Corey's - HW 1 Problems

GUILHERME FRANÇA

Wednesday, February 15, 2017 7:25 PM

2.1 Show that if $(X,Y) \mapsto (T(X),Y)$ then $L^*_{(T(X),Y)} \geq L^*(X,Y)$.

Recall Th. 1 which states that for any classifier g we have $L(g^*) = \mathbb{P}[g^*(X) \neq Y] \leq L(g) = \mathbb{P}[g(X) \neq Y]$, i.e g^* is optimal.

Let $T: X \mapsto Z = T(X)$ be a transformation. In the "Z-space" the Bayes classifier is defined by

$$h^*(z) = \begin{cases} 1 & \text{if } \sigma(z) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

where $\sigma(z) = \mathbb{P}[Y=1 \mid Z]$, and its error is $L(h^*) = \mathbb{P}[h^*(Z) \neq Y]$. Now suppose that $L(h^*) < L(g)$. Then we would be able to construct the following classifier in the original "X-space":

$$g(x) = \begin{cases} 1 & \text{if } \eta \circ T(X) = \eta(T(X)) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

where $\eta = \mathbb{P}[Y=1 \mid X]$ and $L(g) < L(g^*)$ by assumption. But this contradicts the above mentioned Th. 1, therefore we must have $L(h^*) \geq L(g)$.

In other words, we can incorporate the transformation in a classifier but this cannot improve Bayes error since it is optimal.

2.2 Let X' be independent of (X,Y). Show that $L^*_{(X,X')} = L^*_X$.

Recall $L_X(g^*) = \mathbb{P}[g^*(X) \neq Y]$ and $g^*(x) = \begin{cases} 1, \eta(x) > 1/2 \\ 0, \text{otherwise} \end{cases}$ where $\eta(x) = \mathbb{P}[Y=1 \mid X]$.

Analogously, $L_{(X,X')}(h^*(X,X') \neq Y)$ with $h^* = \begin{cases} 1, \sigma(X,X') > 1/2 \\ 0, \text{otherwise} \end{cases}$ where $\sigma(X,X') = \mathbb{P}[Y=1 \mid X,X']$.

Now if $X' \perp (X,Y)$, then $\mathbb{P}[Y \mid X,X'] = \mathbb{P}[Y \mid X]$, thus $h^*(X,X') = g^*(X)$ implying $L^*_{(X,X')} = L^*_X$.

2.5 For $c \in (0, 1/2)$ define

$$g_c(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 + c \\ 0 & \text{if } \eta(x) \leq 1/2 - c \\ \text{"reject"} & \text{otherwise} \end{cases}$$

Show that, for any g, if

$$\mathbb{P}[g(X) = \text{"reject"}] \leq \mathbb{P}[g_c(X) = \text{"reject"}]$$

then

$$\mathbb{P}[g(X) \neq Y \mid g(X) \neq \text{"reject"}] \geq \mathbb{P}[g_c(X) \neq Y \mid g_c(X) \neq \text{"reject"}].$$

Thus decisions under g_c are optimal.

Notice that

$$\mathbb{P}[g(X) \neq Y \mid g(X) \neq R] = 1 - \mathbb{P}[g(X) = Y \mid g(X) \neq R]$$

$$= 1 - \frac{\mathbb{P}[g = Y,\ g \neq R]}{\mathbb{P}[g \neq R]}$$

$$\begin{aligned}\mathbb{P}[g \neq Y \mid g \neq R] &= 1 - \mathbb{P}[g = Y \mid g \neq R] \\ &= 1 - \frac{\mathbb{P}[g = Y,\ g \neq R]}{\mathbb{P}[g \neq R]} \\ &= 1 - \frac{\mathbb{1}(g \neq R)\,\mathbb{P}[g = Y]}{\mathbb{P}[g \neq R]}\end{aligned}$$

Thus

$$\mathbb{P}[g \neq Y \mid g \neq R] - \mathbb{P}[g_c \neq Y \mid g_c \neq R] = \frac{\mathbb{1}(g_c \neq R)\,\mathbb{P}[g_c = Y]}{\mathbb{P}[g_c \neq R]} - \frac{\mathbb{1}(g \neq R)\,\mathbb{P}[g = Y]}{\mathbb{P}[g \neq R]}$$

It is enough to show that

$$\frac{\mathbb{P}[g_c = Y \mid X]}{\mathbb{P}[g_c \neq R]} - \frac{\mathbb{P}[g = Y \mid X]}{\mathbb{P}[g \neq R]} \geq 0 \qquad (\#)$$

Consider $\mathbb{P}[g = Y \mid X] = \mathbb{P}[g = 1, Y = 1 \mid X] + \mathbb{P}[g = 0, Y = 0 \mid X]$
$= \mathbb{1}(g = 1)\,\eta(x) + \mathbb{1}(g = 0)\,(1 - \eta(x))$

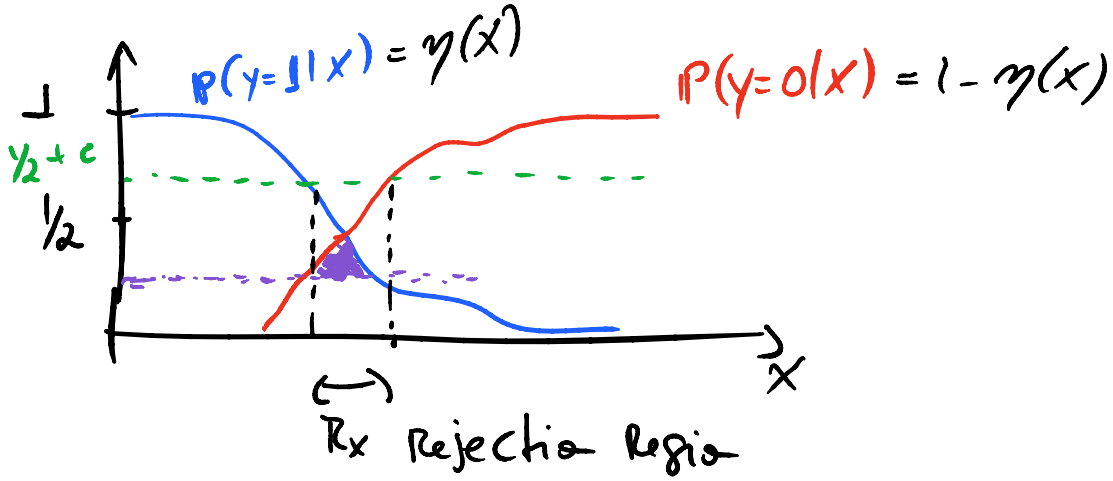
Thus we have

$$\frac{\mathbb{1}(g_c = 1)\,\eta + \mathbb{1}(g_c = 0)(1-\eta)}{\mathbb{P}[g_c \neq R]} - \frac{\mathbb{1}(g = 1)\,\eta + \mathbb{1}(g = 0)(1-\eta)}{\mathbb{P}[g \neq R]} \qquad (*)$$

If $g = g_c$ this gives $\left(\frac{1}{\mathbb{P}[g_c \neq R]} - \frac{1}{\mathbb{P}[g \neq R]}\right) \cdot C \geq 0$,

where $C = \eta$ or $C = 1 - \eta$ (both are > 0). If $g_c = 1$ and $g = 0$ we get $\frac{\eta}{\mathbb{P}[g_c \neq R]} - \frac{(1-\eta)}{\mathbb{P}[g \neq R]} > 0$ since $\eta > 1 - \eta$ and $\mathbb{P}[g_c \neq R] \leq \mathbb{P}[g \neq R]$. Analogous argument works for $g_c = 0$ and $g = 1$. Therefore, in any case, $(*) \geq 0$, thus $(\#) \geq 0$. Integrating over X gives the claim.

The intuition behind this is the following.



we assign x to class $\{1,0\}$ according to $\max\{\eta(x), 1-\eta(x)\} > \frac{1}{2}+c$, and reject any decision otherwise. We only incur in error if we make a wrong assignment when we do not reject. The error comes when $\mathbb{P}[y=1|x]$ and $\mathbb{P}[y=0|x]$ have similar values and overlap, as indicated in the PURPLE shaded area. Thus if we lower the rejection threshold, $\mathbb{P}[g \neq \text{Reject}]$, more points x will have non-reject decisions in the shaded area which increases the misclassification error.

2.9 | $g(x) = \begin{cases} 0 & \text{if } \tilde{\eta}_1(x) \leq \tilde{\eta}_0(x) \\ 1 & \text{otherwise} \end{cases}$

$\tilde{\eta}_1(x) \sim \eta(x)$ It may be the case that
$\tilde{\eta}_0(x) \sim 1 - \eta(x)$ $\tilde{\eta}_0 + \tilde{\eta}_1 \neq 1$

Theo. $L(g) - L(g^*) \leq \int_{\mathbb{R}^d} |(1 - \eta(x)) - \tilde{\eta}_0(x)| \mu(dx)$
$+ \int_{\mathbb{R}^d} |\eta(x) - \tilde{\eta}_1(x)| \mu(dx)$

Proof

$$\begin{aligned} \mathbb{P}[g(x) \neq Y | X] &= 1 - \mathbb{P}[g(x) = Y | X] \\ &= 1 - \mathbb{P}[g=1, Y=1 | X] - \mathbb{P}[g=0, Y=0 | X] \\ &= 1 - \mathbb{1}(g=1)\mathbb{P}[Y=1|X] - \mathbb{1}(g=0)\mathbb{P}[Y=0|X] \\ &= 1 - \mathbb{1}(g=1)\tilde{\eta}_1(x) - \mathbb{1}(g=0)\tilde{\eta}_0(x) \end{aligned}$$

$$\mathbb{P}[g^*(x) \neq Y | X] = 1 - \mathbb{1}(g^*=1)\eta(x) - \mathbb{1}(g^*=0)(1-\eta(x))$$

Therefore,

$$\begin{aligned} L(g) - L(g^*) = &\int \mu(dx)\left(\eta(x)\mathbb{1}(g^*=1) - \tilde{\eta}_1(x)\mathbb{1}(g=1)\right) \\ &+ \int \mu(dx)\left((1-\eta(x))\mathbb{1}(g^*=0) - \tilde{\eta}_0(x)\mathbb{1}(g=0)\right) \end{aligned}$$

$$L(g)-L(g^*)=\int \mu(dx)\left(\eta \mathbb{1}(\eta>1/2)-\tilde{\eta}_1 \mathbb{1}(\tilde{\eta}_1>\tilde{\eta}_0)\right)$$
$$+\int \mu(dx)\left((1-\eta)\mathbb{1}(\eta\le 1/2)-\tilde{\eta}_0\mathbb{1}(\tilde{\eta}_1\le\tilde{\eta}_0)\right)$$

Now we consider each possibility.

1. $\eta>1/2$, $\tilde{\eta}_1>\tilde{\eta}_0$. The above integrand is
$\eta-\tilde{\eta}_1\le|\eta-\tilde{\eta}_1|\le|\eta-\tilde{\eta}_1|+|(1-\eta)-\tilde{\eta}_0|$

2. $\eta\le 1/2$, $\tilde{\eta}_1\le\tilde{\eta}_0$. The integrand is
$(1-\eta)-\tilde{\eta}_0\le|(1-\eta)-\tilde{\eta}_0|\le|(1-\eta)-\tilde{\eta}_0|+|\eta-\tilde{\eta}_1|$

3. $\eta>1/2$, $\tilde{\eta}_1\le\tilde{\eta}_0$. The integrand is
$\eta-\tilde{\eta}_0\le|\eta-\tilde{\eta}_1|\le|\eta-\tilde{\eta}_1|+|(1-\eta)-\tilde{\eta}_0|$

4. $\eta\le 1/2$, $\tilde{\eta}_1>\tilde{\eta}_0$. The integrand is
$(1-\eta)-\tilde{\eta}_1\le|(1-\eta)-\tilde{\eta}_0|\le|(1-\eta)-\tilde{\eta}_0|+|\eta-\tilde{\eta}_1|$

Thus in either case we get

$$L(g)-L(g^*)\le\int\mu(dx)\,|\eta-\tilde{\eta}_1|+\int\mu(dx)\left((1-\eta)-\tilde{\eta}_0\right)$$

as desired.

TRUNK'S PROBLEM

Consider a two-class problem where

$p(x|\omega_1) = N(x|\mu, I)$

$p(x|\omega_2) = N(x|-\mu, I)$

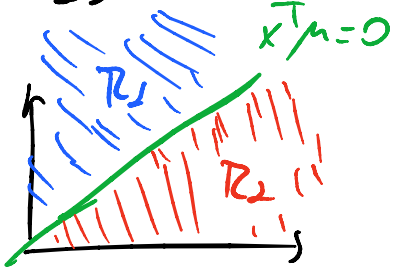
with $p(\omega_1) = p(\omega_2) = 1/2$ and $\mu \in \mathbb{R}^D$ with elements given by $\mu_i = \sqrt{i}$, $i = 1, \ldots, D$.

The decision boundary is determined by:

$$p(\omega_1|x) > p(\omega_2|x) \qquad (x \in C_1)$$

$$\frac{p(x|\omega_1)p(\omega_1)}{p(x)} > \frac{p(x|\omega_2)p(\omega_2)}{p(x)}$$

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > 1$$



or

$$\log p(x|\omega_1) - \log p(x|\omega_2) > 0$$

$$-\|x-\mu\|^2 + \|x+\mu\|^2 > 0$$

$$-\|x\|^2 + 2x^T\mu - \|\mu\|^2 + \|x\|^2 + 2x^T\mu + \|\mu\|^2 > 0$$

$$\boxed{x^T\mu > 0}$$

Thus the decision rule is $\begin{cases} x \in C_1 \text{ if } x^T\mu > 0 \\ x \in C_2 \text{ if } x^T\mu \leq 0 \end{cases}$.

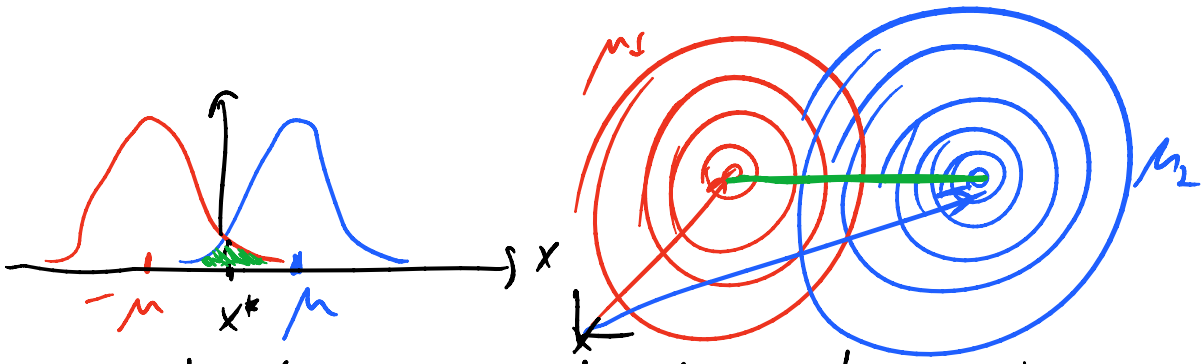
The misclassification error is

$$P_e = P(x \in R_1, x \in C_2) + P(x \in R_2, x \in C_1)$$

$$= \int_{R_1} p(x, \omega_2)\, dx + \int_{R_2} p(x, \omega_1)\, dx$$

$$P_e = \frac{1}{2}\int_{R_1} p(x|\omega_2)dx + \frac{1}{2}\int_{R_2} p(x|\omega_1)dx$$

$$= \frac{1}{2}\int_{R_1} \frac{1}{(2\pi)^{D/2}} e^{-\frac{1}{2}\|x-\mu_2\|^2} dx + \frac{1}{2}\int_{R_2} \frac{1}{(2\pi)^{D/2}} e^{-\frac{1}{2}\|x-\mu_1\|^2} dx$$



we can write x in a coordinate system where one of the axis is in the direction $\mu_1 - \mu_2$, the other ones being orthogonal, i.e, $\mu = (\|\mu\|, 0, \ldots, 0)^T$ and $X = (x_1, x_2, \ldots, x_D)$. Then only the first component contribute. We thus have

$$P_e = \frac{1}{2}\int_0^\infty \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}(x+\mu)^2} dx + \frac{1}{2}\int_{-\infty}^0 \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}(x-\mu)^2} dx$$

$$= \frac{1}{2}\int_\mu^\infty \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}z^2} dz + \frac{1}{2}\int_{-\infty}^{-\mu} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}z^2} dz$$

$$= \int_\mu^\infty \frac{1}{(2\pi)^{1/2}} e^{-\frac{z^2}{2}} dz$$

where $\mu \equiv \|\mu\| = \left(\sum_{m=1}^{D} \frac{1}{m}\right)$. ← divergent series.

Thus, $\boxed{\lim_{D\to\infty} P_e = 0}$. This would not happen if the components of μ decay faster than $1/\sqrt{\cdot}$.

Now suppose we have n_1 samples from $p(x|\omega_1)$ and n_2 samples from $p(x|\omega_2)$:

$\{x_1, \ldots, x_{n_1}\} \rightarrow \hat{\mu}_1 = \frac{1}{n_1}\sum_i x_i = \hat{\mu}$

$\{\tilde{x}_1, \ldots, \tilde{x}_{n_2}\} \quad \hat{\mu}_2 = \frac{1}{n_2}\sum_i \tilde{x}_i = -\hat{\mu}$

or $\tilde{x}_i \rightarrow -\tilde{x}_i$ and we have $\{x_1, \ldots, x_N\}$, $N = n_1 + n_2$, elements. Then

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(MLE estimator).}$$

the prob. of error is

$$\begin{aligned} P_e &= P(x \in R_1, x \in C_2) + P(x \in R_2, x \in C_1) \\ &= \frac{1}{2} P(x^T\hat{\mu} > 0 \mid \omega_2) + \frac{1}{2} P(x^T\hat{\mu} \leq 0 \mid \omega_1) \\ &= P(x^T\hat{\mu} > 0 \mid \omega_2) \\ &\equiv z \end{aligned}$$

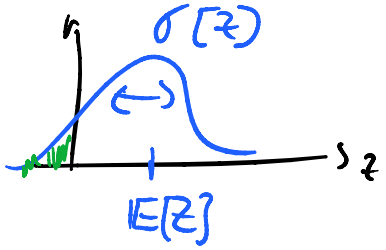
$$\mathbb{E}[z] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[x^T x_i] \overset{\text{iid}}{=} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[x^T]\,\mathbb{E}[x_i]$$

$$= \frac{1}{N} N\, \mu^T\mu = \|\mu\|^2 = \sum_{m=1}^{D} \frac{1}{m}$$

$$\mathrm{Var}[z] = \left(1 + \frac{1}{N}\right)\sum_{m=1}^{D}\frac{1}{m} + \frac{D}{N} \qquad \text{(show this ???)}$$

Thus the error is

$$\hat{P}_e = \int_{-\infty}^{0} \frac{1}{(2\pi)^{1/2}} \frac{e^{-\frac{1}{2}\frac{(z-\mathbb{E}[z])^2}{\sigma^2[z]}}}{\sigma[z]} dz$$



Let $\frac{z-\mathbb{E}[z]}{\sigma[z]} \equiv y$. Notice that $y \sim N(0,1)$ according to the CLT

Then, $\hat{P}_e = \int_{-\infty}^{\frac{-\mathbb{E}[z]}{\sigma[z]}} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}y^2} dy = \int_{\frac{\mathbb{E}[z]}{\sigma[z]}}^{\infty} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}y^2}$

Notice that $\sigma^2[z] = \alpha(D/N)$, thus $\frac{\mathbb{E}[z]}{\sigma[z]} = O\left(\frac{N^{1/2}}{D^{1/2}}\right) \sum_{m=1}^{D} \frac{1}{m}$.

Now, by definition (11)

$$\lim_{D\to\infty} \left(\sum_{m=1}^{D} \frac{1}{m} - \log D \right) = \gamma \quad \text{(Euler-Mascheroni constant)}$$

Therefore,

$$\lim_{D\to\infty} \left(\frac{N}{D}\right)^{1/2} \sum_{m=1}^{D} \frac{1}{m} - \underbrace{\frac{N^{1/2}}{D} \log D}_{\to 0} = \underbrace{\left(\frac{N}{D}\right)^{1/2} \gamma}_{\to 0}$$

$\therefore \frac{\mathbb{E}[z]}{\sigma[z]} \to 0$ as $D \to \infty$, which implies that

$$\boxed{\hat{P}_e \to 1/2 \text{ as } D \to \infty} \; !$$

Comments:

(1) This shows that, for this example, if μ is known theoretically, the Bayes error $P_e \to 0$ as $D \to \infty$. However, if we estimate μ by N-size sample, the plug-in error $\hat{P}_e \to 1/2$ as $D \to \infty$.

(2) The example is of course rather artificial. Does this happe in real data? Notice that if the components of μ decay fast, for instance if $\mu_i = \frac{1}{i^{1/2+\varepsilon}}$ with $\varepsilon > 0$, the series would not diverge and $P_e \not\to 0$.

(3) In the plug-in error, we are assuming $\frac{N}{D} \ll 0$. If we have enough data, $\frac{N}{D} = O(1)$, and $\hat{P}_e < 1/2$. By increasing N, $\hat{P}_e$ should approach P_e.