Machine Learning I

Supervised Learning: Optimal Predictions

Souhaib Ben Taieb

University of Mons

Table of contents

Optimal prediction function

$$\boxed{f = \underset{h:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}}\ E_{\text{out}}(h)} \quad g^* = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ E_{\text{out}}(h), \quad g = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ E_{\text{in}}(h)$$

Recall that the **optimal prediction function** is given by

$$f = \underset{h:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}}\ \mathbb{E}_{x,y}[L(y, h(x))], \tag{1}$$

where $L(\cdot,\cdot)$ is the loss function.

Using the **law of iterated expectation**, we can write

$$\mathbb{E}_{x,y}[L(y, h(x))], = \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[L(y, h(x))|x\right]\right].$$

It sufficed to minimize the error **pointwise**, i.e. compute

$$f(x) = \underset{h(x)\in\mathcal{Y}}{\operatorname{argmin}}\ \mathbb{E}_{y|x}[L(y, h(x))|x], \tag{2}$$

for all $x \in \mathcal{X}$.

Table of contents

Optimal predictions with the squared error loss

Let $\mathcal{Y} \subseteq \mathbb{R}$. With the squared error loss function $L(y, \hat{y}) = (y - \hat{y})^2$, the expected error at x can be decomposed as follows:

$$
\begin{aligned}
\mathbb{E}[L(y, h(x))|x] &= \mathbb{E}[(y - h(x))^2|x] \\
&= \mathbb{E}[y^2 - 2yh(x) + h(x)^2|x] \\
&= \mathbb{E}[y^2|x] - 2h(x)\mathbb{E}[y|x] + h(x)^2 \\
&= \mathrm{Var}(y|x) + (\mathbb{E}[y|x])^2 - 2h(x)\mathbb{E}[y|x] + h(x)^2 \\
&= \mathrm{Var}(y|x) + (\mathbb{E}[y|x] - h(x))^2
\end{aligned}
$$

- The first term corresponds to the inherent unpredictability, or noise of y, and is called the **Bayes error**. It is the smallest error any learning algorithm can achieve.
- The second term is non-negative, and will be equal to zero if $h(x) = \mathbb{E}[y|x]$.

Optimal predictions with the squared error loss

In summary, the optimal prediction at x is given by

$$f(x) = \operatorname*{argmin}_{h(x) \in \mathcal{Y}} \mathbb{E}[(y - h(x))^2 | x] \tag{3}$$

$$= \mathbb{E}[y|x], \tag{4}$$

i.e. the conditional expectation, also known as the **regression function**.

In other words, when *best is measured by expected squared error*, the best prediction for y at any point x is the **conditional expectation** at x.

Table of contents

Optimal predictions with the zero-one loss

Consider a multi-class classification problem with K categories where $\mathcal{Y} = \{C_1, \ldots, C_K\}$. With the zero-one loss $L(y, \hat{y}) = \mathbb{1}\{y \neq \hat{y}\}$, the expected error (error rate) at x can be decomposed as follows:

$$
\begin{aligned}
\mathbb{E}[L(y, h(x))|x] &= \mathbb{E}\left[\mathbb{1}\{y \neq h(x)\}|x\right] \\
&= \mathbb{P}(y \neq h(x)|x) \\
&= 1 - \mathbb{P}(y = h(x)|x).
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
f(x) &= \underset{h(x)\in\mathcal{Y}}{\operatorname{argmin}}\ \mathbb{E}[\mathbb{1}\{y \neq h(x)\}|x] \\
&= \underset{h(x)\in\mathcal{Y}}{\operatorname{argmin}}\ 1 - \mathbb{P}(y = h(x)|x) \\
&= \underset{h(x)\in\mathcal{Y}}{\operatorname{argmax}}\ \mathbb{P}(y = h(x)|x).
\end{aligned}
$$

Optimal predictions with the zero-one loss

In summary, the optimal prediction at x is given by

$$f(x) = \underset{h(x)\in\mathcal{Y}}{\operatorname{argmin}}\ \mathbb{E}[\mathbb{1}\{y \neq h(x)\}|x] \quad (5)$$

$$= \underset{h(x)\in\mathcal{Y}}{\operatorname{argmax}}\ \mathbb{P}(y = h(x)|x). \quad (6)$$

This optimal classifier is called the **Bayes classifier**, and has the following expected error (error rate) at x:

$$1 - \max_{k=1,\dots,K} \mathbb{P}(y = \mathcal{C}_k|x),$$

also called the **Bayes error rate**, which gives the lowest possible error rate that could be achieved <u>if we knew</u> $p_{y|x}$.

Table of contents

Optimal predictions

Consider a binary classification problem where $\mathcal{Y} = \{0, 1\}$ and the general binary classification loss function:

$$L(y, \hat{y}) = \begin{cases} L(0,0) & \text{if } y = 0 \text{ and } \hat{y} = 0; \\ L(0,1) & \text{if } y = 0 \text{ and } \hat{y} = 1; \\ L(1,0) & \text{if } y = 1 \text{ and } \hat{y} = 0; \\ L(1,1) & \text{if } y = 1 \text{ and } \hat{y} = 1, \end{cases}$$

where we assume $L(1,0) > L(1,1) \geq 0$ and $L(0,1) > L(0,0) \geq 0$.

The zero-one loss is a particular case where $L(0,0) = L(1,1) = 0$ and $L(1,0) = L(0,1) = 1$.

The optimal prediction at x is given by

$$f(x) = \underset{h(x) \in \{0,1\}}{\text{argmin}} \; \mathbb{E}[L(y, h(x))|x].$$

Let us consider the two cases: $h(x) = 0$ and $h(x) = 1$.

Optimal predictions

We can expand the expected loss for each of the two possible predictions.

$h(x) = 0$:

$$\mathbb{E}[L(y,0)|x] = L(0,0)\mathbb{P}(y=0|x) + L(1,0)\mathbb{P}(y=1|x).$$

$h(x) = 1$:

$$\mathbb{E}[L(y,1)|x] = L(0,1)\mathbb{P}(y=0|x) + L(1,1)\mathbb{P}(y=1|x).$$

Since we want to minimize the expected loss, the optimal prediction is $h(x) = 1$ ($h(x) = 0$) whenever the second expression is smaller (larger) than the first.

Optimal predictions

In other words, the optimal prediction at x is given by

$$
\begin{aligned}
f(x) &= \mathbb{1}\{\mathbb{E}[L(y,1)|x] \leq \mathbb{E}[L(y,0)|x]\} \\
&= \mathbb{1}\left\{\mathbb{P}(y=1|x) \geq \frac{L(0,1)-L(0,0)}{L(1,0)-L(1,1)}\mathbb{P}(y=0|x)\right\} \\
&= \mathbb{1}\left\{\mathbb{P}(y=1|x) \geq \frac{L(0,1)-L(0,0)}{L(0,1)-L(0,0)+L(1,0)-L(1,1)}\right\}
\end{aligned}
$$

Examples

Consider the following three cases:

- $L(0,0) = 0, L(0,1) = 1, L(1,0) = 1, L(1,1) = 0$ (zero-one loss)

  $$f(x) = \mathbb{1}\{\mathbb{P}(y=1|x) \geq \mathbb{P}(y=0|x)\} = \mathbb{1}\left\{\mathbb{P}(y=1|x) \geq \frac{1}{2} = 0.5\right\}$$

- $L(0,0) = 0, L(0,1) = 1, L(1,0) = 10, L(1,1) = 0$

  $$\begin{aligned} f(x) &= \mathbb{1}\left\{\mathbb{P}(y=1|x) \geq \frac{1}{10} \times \mathbb{P}(y=0|x)\right\} \\ &= \mathbb{1}\left\{\mathbb{P}(y=1|x) \geq \frac{1}{11} \approx 0.09\right\} \end{aligned}$$

- $L(0,0) = 0, L(0,1) = 1000, L(1,0) = 1, L(1,1) = 0$

  $$\begin{aligned} f(x) &= \mathbb{1}\{\mathbb{P}(y=1|x) \geq 1000 \times \mathbb{P}(y=0|x)\} \\ &= \mathbb{1}\left\{\mathbb{P}(y=1|x) \geq \frac{1000}{1001} \approx 0.99\right\} \end{aligned}$$