Bayesian Classification

Aleksandar Dogandžić

April 15, 2017

Contents

Reading: [Hero 2015, §7.3], [Van Trees et al. 2013, §2].

Bayes' Rule for Testing Multiple Hypotheses

Choose a parameter-space partitioning with $M > 2$ partitions:

$$\bigcup_{m=1}^{M} \mathrm{sp}_\Theta(m) = \mathrm{sp}_\Theta, \qquad \mathrm{sp}_{\Theta_i} \cap \mathrm{sp}_{\Theta_j} = \emptyset \qquad \forall i \neq j$$

depicted in Fig. 1. We wish to distinguish among $M > 2$ hypotheses, i.e., identify which hypothesis is true:

$$\begin{aligned}
\mathcal{H}_1 : &\quad \Theta \in \mathrm{sp}_\Theta(1) \qquad \text{versus} \\
\mathcal{H}_2 : &\quad \Theta \in \mathrm{sp}_\Theta(2) \qquad \text{versus} \\
&\quad \vdots \qquad\qquad\qquad \text{versus} \\
\mathcal{H}_M : &\quad \Theta \in \mathrm{sp}_\Theta(M)
\end{aligned}$$

and, consequently, our action space consists of M choices. We design a decision rule $\phi(x) : \mathcal{X} \to (1, 2, \ldots, M)$:

$$\phi(x) = \begin{cases} 1, & \text{decide } \mathcal{H}_1 \\ 2, & \text{decide } \mathcal{H}_2 \\ \vdots & \\ M, & \text{decide } \mathcal{H}_M \end{cases}$$

where $\phi(x)$ partitions the data space $\mathcal{X}$ into M regions:

$$(\mathcal{X}_m)_{m=1}^{M} = \{x \mid \phi(x) = m\}$$

depicted in Fig. 1.

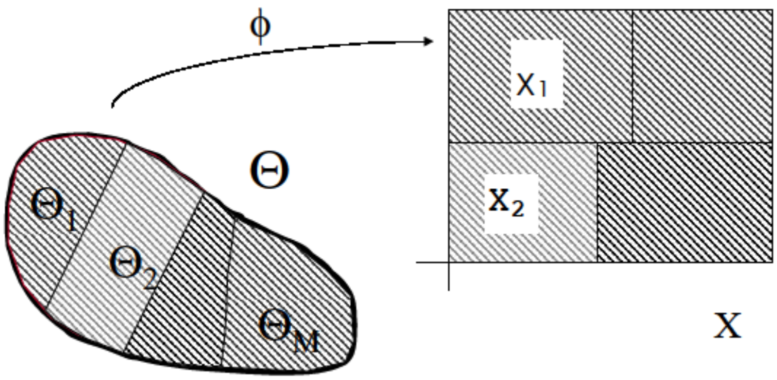

Figure 1: Parameter and measurement space partitioning for classification.

We use a piecewise-constant loss function

$$\mathbb{L}(\theta, \text{say } \mathbb{H}_m) = \sum_{i=1}^{M} \mathbb{L}(m \mid i)\, \mathbb{1}_{\mathrm{sp}_\Theta(i)}(\theta) \tag{1}$$

where $\mathbb{L}(m \mid i)$ is the loss of deciding the mth hypothesis when hypothesis i is true. Now, our posterior expected loss takes M values:

for $m = 1, \ldots, M$

$$\begin{aligned}
\underbrace{\rho_m(x)}_{\rho(\text{say } \mathcal{H}_m \mid x)} &= \int_{\mathrm{sp}_\Theta} \mathbb{L}(\theta, \text{say } \mathcal{H}_m)\, f_{\Theta \mid X}(\theta \mid x)\, \mathrm{d}\theta \\
&= \sum_{i=1}^{M} \int_{\mathrm{sp}_\Theta(i)} \mathbb{L}(m \mid i)\, f_{\Theta \mid X}(\theta \mid x)\, \mathrm{d}\theta \\
&= \sum_{i=1}^{M} \mathbb{L}(m \mid i) \underbrace{\int_{\mathrm{sp}_\Theta(i)} f_{\Theta \mid X}(\theta \mid x)\, \mathrm{d}\theta}_{\Pr(\mathbb{H}_i \mid x)} \\
&= \sum_{i=1}^{M} \mathbb{L}(m \mid i) \Pr(\mathbb{H}_i \mid x)
\end{aligned}$$

where

$$\begin{aligned}
\Pr(\mathbb{H}_i \mid x) &\triangleq \Pr_{\Theta \mid X}\big(\Theta \in \mathrm{sp}_\Theta(1) \mid \boldsymbol{x}\big) \\
&= \frac{f(x \mid \mathbb{H}_i) \Pr(\mathbb{H}_i)}{f_X(x)}.
\end{aligned} \tag{2}$$

☞ NOTE:

$$f(x \mid \mathbb{H}_i) = \frac{\int_{\mathrm{sp}_\Theta(i)} f_{X \mid \Theta}(x \mid \theta)\, f_\Theta(\theta)\, \mathrm{d}\theta}{\Pr(\mathbb{H}_i)}.$$

Then, the Bayes' decision rule $\phi^\star(x)$ is defined by the following data-space partitioning:

$$(\mathcal{X}_m^\star)_{m=1}^{M} = \left\{ x \,\middle|\, m = \arg \min_{1 \le \ell \le M} \rho_\ell(x) \right\}$$

or, equivalently, upon applying the Bayes' rule,

$$\mathcal{X}_m^\star = \left\{ x \,\middle|\, m = \arg\min_{1\le\ell\le M} \underbrace{\sum_{i=1}^{M} \mathbb{L}(\ell \,|\, i)\Pr(\mathbb{H}_i) f(x \,|\, \mathbb{H}_i)}_{\triangleq h_\ell(x)} \right\}. \tag{3}$$

0–1 loss, MAP, and ML rules

0–1 loss:

$$\mathbb{L}(m \,|\, i) = 1 - \delta_{m,i}$$

where $\delta_{m,i} = \begin{cases} 1, & m = i, \\ 0, & m \neq i \end{cases}$ is the Kronecker delta symbol. Hence, the posterior expected loss $\rho_m(\boldsymbol{x})$ can be written as

$$\rho_m(\boldsymbol{x}) = 1 - \Pr(\mathbb{H}_i \,|\, x)$$

which yields the following Bayes' decision rule, called the maximum *a posteriori* (MAP) rule:

$$\mathcal{X}_m^\star = \left\{ x \,\middle|\, m = \arg\max_{0\le\ell\le M-1} \Pr(\mathbb{H}_\ell \,|\, x) \right\}. \tag{4}$$

i.e.,

$$\mathcal{X}_m^\star = \left\{ x \,\middle|\, m = \arg\max_{0\le\ell\le M-1} \Pr(\mathbb{H}_\ell) f(x \,|\, \mathbb{H}_\ell) \right\}. \tag{5}$$

✻ ML RULE. For equiprobable hypotheses:

$$\Pr(\mathbb{H}_m) = \frac{1}{M} \qquad \forall m \tag{6a}$$

the MAP rule (4) is known as the maximum-likelihood (ML) rule. Substituting (6a) into (5) yields

$$\mathcal{X}_m^\star = \left\{ x \,\middle|\, m = \arg\max_{0\le\ell\le M-1} f(x \,|\, \mathbb{H}_\ell) \right\}. \tag{6b}$$

ML rule

Example 1 (Classifying DC level in Gaussian noise with known variance). Consider independent, identically distributed (i.i.d.) Gaussian measurements $\boldsymbol{X} = (X[n])_{n=0}^{N-1} = \boldsymbol{x}$ with unknown means μ and known variances σ^2: $X[n] \sim \mathcal{N}(\mu, \sigma^2)$. Consider simple hypotheses with three values, i.e., $M = 3$:

$$\begin{aligned} \mathbb{H}_1 : &\quad \mu = \mu_1 \\ \text{versus} & \\ \mathbb{H}_2 : &\quad \mu = \mu_2 \\ \text{versus} & \\ \mathbb{H}_3 : &\quad \mu = \mu_3 \end{aligned}$$

The optimal classifier depends on $\boldsymbol{x}$ only through sufficient statistic for μ:

$$\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n].$$

We know

$$\{\bar{X} \mid \mu\} \sim \mathcal{N}(\mu, \sigma^2/N).$$

Assume 0–1 loss and equiprobable hypotheses. Then, the ML test in (6b) applies:

$$\mathcal{X}_m^\star = \left\{\bar{x} \,\middle|\, m = \arg\max_{1\le\ell\le M} f(\bar{x} \mid \mu_\ell)\right\} \tag{7}$$

where

$$f(\bar{x} \mid \mu_\ell) = \frac{1}{\sqrt{2\pi\sigma^2/N}} \exp\Big[-\frac{1}{2\sigma^2/N}(\bar{x} - \mu_\ell)^2\Big]$$

and becomes

$$\mathcal{X}_m^\star = \left\{\boldsymbol{x} \,\middle|\, \bar{x}\mu_m - 0.5\mu_m^2 \ge \bar{x}\mu_\ell - 0.5\mu_\ell^2,\ \forall \ell\right\}. \tag{8}$$

Consider $\mu_1 = -1, \mu_2 = 1, \mu_3 = 2$. By plotting the 3 lines defined by the equalities in (8) as a function of $\bar{x}$, we can easily find the decision regions:

$$\mathcal{X}_1^\star = \{\boldsymbol{x} \mid \bar{x} \le 0\}$$
$$\mathcal{X}_2^\star = \{\boldsymbol{x} \mid 0 < \bar{x} \le 1.5\}$$
$$\mathcal{X}_3^\star = \{\boldsymbol{x} \mid \bar{x} \ge 1.5\}$$

see Fig. 2.

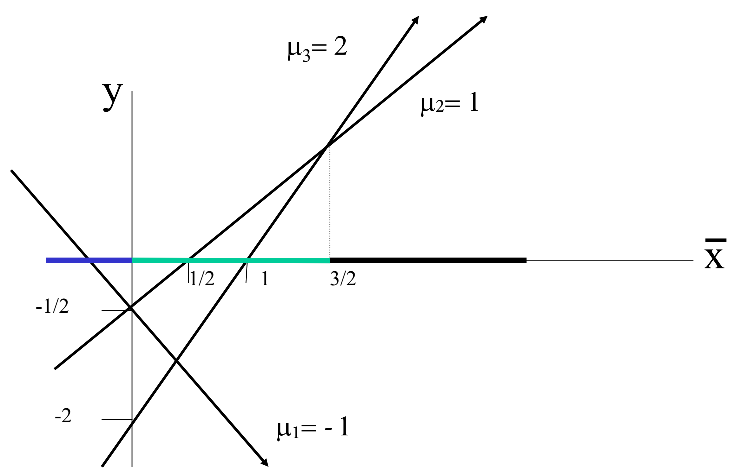

Figure 2: Decision regions.

Bayes Risk

Apply the law of iterated expectations:

$$
\begin{aligned}
&\mathrm{E}_{X,\Theta}\big[\mathbb{L}\big(\Theta,\ \text{decide } \mathbb{H}_{\phi(X)}\big)\big] = \mathrm{E}_X\big\{\mathrm{E}_{\boldsymbol{\Theta}|X}\big[\mathbb{L}\big(\Theta,\ \text{say } \mathbb{H}_{\phi(X)}\big)\,\big|\,X\big]\big\}\\
&= \mathrm{E}_X\big[\rho\big(\text{say } \mathbb{H}_{\phi(X)}\,\big|\,X\big)\big]\\
&= \mathrm{E}_X\Big[\sum_{i=0}^{M-1} \mathbb{L}(\phi(X)\,|\,i)\,\Pr\big(\mathbb{H}_i\,|\,X\big)\Big]\\
&= \int_{\mathcal{X}} \sum_{i=0}^{M-1} \mathbb{L}(\phi(x)\,|\,i)\,\underbrace{\Pr(\mathbb{H}_i\,|\,x)\,f_X(x)}_{\text{joint}}\,\mathrm{d}x\\
&= \int_{\mathcal{X}} \sum_{i=0}^{M-1} \mathbb{L}(\phi(x)\,|\,i)\,\underbrace{\Pr(\mathbb{H}_i)\,f(x\,|\,\mathbb{H}_i)}_{\text{joint}}\,\mathrm{d}x\\
&= \sum_{m=0}^{M-1} \int_{\mathcal{X}_m} \underbrace{\sum_{i=0}^{M-1} \mathbb{L}(m\,|\,i)\,\Pr\{\mathbb{H}_i\}\,f(x\,|\,\mathbb{H}_i)}_{h_m(x)}\ \mathrm{d}x \qquad (9)\\
&= \sum_{m=0}^{M-1} \sum_{i=0}^{M-1} \mathbb{L}(\ell\,|\,i)\,\Pr\{\mathbb{H}_i\}\,\underbrace{\int_{\mathcal{X}_m} f(x\,|\,\mathbb{H}_i)\,\mathrm{d}x}_{\Pr\{X\in\mathcal{X}_m\,|\,\mathbb{H}_i\}}\\
&= \sum_{m=0}^{M-1} \sum_{i=0}^{M-1} \mathbb{L}(m\,|\,i)\,\Pr(\mathbb{H}_i)\,\Pr(X\in\mathcal{X}_m\,|\,\mathbb{H}_i). \qquad (10)
\end{aligned}
$$

Recall (3) and (9):

$$
\mathcal{X}_m^\star = \Big\{x\,\Big|\,m = \arg\min_{0\le\ell\le M-1} h_\ell(x)\Big\} \qquad (11a)
$$

$$
\mathrm{E}_{X,\Theta}\big[\mathbb{L}\big(\Theta,\ \text{decide } \mathbb{H}_{\phi(X)}\big)\big] = \sum_{m=0}^{M-1}\int_{\mathcal{X}_m} h_m(x)\,\mathrm{d}x \qquad (11b)
$$

Then, for an arbitrary rule $\phi(x)$,

see (11a) and (11b)

$$
\sum_{m=0}^{M-1}\int_{\mathcal{X}_m} h_m(x)\,\mathrm{d}x - \sum_{m=0}^{M-1}\int_{\mathcal{X}_m^\star} h_m(x)\,\mathrm{d}x \ge 0
$$

which verifies that the Bayes' decision rule $\phi^\star(x)$ indeed minimizes the Bayes (preposterior) risk.

Average error probability

FOR the 0–1 loss, the Bayes risk for rule $\phi(\boldsymbol{x})$ is the average error probability:

$$\begin{aligned} P_{\text{av}} &= \mathrm{E}_{X,\Theta}\big[\mathbb{L}\big(\Theta,\ \text{decide } \mathbb{H}_{\phi(X)}\big)\big] \\ &= \sum_{m=0}^{M-1}\sum_{i=0}^{M-1} \mathbb{L}(m \mid i)\Pr(\mathbb{H}_i)\Pr(X \in \mathcal{X}_m \mid \mathbb{H}_i) \\ &= 1 - \underbrace{\sum_{m=0}^{M-1} \Pr(\mathbb{H}_m)\Pr(X \in \mathcal{X}_m \mid \mathbb{H}_m)}_{\Pr\left(\substack{\text{correct}\\ \text{decision}}\right)}\ . \end{aligned} \tag{12}$$

✻ UNION bound. Suppose we wish to bound from above the minimum average error probability achieved by the Bayes' rule. If we had a binary hypothesis problem, say testing $\mathbb{H}_i$ versus $\mathbb{H}_j$, then the minimum average *pairwise* error probability for this binary problem was obtained in handout `Chernoffbound`:

$$\int_{\mathcal{X}} \min\left\{f(\boldsymbol{x} \mid \mathbb{H}_i)\Pr(\mathbb{H}_i),\, f(\boldsymbol{x} \mid \mathbb{H}_j)\Pr(\mathbb{H}_j)\right\} \mathrm{d}\boldsymbol{x}.$$

Now,

$$\min P_{\text{av}} \le \sum_{j=1}^{M-1}\sum_{i=0}^{j-1} P(i,j)$$

which follows by applying the union-bound inequality[1] on

$$\text{error event} = \bigcup_{j=1}^{M-1}\bigcup_{i=0}^{j-1} A(i,j)$$

where $A(i,j)$ is the event of mistakingly deciding $\mathbb{H}_i$ instead of $\mathbb{H}_j$ or vice versa.

If we cannot easily compute $P(i,j)$, we can try to find an upper bound for it using the Chernoff bound, see handout `Chernoffbound`.

[1] see the review of union bound at `https://youtu.be/3gV4LWWhWwo?t=335` by Prof. Tsitsiklis, edX

Acronyms

i.i.d. independent, identically distributed. 3

MAP maximum *a posteriori*. 3

ML maximum-likelihood. 3, 4

References

Hero, Alfred O. (2015). *Statistical Methods for Signal Processing*. Lecture notes. Univ. Michigan, Ann Arbor, MI (cit. on p. 1).

Van Trees, Harry L., Kristine L. Bell, and Zhi Tian (2013). *Detection, Estimation, and Modulation Theory, Part I*. 2nd ed. New York: Wiley (cit. on p. 1).