---DGL - CH 2

BAYES ERROR

$(X,Y) \in \mathbb{R}^d \times \{0,1\} \sim (\mu, \eta)$. $\mu(A) = \mathbb{P}(X \in A)$

$\eta(x) = \mathbb{P}\{Y=1 | X\} = \mathbb{E}[Y|X]$. This is enough to describe $\mathbb{P}_{X,Y}$ since if $C \subseteq \mathbb{R}^d \times \{0,1\}$, then

$$C = (C \cap \mathbb{R}^d \times \{0\}) \cup (C \cap \mathbb{R}^d \times \{1\})$$
$$\equiv C_0 \times \{0\} \cup C_1 \times \{1\}$$

Thus $\mathbb{P}[(X,Y) \in C] = \mathbb{P}[X \in C_0, Y=0] + \mathbb{P}[X \in C_1, Y=1]$ (disjoint)

$$= \mathbb{P}[Y=0 | X \in C_0]\,\mathbb{P}[X \in C_0] + \mathbb{P}[Y=1 | X \in C_1]\,\mathbb{P}[X \in C_1]$$
$$= \int_{C_0} (1-\eta(x))\,\mu(dx) + \int_{C_1} \eta(x)\,\mu(dx)$$

classifier $\rightarrow$ $g : \mathbb{R}^d \to \{0,1\}$. Error is $L(g) = \mathbb{P}[g(X) \neq Y]$

Bayes classifier $\rightarrow$ $g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$

<u>Theorem</u> g^* is optimal, i.e. for any g we have that

$$\mathbb{P}[g^*(X) \neq Y] \leq \mathbb{P}[g(X) \neq Y]$$
$$L(g^*) \leq L(g)$$

<u>Proof</u>. $\mathbb{P}[g(X) \neq Y | X] = 1 - \mathbb{P}[g(X) = Y | X]$

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

Notice that $\mathbb{1}(g(X)=0) = 1 - \mathbb{1}(g(X)=1)$, thus

$$\mathbb{P}(g(X) \neq Y | X) = 1 - ((1-\eta(X)) + \mathbb{1}(g(X)=1)(2\eta(X)-1))$$
$$= \eta(X) - (2\eta(X)-1)\,\mathbb{1}(g(X)=1)$$

$$\mathbb{P}(g(X)\neq Y|X) - \mathbb{P}(g^*(X)\neq Y|X) = (2\eta(X)-1)\left(\mathbb{1}(g^*(X)=1) - \mathbb{1}(g(X)=1)\right) \quad (2)$$

If $\eta > 1/2 \Rightarrow \underbrace{(2(1/2+\delta)-1)}_{>0}\underbrace{(1-\mathbb{1}(g(X)=1))}_{\geq 0} \geq 0$

If $\eta \leq 1/2 \Rightarrow \underbrace{(2(1/2-\delta)-1)}_{\leq 0}\underbrace{(-\mathbb{1}(g(X)=1))}_{\leq 0} \geq 0$

thus in either case $\mathbb{P}(g(X)\neq Y|X) - \mathbb{P}(g^*(X)\neq Y|X) \geq 0$.
Using $\mathbb{P}(g(X)\neq Y) = \int \mathbb{P}(g(X)\neq Y|X)\,\mu(dX)$ we have the integral of a positive integrand over a positive range, so $L(g) - L(g^*) \geq 0$. □

Note that $\mathbb{P}(g(X)\neq Y) = \int \mathbb{P}(g(X)\neq Y|X)\,\mu(dX)$

$$= \int \mu(dX)\{1 - \eta(X)\mathbb{1}(g(X)=1) - (1-\eta(X))\mathbb{1}(g(X)=0)\}$$
$$= 1 - \mathbb{E}_X[\eta(X)\mathbb{1}(g(X)=1) + (1-\eta(X))\mathbb{1}(g(X)=0)]$$

Also $\eta(X) = \mathbb{P}(Y=1|X) = \mathbb{E}[Y|X]$

Suppose $Y = f(X)$ for some predictor $f: \mathbb{R}^d \to \mathbb{R}$ and consider $\mathbb{E}[(f(X)-Y)^2]$. We have $\mathbb{E}[(f(X)-Y)^2|X] = \mathbb{E}[(f-\eta+\eta-Y)^2|X]$,

$$\mathbb{E}[(f(X)-Y)^2|X] = \mathbb{E}[(f-\eta)^2 + 2(f-\eta)(\eta-Y) + (\eta-Y)^2|X]$$
$$= \underbrace{(f-\eta)^2}_{\geq 0} + 2(f-\eta)\underbrace{(\eta - \mathbb{E}[Y|X])}_{=0} + \mathbb{E}[(\eta-Y)^2|X]$$

$\mathbb{E}[(f(X)-Y)^2|X] \geq \mathbb{E}[(\eta(X)-Y)^2|X]$ or
$\mathbb{E}[(f(X)-Y)^2] \geq \mathbb{E}[(\eta(X)-Y)^2]$ // Therefore, η is a minimizer of the squared error function.

Let us show other formulas.

Since g^* is the best classifier, $L^* = \inf_g \mathbb{P}[g(X) \neq Y]$ //

- Another formula is $L^* = \mathbb{E}\{\min(\eta(X), 1-\eta(X))\}$. //

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

Now $L(g) = \mathbb{P}[g(X) \neq Y] = \int \mathbb{P}[g(x) \neq Y | X]\,\mu(dX) = \mathbb{E}[\mathbb{P}[g(X) \neq Y | X]]$
So $L^* = \mathbb{E}[\mathbb{1}(\eta \leq 1/2)\,\eta(X) + \mathbb{1}(\eta > 1/2)(1-\eta(X))]$

Now observe that
if $\eta \leq 1/2 \Rightarrow 1-\eta \geq 1/2 \geq \eta$
if $\eta > 1/2 \Rightarrow 1-\eta < 1/2 < \eta$

So if $\eta \leq 1-\eta$ we pick the first term $L^* = \mathbb{E}[\eta]$, and if $1-\eta < \eta$ we pick the second $L^* = \mathbb{E}[1-\eta]$. In other words, we pick the minimum between the two. □

- Another formula is $L^* = \frac{1}{2} - \frac{1}{2}\mathbb{E}[|2\eta(X) - 1|]$ //

Proof we showed $L^* = \mathbb{E}\{\min(\eta, 1-\eta)\}$.

Suppose $\eta > 1/2$. Then $L^* = \mathbb{E}\{1-\eta\} = \mathbb{E}\{1/2 - \frac{1}{2}(2\eta - 1)\}$
$= \mathbb{E}\{1/2 - \frac{1}{2}|2\eta - 1|\}$
$= \frac{1}{2} - \frac{1}{2}\mathbb{E}\{|2\eta - 1|\}$.

Suppose $\eta \leq 1/2$. Then $L^* = \mathbb{E}\{\eta\} = \mathbb{E}\{1/2 - \frac{1}{2}(-2\eta + 1)\}$
$= \mathbb{E}\{1/2 - \frac{1}{2}|2\eta - 1|\}$
$= \frac{1}{2} - \frac{1}{2}\mathbb{E}\{|2\eta - 1|\}$ □

If X has density f we have

$$L^* = \int \min(\eta(x), 1-\eta(x)) f(x)\,dx$$

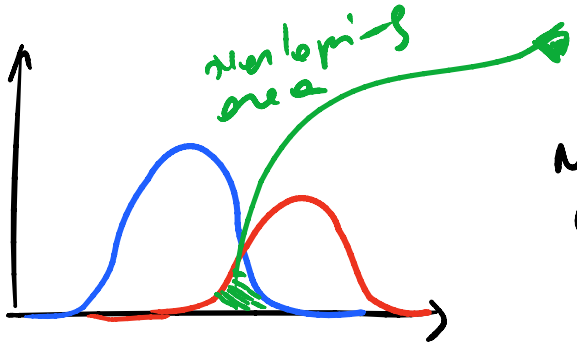
Now note that $\eta(x) = \mathbb{P}(Y=1|X) = \dfrac{p(x|Y=1)\,\mathbb{P}(Y=1)}{p(x)}$

$$\equiv \frac{f_1(x)\,p}{f(x)}$$

Analogously, $1-\eta(x) = \mathbb{P}(Y=0|X) = \dfrac{p(x|Y=0)\,\mathbb{P}(Y=0)}{p(x)}$

$$\equiv \frac{f_0(x)(1-p)}{f(x)}$$

Replacing this we have, $L^* = \int \min(pf_1(x), (1-p)f_0(x))\,dx$



If f_1 and f_0 do not overlap, then $L^* = 0$.

Suppose $p = 1/2$. $L^* = \frac{1}{2}\int \min(f_1, f_0)\,dx$

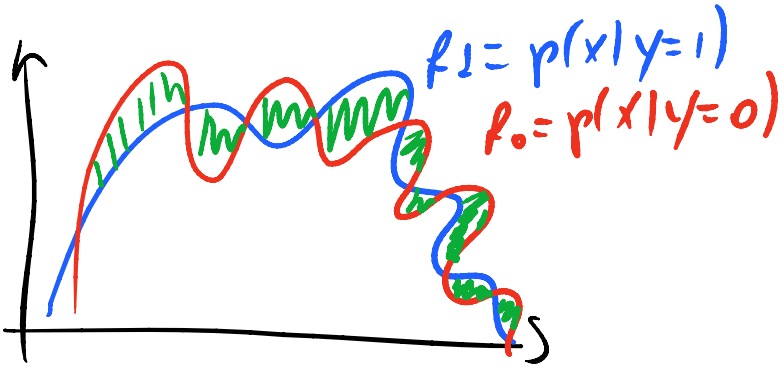
if $f_1 > f_0$: $f_1 - (f_1 - f_0) = f_0$

if $f_1 < f_0$: $f_1 - \underbrace{(f_1 - f_0)_+}_{=0} = f_1$ ← positive part

$$\min(f_1, f_0) = f_1 - (f_1 - f_0)_+$$
$$= f_1 - \frac{1}{2}|f_1 - f_0|$$

$$L^* = \frac{1}{2}\underbrace{\int f_1\,dx}_{=1} - \frac{1}{4}\int |f_1 - f_0|\,dx$$

$$= \frac{1}{2} - \frac{1}{4}\int |f_1 - f_0|\,dx$$ ← L_1 distance between the class densities.



Plugin Decision

$$g^*(x) = \begin{cases} 0 & \text{if } \eta(x) \le 1/2 \ (\text{or } 1-\eta \ge \eta) \\ 1 & \text{otherwise.} \end{cases}$$

η is unknown. Suppose we have an approximation $\tilde{\eta}(x)$. Then we define

$$g(x) = \begin{cases} 0 & \text{if } \tilde{\eta}(x) \le 1/2 \\ 1 & \text{otherwise} \end{cases}$$

Theo. $L(g) - L(g^*) = 2\int_{\mathbb{R}^d} |\eta - 1/2|\, \mathbb{1}(g(x) \neq g^*(x))\, \mu(dx)$

$$\le 2\int_{\mathbb{R}^d} |\eta - \tilde{\eta}|\, \mu(dx) = 2\mathbb{E}(|\eta - \tilde{\eta}|)$$

So if $\tilde{\eta}$ is close to η in L_1 sense, L is close to L^*.

proof
the call + hat

$$\mathbb{P}(g(X)\neq Y|X) - \mathbb{P}(g^*(X)\neq Y|X) = (2\eta(X)-1)\left(\mathbb{1}(g^*(X)=1) - \mathbb{1}(g(X)=1)\right)$$

$$= |2\eta(X)-1|\,\mathbb{1}(g(X)\neq g^*(X))$$

This gives the first equality upon integration.
If $g \neq g^*$ then $|\tilde{\eta}-\eta| \geq |\eta - 1/2|$. Indeed:

1. $g^*=0,\ g=1 \Rightarrow \eta \leq 1/2,\ \tilde{\eta} > 1/2$
2. $g^*=1,\ g=0 \Rightarrow \eta > 1/2,\ \tilde{\eta} \leq 1/2$

Case 1: $\eta = 1/2 - \delta,\ \tilde{\eta} = 1/2 + \tilde{\delta} \therefore$ $|\eta - 1/2| = \delta$, $|\tilde{\eta}-\eta| = \delta + \tilde{\delta}$]

Case 2: $\eta = 1/2 + \delta,\ \tilde{\eta} = 1/2 - \tilde{\delta} \therefore$ $|\eta - 1/2| = \delta$, $|\tilde{\eta}-\eta| = \delta + \tilde{\delta}$] □

Suppose $\tilde{\eta}_1(x) \approx \eta(x)$, $\tilde{\eta}_0(x) \approx 1-\eta(x)$ (approximations).
Define $g(x) = \begin{cases} 0 \text{ if } \tilde{\eta}_1 \leq \tilde{\eta}_0. \\ 1 \text{ otherwise} \end{cases}$

We don't have necessarily $\tilde{\eta}_1 + \tilde{\eta}_0 = 1$. Thus the previous result differs slightly:

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

Recall that $\eta(x) = \mathbb{P}(Y=1|X) = \frac{p(x|Y=1)\,\mathbb{P}(Y=1)}{p(x)} = \frac{f_1 p}{f}$

$1-\eta(x) = \mathbb{P}(Y=0|x) = \ldots = \frac{f_0(1-p)}{f}$

Let $\tilde{f}_1 \approx f_1$, $\tilde{p}_1 \approx p$, $\tilde{f}_0 \approx f_0$, $\tilde{p}_0 \approx 1-p$.

Define $g(x) = \begin{cases} 0 \text{ if } \tilde{p}_1 \tilde{f}_1 \leq \tilde{p}_0 \tilde{f}_0 \\ 1 \text{ otherwise.} \end{cases}$

Then from the previous theorem we have

$$|(1-\eta) - \tilde{\eta}_0| = \left|\frac{(1-p)f_0 - \tilde{p}_0 \tilde{f}_0}{f}\right|$$

$$|\eta - \tilde{\eta}_1| = \left|\frac{p f_1 - \tilde{p}_1 \tilde{f}_1}{f}\right|$$

$\mu(dx) = f(x)\,dx$

So $L - L^* \leq \int_{\mathbb{R}^d} |(1-p)f_0 - \tilde{p}_0 \tilde{f}_0|\,dx + \int_{\mathbb{R}^d} |p f_1 - \tilde{p}_1 \tilde{f}_1|\,dx$

Role of dimension: If we knew $\eta(x)$ the prob. would be 1D since we could just forget about X altogether.

However, η is unknown. In general, every transformation $X \to T(X)$ increases the Bayes risk, since it destroys information. Nevertheless, there are transformations that leaves the Bayes error untouched.