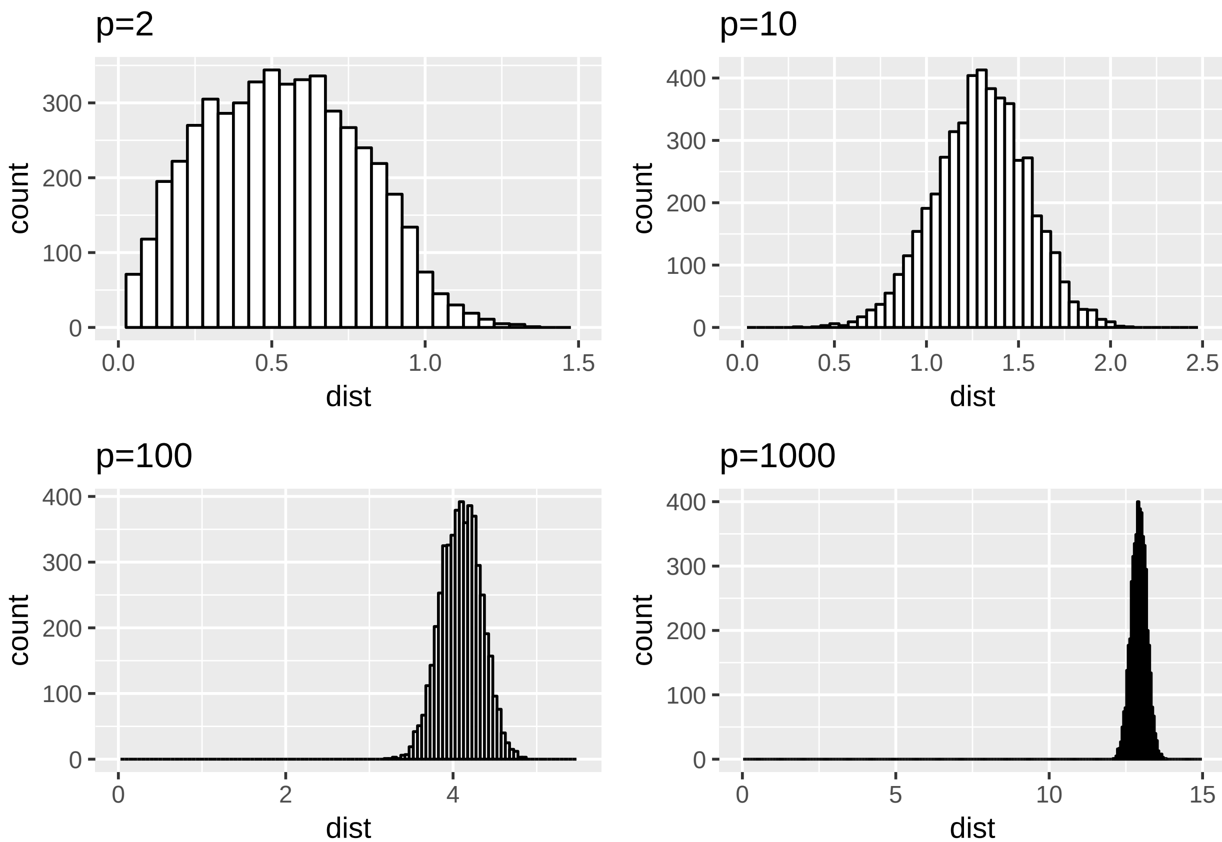
Homework 1

Xingche Guo and Yueying Wang

2/17/2019

1.6



It can be noticed from the histograms that when p becomes larger, the average distance between two points also increases. Meanwhile, when p is really large, the pairwise distances concentrate around a certian value and it will be hard to distinguish the nearest points or 'neighbors' given certian x. The 'local' prediction methods won't work well.

1.7

A lower bound for sample size n is

$$n \geq (V_p(\varepsilon))^{-1} = \left(\frac{1}{\pi\varepsilon^2}\right)^{p/2} \Gamma(\frac{p}{2} + 1)$$

When $p \to \infty$, since Giraud notes that $\frac{V_p(r)}{\left(\frac{2\pi e r^2}{p}\right)^{p/2} (p\pi)^{-1/2}} \to 1$, an approximate lower bound for n is

$$n \geq \left(\frac{p}{2\pi e\varepsilon^2}\right)^{p/2} (p\pi)^{1/2}$$

Plug in $p = 20, 50, 200$ and $\varepsilon = 1, 0.1, 0.01$, we have

```
##                  1          0.1          0.01
## 20   3.874934e+01 3.874934e+21  3.874934e+41
## 50   5.779614e+12 5.779614e+62 5.779614e+112
## 200 1.798939e+108          Inf           Inf
```

2.2

Proof:

$$L(y, \hat{y}) = \left[\ln\left(\frac{\hat{y}+1}{y+1}\right)\right]^2$$

. To find the optimal predictor, note that

$$\begin{aligned} E\Big[\Big(\ln\left(\frac{a+1}{y+1}\right)\Big)^2\Big|x\Big] &= E\Big[\left\{\ln(y+1) - E[\ln(y+1)|x]\right\}^2 + \left\{E[\ln(y+1)|x] - \ln(a+1)\right\}^2\Big|x\Big] \\ &= Var(\ln(y+1)|x) + \left\{E[\ln(y+1)|x] - \ln(a+1)\right\}^2 \end{aligned}$$

Therefore, the optimal predictor $f(x)$ satisfies that

$$\ln(f(x)+1) = E[\ln(y+1)|x]$$

, which leads to

$$f(x) = \exp\left\{E[\ln(y+1)|x]\right\} - 1$$

2.3

(i). $h_1(v) = \ln(1+\exp(-v))/\ln(2)$.

$$\begin{aligned} \because g_1(\mathbf{x}) &= \arg\min_g E\Big[h_1(yg)\Big|\mathbf{x}\Big] = \arg\min_g E\Big[\ln(1+\exp(-yg))\Big|\mathbf{x}\Big] \\ &= \arg\min_g \Big\{P(y=1|\mathbf{x})\ln(1+exp(-g)) + P(y=-1|\mathbf{x})\ln(1+exp(g))\Big\} = \arg\min_g R(g) \end{aligned}$$

Let $c = P(y=1|\mathbf{x})$, then solve equation

$$\frac{\partial R(g)}{\partial g} = \frac{\exp(g)}{1+\exp(g)} - c = 0$$

we have

$$g_1(\mathbf{x}) = \ln\Big(\frac{c}{1-c}\Big) = \ln\Big(\frac{P(y=1|\mathbf{x})}{P(y=-1|\mathbf{x})}\Big).$$

And $\frac{\partial^2 R(g)}{\partial g^2}\Big|_{g=g_1} = \frac{\exp(g)}{(1+\exp(g))^2}\Big|_{g=g_1} > 0$. Therefore, the optimizer of $E\Big[h_1(yg(\mathbf{x}))\Big]$ is

$$g_1(\mathbf{x}) = \ln\Big(\frac{P(y=1|\mathbf{x})}{P(y=-1|\mathbf{x})}\Big).$$

(ii). $h_2(v) = \exp(-v)$.

Similarly, to find optimizer of $E\Big[h_2(yg(\mathbf{x}))\Big]$, i.e.

$$g_2(\mathbf{x}) = \arg\min_g E\Big[h_2(yg)\Big|\mathbf{x}\Big] = \arg\min_g \Big\{P(y=1|\mathbf{x})\exp(-g) + \exp(g)\Big\} = \arg\min_g R(g).$$

Then let $c = P(y=1|\mathbf{x})$, solve equation

$$\frac{\partial R(g)}{\partial g} = (1-c)\exp(g) - c\exp(-g) = 0,$$

we have

$$g_2(\mathbf{x}) = \frac{1}{2}\ln\Big(\frac{P(y=1|\mathbf{x})}{P(y=-1|\mathbf{x})}\Big).$$

Meanwhile, $\frac{\partial^2 R(g)}{\partial g^2}\Big|_{g=g_2} > 0$. Hence, $g_2(\mathbf{x}) = \frac{1}{2}\ln\Big(\frac{P(y=1|\mathbf{x})}{P(y=-1|\mathbf{x})}\Big)$ is an optimizer for $E\Big[h_2(yg(\mathbf{x}))\Big]$.

(iii). $h_3(v) = (1-v)_+$.

$$\begin{aligned} g_3(\mathbf{x}) &= \arg\min_g E\Big[h_3(yg)\Big|\mathbf{x}\Big] = \arg\min_g E\Big[(1-yg)_+\Big|\mathbf{x}\Big] \\ &= \arg\min_g \Big\{P(y=1|\mathbf{x})(1-g)_+ + P(y=-1|\mathbf{x}(1+g)_+)\Big\} \\ &= \arg\min_g \Big\{P(y=1|\mathbf{x})(1-g)I_{\{g\le 1\}} + P(y=-1|\mathbf{x})(1+g)I_{\{g\ge -1\}}\Big\} \end{aligned}$$

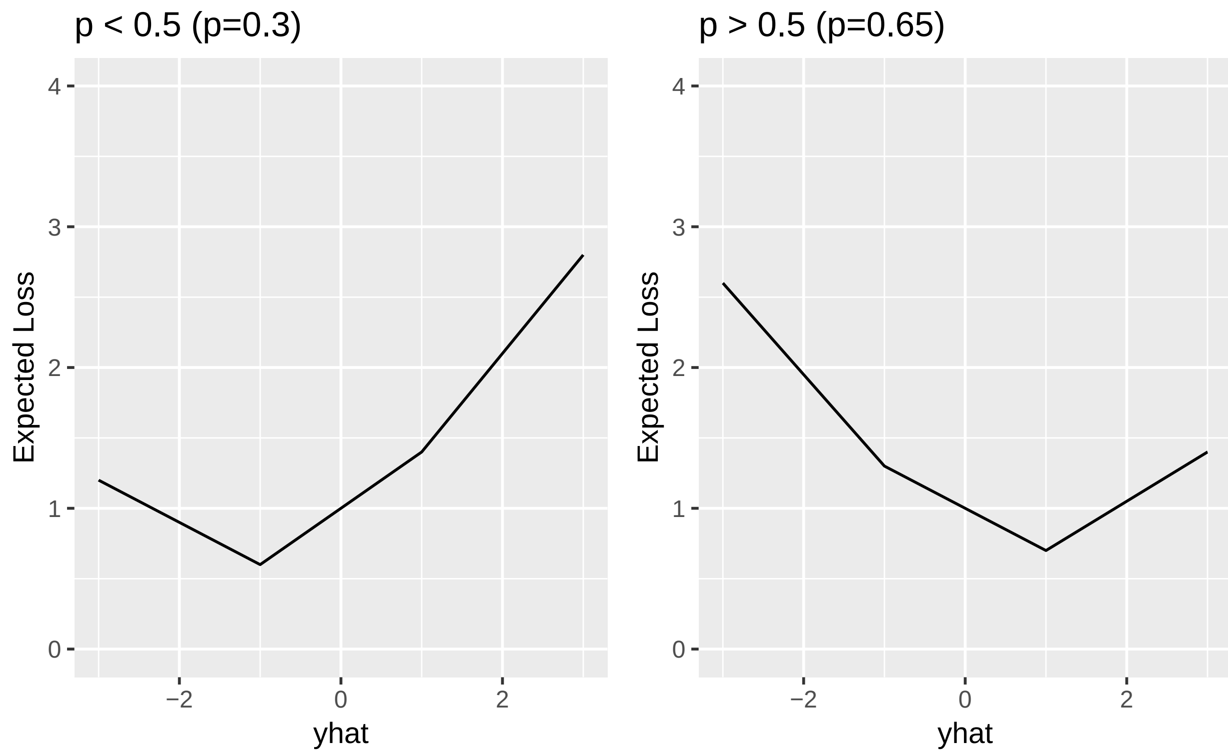
If $P(y=1|\mathbf{x}) < \frac{1}{2}$, $g_3(\mathbf{x}) = -1$. Otherwise when $P(y=1|\mathbf{x}) > \frac{1}{2}$, we have optimizer $g_3(\mathbf{x}) = 1$. Hence, an optimizer for $E\Big[h_3(yg(\mathbf{x}))\Big]$ is

$$g_3(\mathbf{x}) = \text{sign}\Big\{P(y=1|\mathbf{x}) - P(y=-1|\mathbf{x})\Big\}.$$

2.4

Similar as the loss function h_3 we considered in Question 2.3, given loss function $L(y,\hat{y}) = (1-y\hat{y})_+$ and suppose that $P[y=1] = p$, we have

$$E\Big[L(y,\hat{y})\Big] = p(1-\hat{y})_+ + (1-p)(1+\hat{y})_+ = p(1-\hat{y})I_{\{\hat{y}\le 1\}} + (1-p)(1+\hat{y})I_{\{\hat{y}\ge -1\}}$$



Hence, the optimal choice of $\hat{y}$ is

$$\hat{y} = \text{sign}\Big\{p - \frac{1}{2}\Big\} = \text{sign}\Big\{P(y = 1|\mathbf{x}) - P(y = -1|\mathbf{x})\Big\}$$

2.5

Consider $\pi_0 = \pi_1 = \frac{1}{2}$ and $g(x|0) = I[-0.5 < x < 0.5],\ g(x|1) = 12x^2 I[-0.5 < x < 0.5]$.

(a).

Given $0 - 1$ loss $L(y, f(x)) = I_{\{y \neq f(x)\}}$, the optimal classification rule is

$$\begin{aligned} f(x) &= \arg\min_a E\Big[I_{\{y \neq a\}}\Big|x\Big] = \arg\min_a \Big\{P(y = 1|x)I_{\{a=0\}} + P(y = 0|x)I_{\{a=1\}}\Big\} \\ &= I_{\{P(y=1|x) > P(y=0|x)\}} \end{aligned}$$

$$\because P(y = 1|x) = \frac{g(x|1)\pi_1}{g(x|0)\pi_0 + g(x|1)\pi_1} = \frac{12x^2}{12x^2 + 1} I_{\{-0.5 < x < 0.5\}}$$

$$\therefore f(x) = I_{\{P(y=1|x) > P(y=0|x)\}} = I_{\{\frac{12x^2}{12x^2+1} > \frac{1}{2}\}} I_{\{-0.5 < x < 0.5\}} = I_{\{\frac{1}{12} < x^2 < \frac{1}{4}\}}$$

The minimum expected loss is

$$E\Big[I_{\{y\neq f(x)\}}\Big] = P(y=1, f(x)=0) + P(y=0, f(x)=1) = \pi_1 P(f(x)=0|y=1) + \pi_0 P(f(x)=1|y=0)$$
$$= \frac{1}{2}P\Big(x^2 < \frac{1}{12}\Big|y=1\Big) + \frac{1}{2}P\Big(\frac{1}{12} < x^2 < \frac{1}{4}\Big|y=0\Big) = \frac{9-2\sqrt{3}}{18}.$$

(b).

$t(x) = x^2$ would be a good choice. Because in (a) we have shown that the ratio of $\frac{P(y=1|x)}{P(y=0|x)} \propto x^2$ and the optimal classifier $I_{\{\frac{1}{12}<x^2<\frac{1}{4}\}}$ is also closely related to x^2.

3.3

(a)

$$\begin{aligned}
\because\ E(y-f(x))^2 &= E[E\{(y-f(x))^2|x\}] \\
&= E[E\{y^2 - 2yf(x) + f^2(x)|x\}] \\
&= E[f^2(x) - 2f(x)E(y|x) + E(y^2|x)] \\
&= E[f^2(x) - 2f(x)\sin(x) + \sin^2(x) + \frac{1}{4}(|x|+1)^2] \\
&= E[(f(x)-\sin(x))^2 + \frac{1}{4}(|x|+1)^2] \\
&\geq E[\frac{1}{4}(|x|+1)^2]
\end{aligned}$$

$$\begin{aligned}
\therefore \min_f\{E(y-f(x))^2\} &= E[\frac{1}{4}(|x|+1)^2] \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{4}(|x|+1)^2dx \\
&= \frac{1}{4}(1+\pi+\frac{1}{3}\pi^2)
\end{aligned}$$

$$\therefore\ f^*(x) = \text{argmin}_f\{E(y-f(x))^2\} = \sin(x)$$

(b)

For linear predictor, define:

$$l_1 = E(\sin(x) - g_1(x))^2 = E(\sin(x) - a - bx)^2$$

Solve:

$$\begin{cases} \dfrac{\partial l_1}{\partial a} = -E(\sin(x) - a - bx) = 0 \\ \dfrac{\partial l_1}{\partial b} = -E[x(\sin(x) - a - bx)] = 0 \end{cases}$$

then:

$$\begin{cases} \hat{a} = 0 \\ \hat{b} = \dfrac{Ex\sin(x)}{Ex^2} = \dfrac{3}{\pi^2} \end{cases}$$

Thus,

$$g_1^*(x) = \frac{3x}{\pi^2}$$

Similarly, for cubic predictor, define:

$$l_3 = E(\sin(x) - g_3(x))^2 = E(\sin(x) - a - bx - cx^2 - dx^3)^2$$

Solve:

$$\begin{cases} \frac{\partial l_3}{\partial a} = -E(\sin(x) - a - bx - cx^2 - dx^3) = 0 \\ \frac{\partial l_3}{\partial b} = -E[x(\sin(x) - a - bx - cx^2 - dx^3)] = 0 \\ \frac{\partial l_3}{\partial c} = -E[x^2(\sin(x) - a - bx - cx^2 - dx^3)] = 0 \\ \frac{\partial l_3}{\partial d} = -E[x^3(\sin(x) - a - bx - cx^2 - dx^3)] = 0 \end{cases}$$

Since: $E(\sin(x)) = E(x) = E(x^3) = E(x^5) = E(x^2 \sin(x)) = 0$, $E(x^2) = \frac{\pi^2}{3}$, $E(x^4) = \frac{\pi^4}{5}$, $E(x^6) = \frac{\pi^6}{7}$, $E(x \sin(x)) = 1$, $E(x^3 \sin(x)) = \pi^2 - 6$.

Therefore, the above equation system is equivalent to:

$$\begin{cases} a + \frac{\pi^2}{3}c = 0 \\ \frac{\pi^2}{3}b + \frac{\pi^4}{5}d = 1 \\ \frac{\pi^2}{3}a + \frac{\pi^4}{5}c = 0 \\ \frac{\pi^4}{5}b + \frac{\pi^6}{7}d = \pi^2 - 6 \end{cases}$$

and:

$$\begin{cases} \hat{a} = 0 \\ \hat{b} = \frac{15(21 - \pi^2)}{2\pi^4} \\ \hat{c} = 0 \\ \hat{d} = -\frac{35(15 - \pi^2)}{2\pi^6} \end{cases}$$

Thus,

$$g_3^*(x) = \frac{15(21 - \pi^2)}{2\pi^4}x - \frac{35(15 - \pi^2)}{2\pi^6}x^3$$

Using cubic predictor cannot eliminate model bias since $\sin(x)$ does not belongs to the cubic functions group.

3.4

```
formula1.0 <- "y~ 1"
cv_10fd(D3_4, formula1.0)
```

```
## [1] 2.078989
```

```r
formula1.1 <- "y~ x"
cv_10fd(D3_4, formula1.1)
```

```
## [1] 2.124108
```

```r
formula1.2 <- "y~ x + I(x^2)"
cv_10fd(D3_4, formula1.2)
```

```
## [1] 2.064061
```

```r
formula1.3 <- "y~ x + I(x^2) + I(x^3)"
cv_10fd(D3_4, formula1.3)
```

```
## [1] 1.906987
```

```r
formula1.4 <- "y~ x + I(x^2) + I(x^3) + I(x^4)"
cv_10fd(D3_4, formula1.4)
```

```
## [1] 1.873633
```

```r
formula1.5 <- "y~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5)"
cv_10fd(D3_4, formula1.5)
```

```
## [1] 1.93059
```

```r
formula2 <- "y~ sin(x) + cos(x)"
cv_10fd(D3_4, formula2)
```

```
## [1] 1.787845
```

```r
formula3 <- "y~ sin(x) + cos(x) + sin(2*x) + cos(2*x)"
cv_10fd(D3_4, formula3)
```

```
## [1] 1.844818
```

```r
formula4 <- "y~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + sin(x) + cos(x) + sin(2*x) + cos(2*x)"
cv_10fd(D3_4, formula4)
```

```
## [1] 2.028212
```

We can see from the R output that the basis: $\{1, \sin(x), \cos(x)\}$ has the best performance (say, the smallest CV scores) among all sets of basis. All basis that consist $\sin(x)$ do not have model bias.

3.5

(a)

See the result of Problem 2.5, the minimum expect loss is: $\frac{1}{2} - \frac{\sqrt{3}}{9}$, in this case, $f^*(x) = I(12x^2 \geq 1) = I(x \geq \frac{\sqrt{3}}{6} \ or \ x \leq -\frac{\sqrt{3}}{6})$

(b)

To minimize $E\{I(y \neq g_c(x))\}$, it's equivalent to maximize $E\{I(y = g_c(x))\}$,

$$\begin{aligned}
\because \; E\{I(y = g_c(x))\} &= P(y = g_c(x)) \\
&= \frac{1}{2}P(g_c(x) = 1|y = 1) + \frac{1}{2}P(g_c(x) = 0|y = 0) \\
&= \frac{1}{2}P(x \geq c|y = 1) + \frac{1}{2}P(x < c|y = 0)
\end{aligned}$$

$$\therefore \; E\{I(y = g_c(x))\} = \begin{cases} \frac{1}{2}, & c \notin (-\frac{1}{2}, \frac{1}{2}) \\ \frac{1}{2}\int_c^{1/2} 12x^2 dx + \frac{1}{2}\int_{-1/2}^{c} dx, & c \in (-\frac{1}{2}, \frac{1}{2}) \end{cases}$$

It's easy to find $\hat{c} = \frac{\sqrt{3}}{6}$ that maximize $E\{I(y = g_c(x))\}$. Thus,

$$g_{\hat{c}}(x) = I(x \geq \frac{\sqrt{3}}{6})$$

$$E\{I(y \neq g_{\hat{c}}(x))\} = \frac{1}{2} - \frac{\sqrt{3}}{18}$$

The model penalty is:

$$E\{I(y \neq g_{\hat{c}}(x))\} - E\{I(y \neq f^*(x)\} = \frac{\sqrt{3}}{18}$$

(c)

In part (b), we maximize:

$$E\{I(y = g_c(x))\} = P(y = 1)P(x \geq c|y = 1) + P(y = 0)P(x < c|y = 0)$$

now $P(y = 1)$, $P(y = 0)$, $G(x < x_0|y = 1)$, and $G(x < x_0|y = 0)$ are unknown, we can estiamte them by:

$$\hat{P}(y = 0) = \hat{\pi}_0 = \frac{\sum_{i=1}^{N} I(y_i = 0)}{N}; \quad \hat{P}(y = 1) = 1 - \hat{\pi}_0$$

$$\hat{G}(x < x_0|y) = \frac{\#\text{ training cases with } x_i \leq x_0 \text{ and } y_i = y}{\#\text{ training cases with } y_i = y}$$

Thus, we can estimate $E\{I(y = g_c(x))\}$ by:

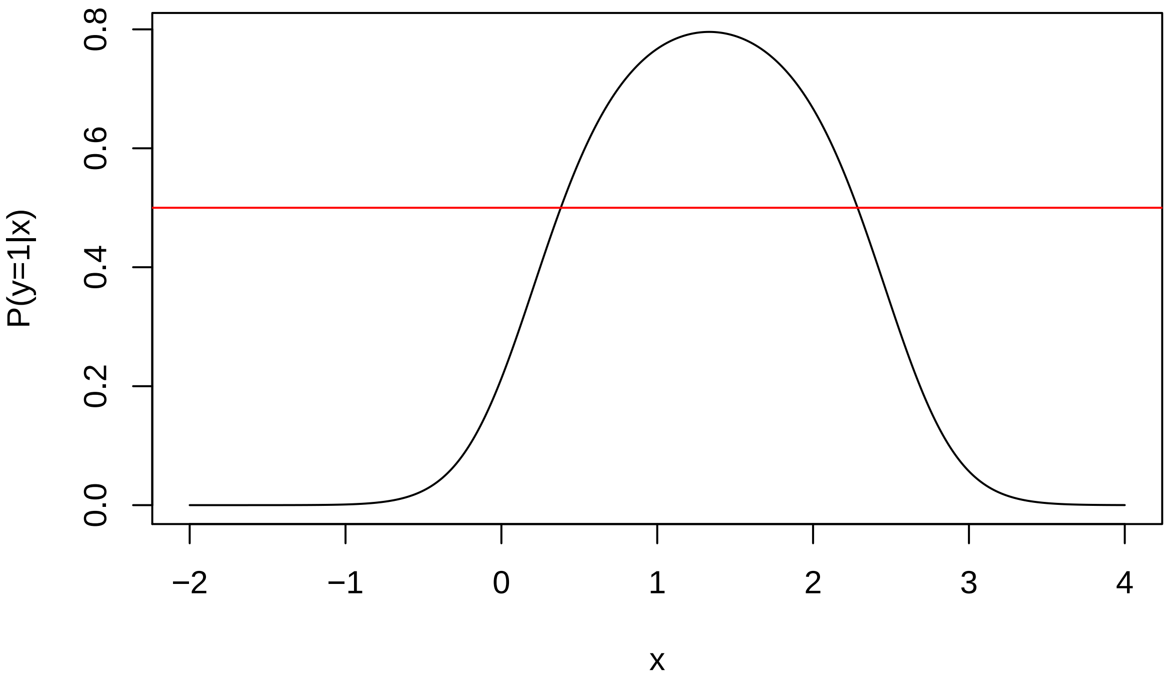
$$\hat{E}\{I(y = g_c(x))\} = (1 - \hat{\pi}_0)(1 - \hat{G}(c|1)) + \hat{\pi}_0\hat{G}(c|0)$$

We can numerically find $\hat{c}$ that maximize $\hat{E}\{I(y = g_c(x))\}$.

3.11

(a)

$$
\begin{aligned}
P(y=1|x) &= \frac{f(x|y=1)\pi_1}{f(x|y=0)\pi_0 + f(x|y=1)\pi_1} \\
&= \frac{f(x|y=1)}{f(x|y=0) + f(x|y=1)} \\
&= \frac{2\exp\{-2(x-1)^2\}}{\exp\{-x^2/2\} + 2\exp\{-2(x-1)^2\}}
\end{aligned}
$$



(b)

Similar to Problem 2.5, the optimal classfier is:

$$
f^*(x) = I(P(y=1|x) > P(y=0|x)) = I(P(y=1|x) > \frac{1}{2})
$$

```
range( x[which( P1x(x)>0.5 )] )
```

```
## [1] 0.382 2.285
```

By R output, we find that $f^*(x) = I(0.382 < x < 2.285)$

In this case, the expected loss is:

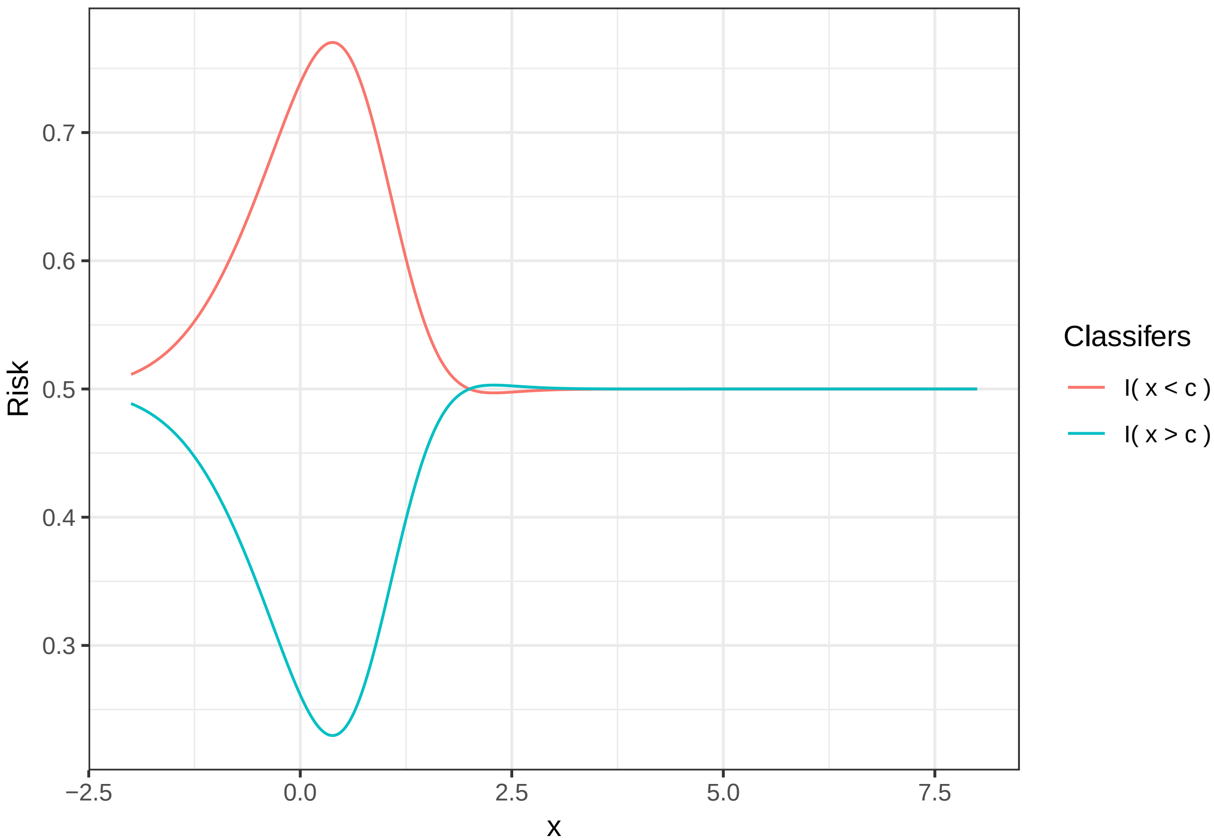
$$
\begin{aligned}
EI(y \neq f^*(x)) = P(y \neq f^*(x)) \\
&= \frac{1}{2}P(f^*(x) = 1|0) + \frac{1}{2}P(f^*(x) = 0|1) \\
&= \frac{1}{2}P(0.382 < x < 2.285|0) + \frac{1}{2}P(x \geq 2.285 \ or \ x \leq 0.382|1)
\end{aligned}
$$

We can compute the result by R:

```
R1 <- (1/2)*( pnorm(2.285,0,1)-pnorm(0.382,0,1) ) +
  (1/2)*( 1-pnorm(2.285,1,1/2) + pnorm(0.382,1,1/2) )
R1
```

```
## [1] 0.2266942
```

(c)



From the plot, it's easy to find that classifier $I(x > c)$ is prefered, we can find $\hat{c}$ by:

```
x[which.min(Ruc(x))]
```

```
## [1] 0.381
```

The expected loss is:

$$
\begin{aligned}
EI(y \neq I(x > 0.381))) &= P(y \neq I(x > 0.381)) \\
&= \frac{1}{2}P(x > 0.381|0) + \frac{1}{2}P(x \leq 0.381|1)
\end{aligned}
$$

We can compute the result by R:

```
R2 <- (1/2)*( 1-pnorm(0.381,0,1) ) +
  (1/2)*(  pnorm(0.381,1,1/2) )
R2
```

```
## [1] 0.2297298
```

The modeling penalty is just:

```r
R2 - R1
```

```
## [1] 0.003035589
```

(d)

For $N = 100$ case, the mean of estimated $\hat{c}$, the frequency of the classifier with the form $I(x < c)$, the mean of the conditonal error rate for the 10000 training samples are given in the following R output, the fitting penalty are also computed:

```r
## N=100
Out100 <- classifier_3_11d(M,N1)

## mean c_hat
mean(Out100$C)
```

```
## [1] 0.4003703
```

```r
## model frequency
mean(Out100$Model_ind)
```

```
## [1] 0
```

```r
## Risk for N=100
R3_100 <- mean(Out100$Con_err)
R3_100
```

```
## [1] 0.2390229
```

```r
## fitting penalty
R3_100 - R2
```

```
## [1] 0.009293059
```

We computed the same statistics for $N = 50$ case:

```r
## N=50
Out50 <- classifier_3_11d(M,N2)
## mean c_hat
mean(Out50$C)
```

```
## [1] 0.4141971
```

```r
## model frequency
mean(Out50$Model_ind)
```

```
## [1] 0
```

```r
## Risk for N=100
R3_50 <- mean(Out50$Con_err)
R3_50
```

```
## [1] 0.2442063
```

```r
## fitting penalty
R3_50 - R2
```

```
## [1] 0.01447647
```

Thus, the fitting penalty indeed increases when N decreases from 100 to 50.

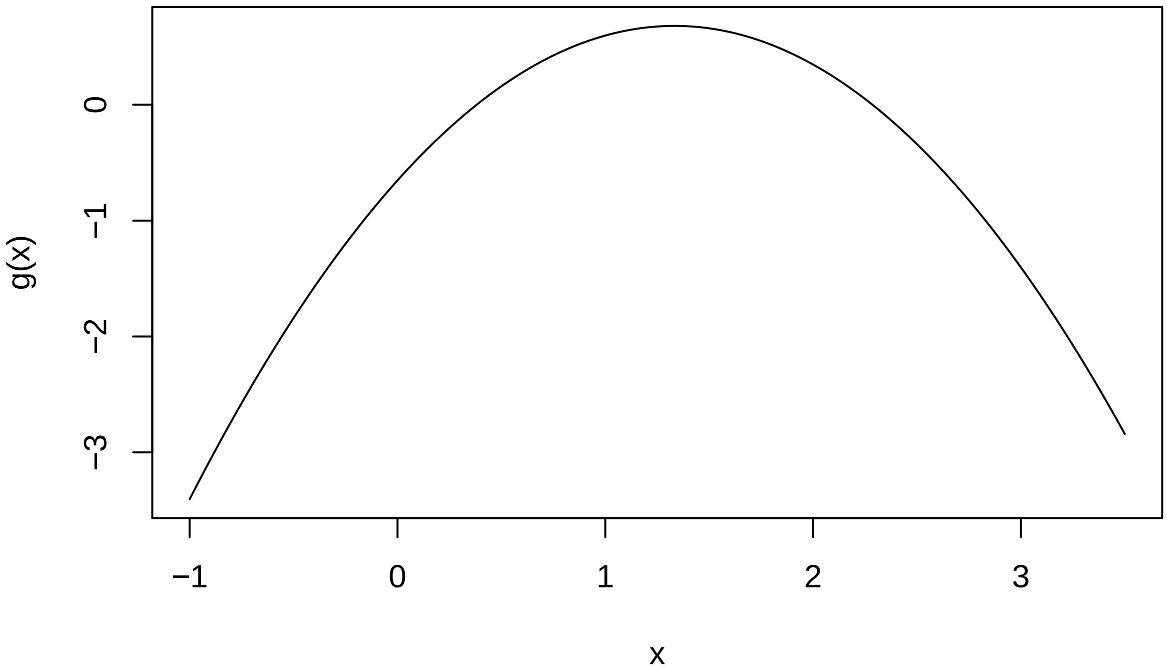
3.12

(a)

By problem 2.3:

$$
\begin{aligned}
g^*(x) &= \frac{1}{2}\log\frac{P(y=1|x)}{P(y=-1|x)} \\
&= \frac{1}{2}\log\frac{f(x|y=1)}{f(x|y=0)} \\
&= \frac{1}{2}\log\frac{2e^{-2(x-1)^2}}{e^{-x^2/2}} \\
&= -\frac{3}{4}x^2 + 2x - 1 + \frac{1}{2}\log 2
\end{aligned}
$$

The plot of $g^*(x)$ is given below:



(b)

Using R function "nlm":

```
fit2 <- nlm(f=Risk,p=c(0,0,0),x=D3_12$x, y=D3_12$y)
fit2$estimate
```

```
## [1]  0.2259491  1.5666600 -0.7336744
```

(c)

Using "nlm" for $\lambda \in \{1, 0.1, 0.01, 0.001\}$:

```
fit3_0 <- nlm(f=PRisk,p=c(0,0,0),x=D3_12$x, y=D3_12$y,lambda=1)
fit3_0$estimate
```

```
## [1]  0.004986373  1.072750906 -0.119046792
```

```
fit3_1 <- nlm(f=PRisk,p=c(0,0,0),x=D3_12$x, y=D3_12$y,lambda=0.1)
fit3_1$estimate
```

```
## [1]  0.1282131  1.3466865 -0.4340556
```

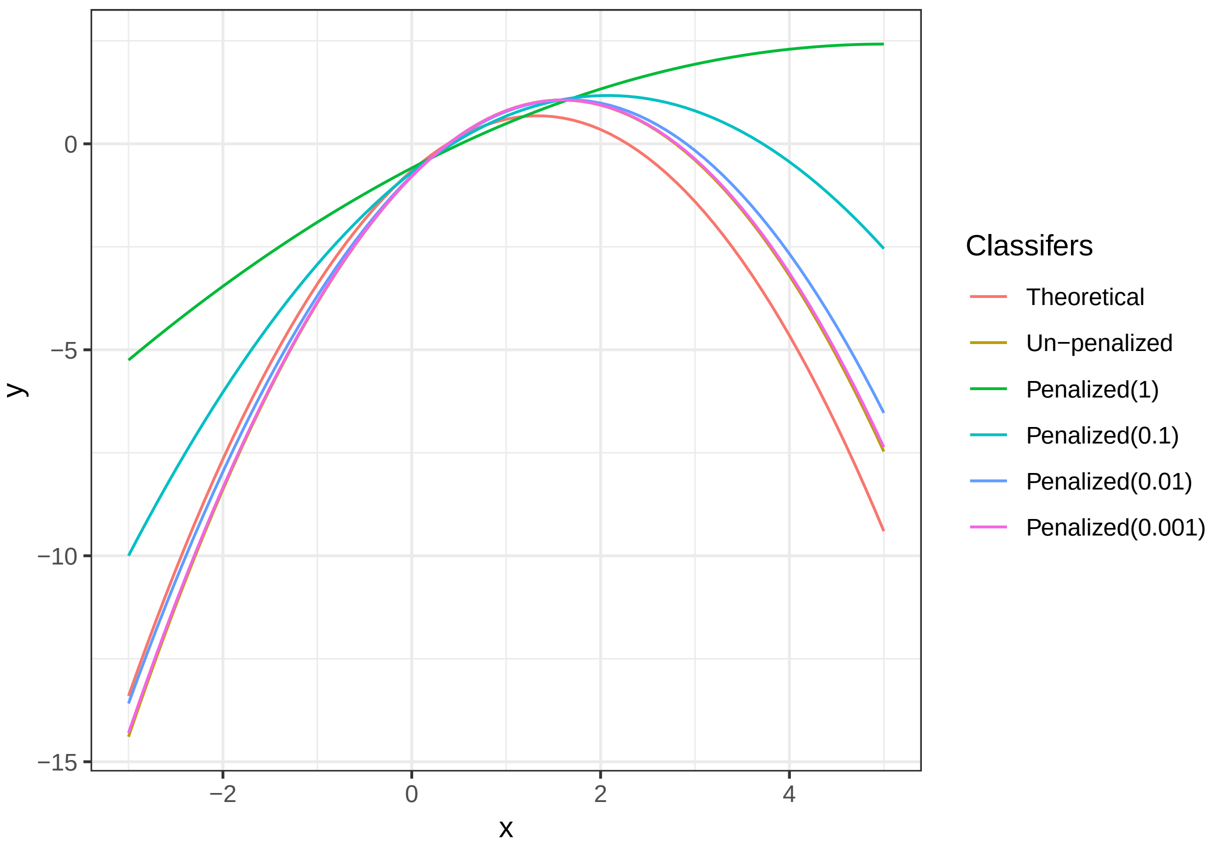
```
fit3_2 <- nlm(f=PRisk,p=c(0,0,0),x=D3_12$x, y=D3_12$y,lambda=0.01)
fit3_2$estimate
```

```
## [1]  0.2082287  1.5295405 -0.6777548
```

```
fit3_3 <- nlm(f=PRisk,p=c(0,0,0),x=D3_12$x, y=D3_12$y,lambda=0.001)
fit3_3$estimate
```

```
## [1]  0.2239894  1.5626297 -0.7274714
```

The plot is given below:



5.7

First of all, it's easy to find that $\{1, t, \cos(t)\}$ are pairwise perpendicular, since $< 1, t >= 0$, $< 1, \cos(t) >= 0$, $< t, \cos(t) >= 0$. Also note that $< 1, 1 >= 2\pi$, $< t, t >= \frac{2\pi^3}{3}$, $< \cos(t), \cos(t) >= \pi$, thus: $q_1 = \frac{1}{\sqrt{2\pi}}$, $q_2 = t/\sqrt{\frac{2\pi^3}{3}}$, $q_3 = \cos(t)/\sqrt{\pi}$.

On the other hand, since $< 1, \sin(t) >= 0$, $< t, \sin(t) >= 2\pi$, $< \sin(t), \cos(t) >= 0$, hence:

$$z_4 = \sin(x) - \frac{< t, \sin(t) >}{< t, t >} t = \sin(x) - \frac{3t}{\pi^2}$$

Finally, given that:

$$< z_4, z_4 >= \int_{-\pi}^{\pi} (\sin(x) - \frac{3t}{\pi^2})^2 dt = \pi - \frac{6}{\pi}$$

we have: $q_4 = (\sin(x) - \frac{3t}{\pi^2})/\sqrt{\pi - \frac{6}{\pi}}$. Thus, one of the 4-dimensional orthonormal basis is:

$$\{1/\sqrt{2\pi}, \quad t/\sqrt{\frac{2\pi^3}{3}}, \quad \cos(t)/\sqrt{\pi}, \quad (\sin(x) - \frac{3t}{\pi^2})/\sqrt{\pi - \frac{6}{\pi}}\}$$