Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 2 — August 23

Instructor: Stas Minsker *Scribe: Mose Wintner*

2.1 Binary Classification, continued

Goal: Predict Y from X.

$$g : \mathbb{S} \to \{\pm 1\}$$
$$L(g) = \mathbb{E}\left(I\{Y \neq g(X)\}\right) = P(Y \neq g(X)) = P((x,y) \in S \times \{\pm 1\} : y \neq g(x)).$$

The **Bayes classifier** g_* minimizes the loss, i.e.

$$g_* = \underset{\substack{g: S \to \{\pm 1\} \\ g \text{ measurable}}}{\operatorname{argmin}} L(g).$$

To compute it we need the conditional expectation

$$\eta(x) := \mathbb{E}(Y|X=x).$$

In the case of binary classifiation,

$$\eta(x) = 1 \cdot P(Y=1|X=x) - 1 \cdot P(Y=-1|X=x).$$

Since $P(Y=1|X=x) + P(Y=-1|X=x) = 1$, we have that

$$P(Y=1|X=x) = \frac{1+\eta(x)}{2}$$
$$P(Y=-1|X=x) = \frac{1-\eta(x)}{2}.$$

In other words,

$$P(Y=t) = \frac{1+t\eta(x)}{2}.$$

This suggests $g_*(x) = 1$ if and only if $\frac{1+\eta(x)}{2} > \frac{1}{2}$, or $\eta(x) > 0$. Luckily, in the case of binary classification we can compute the Bayes classifier directly.

$$\begin{aligned}
L(g) &= \int_{\mathbb{S}} P(Y \neq g(x)|X=x)\, d\Pi(x) \\
&= \int (1 - P(Y = g(x)|X=x))\, d\Pi(x) \\
&= \int \frac{1 - g(x)\eta(x)}{2}\, d\Pi(x) \\
&\geq \int \frac{1 - |\eta(x)|}{2}\, d\Pi(x).
\end{aligned}$$

We see that the loss is minimized when equality is achieved, i.e. the sign of $g(x)$ agrees with the sign of $\eta(x)$, which confirms our guess.

Example 1. Treat Y as a "parameter".

$$\begin{aligned}\Pi_+ &= Prob(Y=1)\\ \Pi_- &= Prob(Y=-1)\\ \Pi_+(dx) &= P(X\in dx|Y=1) = p_+(x)\\ \Pi_-(dx) &= P(X\in dx|Y=-1) = p_-(x).\end{aligned}$$

Bayes formula:

$$P(Y=1|X=x) = \frac{\Pi_+p_+(x)}{\Pi_+p_+(x)+\Pi_-p_-(x)}P(Y=-1|X=x) \quad = \frac{\Pi_-p_-(x)}{\Pi_+p_+(x)+\Pi_-p_-(x)}$$

Bayes classifier:

$$\begin{aligned}P(Y=1|X=x) > P(Y=-1|X=x) &\Rightarrow \text{predict } 1.\\ &\Leftrightarrow \Pi_+p_+(x) > \Pi_-p_-(x)\\ &\Leftrightarrow \frac{p_+(x)}{p_-(x)} > \frac{\Pi_-}{\Pi_+}.\end{aligned}$$

Exercise: Assume that

$$p_+(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det\Sigma}}\exp\left(-\frac{\langle\Sigma^{-1}(x-\mu_+),(x-\mu_+)\rangle}{2}\right)$$

and similarly for p_- but with μ_- instead of μ_+. Compute the Bayes classifier.

Definition 1. The **excess risk** fo a classifier g is defined to be

$$\mathcal{E}(g) = L(g) - L(g_*) \geq 0.$$

In binary classification,

$$\begin{aligned}\mathcal{E}(g) &= \int_{\mathbb{S}} \frac{1-g(x)\eta(x)}{2} - \frac{1-\text{sgn}(\eta(x))\eta(x)}{2}\,d\Pi\\ &= \int_{g(x)\neq\text{sgn}(\eta(x))} |\eta(x)|\,d\Pi.\end{aligned}$$

In practice, the distribution of (X,Y) is unknown, so computation of the Bayes classifier is impossible. A naïve approach is to make an estimate $\widehat{\eta}(x)$ of the conditional probability based on training data $(X_j,Y_j)_{j=1}^n$ and use the estimate

$$\widehat{g}_*(x) = \text{sign}(\widehat{\eta}(x)).$$

We calculate the excess risk for such a naïve estimate.

$$\begin{aligned}
\mathcal{E}(g) &= \int_{\{\mathrm{sgn}(\widehat{\eta})=\mathrm{sgn}(\eta)\}} |\eta(x)|\, d\Pi(x) \\
&\leq \int_{\mathbb{S}} |\eta(x) - \widehat{\eta}(x)|\, d\Pi(x) \\
&\leq \left(\int_{\mathbb{S}} (\eta(x) - \widehat{\eta}(x))^2\, d\Pi(x) \right)^{1/2}
\end{aligned}$$

by Cauchy-Schwarz.