Bayes Classifiers

Matthieu R. Bloch

1 Bayes classifier

For ease of notation, let us revisit our learning model with a slight change in notation to clearly indicate the random variables. Our supervised learning problem consists of:

1. A dataset $\mathcal{D} \triangleq \{(X_1, Y_1), \cdots, (X_N, Y_N)\}$
   - $\{X_i\}_{i=1}^N$ drawn i.i.d. from an unknown probability distribution P_X on $\mathcal{X}$;
   - $\{Y_i\}_{i=1}^N$ with $\mathcal{Y} = \{0, 1, \cdots, K\}$.
2. An a priori unknown labeling probability $P_{Y|X}$
3. A *binary* loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+ : (y_1, y_2) \mapsto \mathbb{1}\{y_1 \neq y_2\}$.

Since our goal is to characterize the minimum true risk, we need to specify a class of hypotheses $\mathcal{H}$ at this point. Note that the (true) risk of a classifier h is

$$R(h) \triangleq \mathbb{E}_{XY}(\mathbb{1}\{h(X) \neq Y\}) = \mathbb{P}_{XY}(h(X) \neq Y) \tag{1}$$

To estimate the smallest risk that we can ever hope to achieve, we assume for now that we *know* P_X and $P_{Y|X}$. This is not a realistic assumption since the whole point of learning is to figure out what $P_{Y|X}$ is and P_X might never be learned at all; however, the risk of any realistic classifier can certainly be no less than the risk of the best classifier that knows P_X and $P_{Y|X}$, which can therefore serve as the ultimate benchmark of performance. For notational convenience, we introduce the following:

- the *a priori* class probabilities are denoted $\pi_k \triangleq \mathbb{P}_Y(k)$.
- the *a posteriori* class probabilities are denoted $\eta_k(x) \triangleq \mathbb{P}_{Y|X}(k|x)$ for all $x \in \mathcal{X}$.

Lemma 1.1. *The classifier $h^B(\mathbf{x}) \triangleq argmax_{k\in[0;K-1]}\, \eta_k(\mathbf{x})$ is optimal, i.e., for any classifier h, we have $R(h^B) \leqslant R(h)$. In addition*

$$R(h^B) = \mathbb{E}_X\left(1 - \max_k \eta_k(X)\right)$$

Proof. For a classifier h and for each $0 \leqslant k \leqslant K-1$, let us define the corresponding decision region $\Gamma_k(h) \triangleq \{x : h(x) = k\}$. Then note that

$$1 - R(h) = \mathbb{P}(h(X) = Y) = \sum_{k=0}^{K-1} \pi_k \mathbb{P}(h(X) = k|Y = k) = \sum_{k=0}^{K-1} \int_{\Gamma_k(h)} \pi_k p_{X|Y}(x|k)dx. \tag{2}$$

To minimize the risk, we should maximize (2). The expression is maximum when the regions are such that $\pi_k p_{X|Y}(x|k)$ takes the maximum possible value (over the K possibilities) in the region $\Gamma_k(h)$. Said differently, the region $\Gamma_k(h)$ must be defined as

$$\Gamma_k(h) = \{x \in \mathcal{X} : \forall \ell \in [\![0, K-1]\!]\, \pi_\ell p_{X|Y}(x|\ell) \leqslant \pi_k p_{X|Y}(x|k)\}. \tag{3}$$

The case of equality can be broken arbitrarily. The classifier leading to these decision regions is therefore

$$h^B(x) = \operatorname*{argmax}_k \pi_k p_{X|Y}(x|k) = \operatorname*{argmax}_k \eta_k(x) p_X(x) = \operatorname*{argmax}_k \eta_k(x). \tag{4}$$

The risk associated with h^B is then

$$\begin{aligned} R_B = \mathbb{E}_{XY}\left(\mathbb{1}\{h^B(X) \neq Y\}\right) &= 1 - \mathbb{E}_{XY}\left(\mathbb{1}\{h^B(X) = Y\}\right) &(5)\\ &= 1 - \mathbb{E}_{XY}\left(\mathbb{1}\left\{Y = \operatorname*{argmax}_k \eta_k(X)\right\}\right) &(6)\\ &= 1 - \mathbb{E}_X\left(\max_k \eta_k(X)\right). &(7)\end{aligned}$$

In the last step, we have used that

$$\begin{aligned} \mathbb{E}_{XY}\left(\mathbb{1}\left\{Y = \operatorname*{argmax}_k \eta_k(X)\right\}\right) &= \mathbb{E}_X\left(\sum_y P_{Y|X}(y|X)\mathbb{1}\left\{y = \operatorname*{argmax}_k \eta_k(X)\right\}\right)\\ &= \mathbb{E}_X\left(P_{Y|X}(\operatorname*{argmax}_k \eta_k(X)|X)\right)\\ &= \mathbb{E}_X\left(\max_k P_{Y|X}(k|X)\right).\end{aligned}$$

Note that we are implicitly assuming that ties have been broken with some arbitrary but fixed choice when defining the argmax. ■

The classifier h^B is called the *Bayes classifier* and $R_B \triangleq R(h^B)$ is called the *Bayes risk*.

2 Alternative forms of the Bayes classifier

You might have encountered several different forms of the Bayes classifier.

- $h^{\text{B}}(\mathbf{x}) \triangleq \operatorname{argmax}_{k\in[0;K-1]} \eta_k(\mathbf{x})$
- $h^{\text{B}}(\mathbf{x}) \triangleq \operatorname{argmax}_{k\in[0;K-1]} \pi_k p_{X|Y}(\mathbf{x}|k)$
- For $K = 2$ (binary classification), the Bayes classifier can be expressed as a log-likelihood ratio test
$$\log \frac{p_{X|Y}(\mathbf{x}|1)}{p_{X|Y}(\mathbf{x}|0)} \gtrless \log \frac{\pi_0}{\pi_1}$$
- If all classes are equally likely $\pi_0 = \pi_1 = \cdots = \pi_{K-1}$
$$h^{\text{B}}(\mathbf{x}) \triangleq \operatorname*{argmax}_{k\in[0;K-1]} p_{X|Y}(\mathbf{x}|k)$$

Example 2.1. *Assume $X|Y = 0 \sim \mathcal{N}(0,1)$ and $X|Y = 1 \sim \mathcal{N}(1,1)$. Let us compute the Bayes risk for $\pi_0 = \pi_1$. From Lemma 1.1, we have*

$$R_B = 1 - \mathbb{E}_X\left(\max_k \eta_k(X)\right) \tag{8}$$

$$= 1 - \int_{-\infty}^{\infty} p_X(x) \max_k \eta_k(x) dx \tag{9}$$

$$= 1 - \int_{-\infty}^{\infty} \max_k p_{X|Y}(x|k)\pi_k dx \tag{10}$$

$$= 1 - \frac{1}{2}\int_{-\infty}^{\frac{1}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx - \frac{1}{2}\int_{\frac{1}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} dx \tag{11}$$

$$= \frac{1}{2}(1 - \Phi(\tfrac{1}{2})) + \frac{1}{2}\int_{-\infty}^{\frac{1}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} dx \tag{12}$$

$$= \frac{1}{2}\Phi(-\tfrac{1}{2})) + \frac{1}{2}\int_{-\infty}^{-\frac{1}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv \tag{13}$$

$$= \Phi(-\tfrac{1}{2}), \tag{14}$$

where we have made use of $\Phi \triangleq$ *Normal CDF.*

In practice we do not know P_X and $P_{Y|X}$, so what is the use of the Bayes classifier? A natural, but not always wise, solution consists in using *plugin methods*, in which we use the data to learn the distributions and plug the estimates in the corresponding Bayes classifier. We will see examples of such methods in the next lecture.

3 Beyond the binary loss function

The previous discussion extends beyond the binary loss function. Given a valid loss $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}^+$, the risk of a hypothesis h is $R(h) \triangleq \mathbb{E}_{XY}(\ell(h(X), Y))$. Following the reasoning of the proof of Lemma 1.1, we can derive the Bayes classifier as follows.

$$R(h) = \sum_x \sum_y p_{X,Y}(x, y)\ell(h(x), y) \tag{15}$$

$$= \sum_k \sum_{x\in\Gamma_k(h)} \sum_m p_{X|Y}(x|m)\pi_m \ell(k, m). \tag{16}$$

Hence the Bayes's classifier is then

$$h^B(x) = \operatorname*{argmin}_k \left(\sum_m \pi_m p_{X|Y}(x|m)\ell(k, m) \right). \tag{17}$$

Without additional assumptions, this expression does not simplify much further. The elegant form obtained in Lemma 1.1 is largely the consequence of using a binary loss function.

4 To go further

A discussion of the Bayes classifier can be found in [1, Section 2.4].

References

[1] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer series in statistics. Springer, 2009.