The oracle or optimal discriminator $D^*$ reminds me of bayesian updating.

Let's say we have a prior $p = \Pr(x \in R)$ and we observe $x_i = m$ what is the posterior probability $D(x_i) = \Pr(x_i \in R | x_i = m)$? We also know that when $x \in R$, it follows a distribution $p_r(x)$ and when $x \in G$, it follows a distribution $p_g(x)$. The posterior probability is given by Bayes' theorem:

$$D(x_i) = \frac{\Pr(x_i = m | x_i \in R)p}{\Pr(x_i = m | x \in R)p + \Pr(x_i = m | x \in G)(1-p)}$$

Since we have a uniform prior $p = 1/2$, then

$$D^*(x_i) = \frac{p_r(x_i)}{p_r(x_i) + p_g(x_i)}$$

This is the same as

$$D^* = \max_D E_{p_r}[\log D(x)] + E_{p_g}[\log(1 - D(x))]$$

How about bayesian estimator?

The same setting as before. A prior $p = \Pr(x \in R)$, and two distribution $p_r(x)$ and $p_g(x)$. We observe a sequence of $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and we want to classify them into $R$ or $G$ correctly (discriminator). Let $t_i \in \{R, G\}$ be the true label of $x_i$. We denote the loss function as $L(x, t)$. For data coming from $R$, the loss is $L(x, R)$. Therefore, in expectation, a data point from group $R$ has a loss

$$E_{p_r}[L(x, R)] = \int L(x, R) p_r(x) dx$$

Similary for a data point from group $G$.

$$E_{p_g}[L(x, G)] = \int L(x, G) p_g(x) dx$$

However, for each data point, we don't know the true label. But we hold a prior that each data point has a $p$ probability of coming from $R$. Therefore, for a given data point $x_i$, the expected loss is

$$\Pr(x_i \in R) E_{p_r}[L(\delta(x_i), R)] + \Pr(x_i \in G) E_{p_g}[L(\delta(x_i), G)]$$

To write it more compactly,

$$\begin{aligned} E(L(x_i, t_i)) &= E[L(\delta(x_i), t_i) | t_i = R] \Pr(t_i = R) + E[L(\delta(x_i), t_i) | t_i = G] \Pr(t_i = G) \\ &= E_{p_r}[L(\delta(x_i), R)]p + E_{p_g}[L(\delta(x_i), G)](1-p) \end{aligned}$$

This is the expected loss for a single data point, without knowing the true label.

Notice that this is equivalent to the discriminator's objective function if we set $p = 1/2$ and $L(\delta(x_i), R) = \log(D(x_i))$, $L(\delta(x_i), G) = \log(1 - D(x_i))$.

Now let us compare $D$ and $\delta$. First we define that for a single data point, the loss is

$$L(\delta(x_i), t_i) = \delta(x_i) 1_{t_i = R} + \tau(1 - \delta(x_i))(1 - 1_{t_i = R})$$

where $\delta(x_i)$ is the decision function, the classifier that maps $x_i$ to $R$ or $G$. Unlike $D(x_i) \in [0, 1]$, the small $\delta(x_i) \in \{0, 1\}$.

If $x_i$ is from $R$, - with $D(x_i) = 0.8$, the loss is

$$\log(D(x_i)) = \log(0.8)$$

- with $\delta(x_i)$, the loss is

$$L(\delta(x_i), R) = \delta(x_i)$$

In this case, the difference between wrong and right classification is 1. A gain of 1 if you classify correctly. With respect to $D$, you gain $\log(0.8) - \log(0.5) > 0$ by saying that it is more likely from $R$.

if $x_i$ is from $G$, - with $D(x_i) = 0.8$, the loss is

$$\log(1 - D(x_i)) = \log(0.2)$$

- with $\delta(x_i)$, the loss is

$$L(\delta(x_i), G) = \tau(1 - \delta(x_i))$$

A loss of $\tau$ if you don't classify correctly. With respect to $D$, you lose $|\log(0.2) - \log(0.5) < 0|$ by saying that it is more likely from $R$.

We are both minimizing bayesian risk! The difference is the loss function and the prior!

Notice that for discriminator, in reality

$$D^* = \max_D \frac{1}{n} \sum_{i=1}^{n} \log D(x_i) + \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(x_i))$$

where $n$ is the number of data points from $R$ and $m$ is the number of data points from $G$. If $n \neq m$, then we are weighting the point from $R$ and $G$ differently. While in maximum likelihood, we are taking weight $1/(n+m)$ for each point.

$$D^* = \max_D \sum_{i=1}^{n+m} t_i \log D(x_i) + (1 - t_i) \log(1 - D(x_i))$$

where $t_i = 1$ if $x_i$ is from $R$ and $t_i = 0$ if $x_i$ is from $G$.

For the discriminator, if we have a prior $p = \frac{n}{n+m}$, then

$$D^* = \max_D \frac{1}{n} \sum_{i=1}^{n} \log D(x_i) \frac{n}{m+n} + \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(x_i)) \frac{m}{m+n}$$

Then every point has the same weight. The function is equivalent to the ML.

If we start with equal wight ($n < m$), then we need to add preference (prior?) for real point $\frac{m}{n+m}$ such that we recover the discriminator's objective function.

Derive the optimal $\delta^*$, the bayesian decision rule.

$$\max_\delta pE_{p_r}[\delta(x)] + (1-p)E_{p_g}[\tau(1 - \delta(x))]$$

$$\max_\delta \int p\delta(x)p_r(x)dx + (1-p)\tau(1 - \delta(x))p_g(x)dx$$

First assuming $p = 1/2, \tau = 1$, then the optimal decision rule takes the form

$$\delta^*(x) = 1\{p_r(x) > p_g(x)\}$$

as simple as it can be. This is equivalent to

$$\delta^*(x) = 1\{D^*(x) > 0.5\}$$

since

$$D^*(x) = \frac{p_r(x)}{p_r(x) + p_g(x)} > 1/2 \Leftrightarrow p_r(x) > p_g(x)$$

Others stuff: - Let me think about the form of empirical distribution $G_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1x_i \leq x$ and empirical expectation $E_{G_n}(f(x)) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$ again. - How to show that posterior tail probability $P(\theta > \theta_\alpha | Y = y)$ is a monotonically increasing in $Y$ and decreasing in $\theta_\alpha$? Define

$$P(\theta > \theta_\alpha | Y = y) = \frac{P(\theta > \theta_\alpha, Y = y)}{P(Y = y)}$$

$$= \frac{\int_{\theta_\alpha}^{\infty} p(\theta, y) d\theta}{\int p(\theta, y) d\theta} = \frac{\int_{\theta_\alpha} p(y|\theta) p(\theta) d\theta}{\int p(y|\theta) p(\theta) d\theta}$$

Or one can write the Bayes' rule

$$\frac{\Pr(\theta > \theta_\alpha) \Pr(Y = y | \theta > \theta_\alpha)}{\Pr(\theta > \theta_\alpha) \Pr(Y = y | \theta > \theta_\alpha) + \Pr(\theta \leq \theta_\alpha) \Pr(Y = y | \theta \leq \theta_\alpha)}$$

where

$$\Pr(Y = y | \theta > \theta_\alpha) = \frac{\int_{\theta_\alpha}^{\infty} p(y|\theta) p(\theta) d\theta}{\Pr(\theta > \theta_\alpha)}$$