $(a_1, \ldots, a_d)$ s.t $\sum_i a_i^2 = 1$ with margin $\frac{1}{f(d)}$, then Take $e_i = (0, \ldots, \underbrace{1}_{i}, \ldots, 0)$ $\sum_{i=1}^{d} a_i e_i = a_j$ should have the same sign as $(-1)^{j-1}$

- For all x in the d -cube, $a_{2j} + \sum_{i=2j+1}^{d} a_i x_i \geq \frac{1}{f(d)} \implies |a_{2j}| \geq \sum_{i=j+1}^{d} |a_{2i}| + \frac{1}{f(d)}$
- Using the same argument: $|a_{2j+1}| \geq \sum_{i=j+1}^{d} |a_{2i+1}| + \frac{1}{f(d)}$

  Consider the sequence u_j defined by: $u_{d/2} = a_d, u_j = \sum_{i=j+1}^{d/2} u_i + \frac{1}{f(d)}$

By induction we can easily see that:

- $u_j \leq |a_{2j-1}|$ and
- $u_1 = 2^{d/2-1} u_d + (2^{d/2-1} - 1)\frac{1}{f(d)} \geq (2^{d/2-1} - 1)\frac{1}{f(d)}$

Since $|a_1| \leq 1$: $1 \geq a_1 \geq u_1 \geq (2^{n/2-1} - 1)\frac{1}{f(d)}$ Therefore $f(d) \geq 2^{d/2-1} - 1$
So $f(d)$ cannot be bounded by a polynomial from above.