Bayes Decision Theory

Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr

Bayes Decision Theory

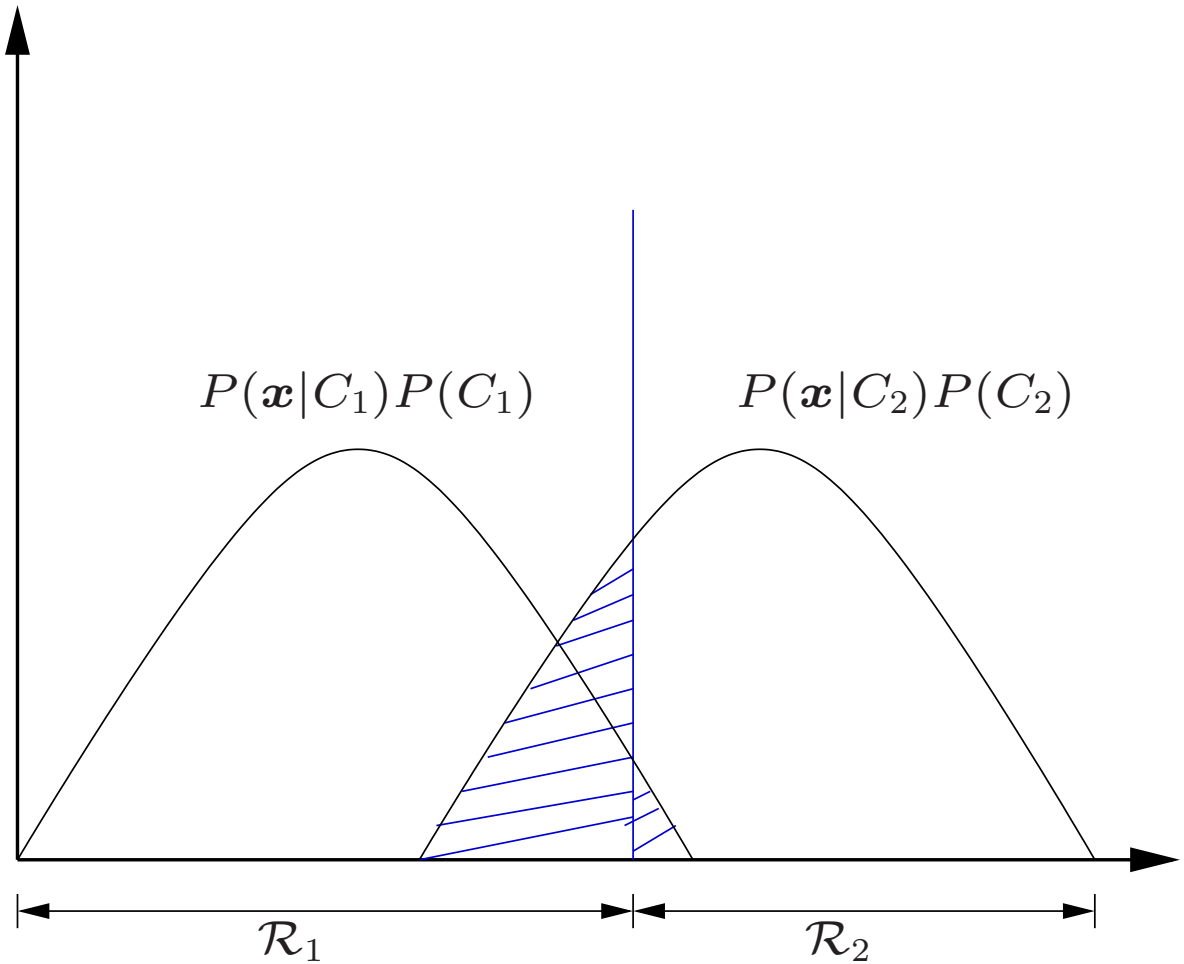
- Consider a set of feature vectors, $\{\boldsymbol{x}\}$, which belong to either class C_1 or C_2 with prior probability $P(C_i)$.
- A fundamental question arises: What would be a best way to assign an appropriate class label to a data point $\boldsymbol{x}$?
- Decide C_1 if $P(C_1) > P(C_2)$? $\Rightarrow$ Not a good idea (little information)
- Bayes decision theory gives an answer to this fundamental question.
- Bayes decision rule: Decide C_1 if $P(C_1|\boldsymbol{x}) > P(C_2|\boldsymbol{x})$.
- Question: Does this Bayes decision rule give the minimal probability of misclassification? $\Rightarrow$ Yes!

Bayes Rule

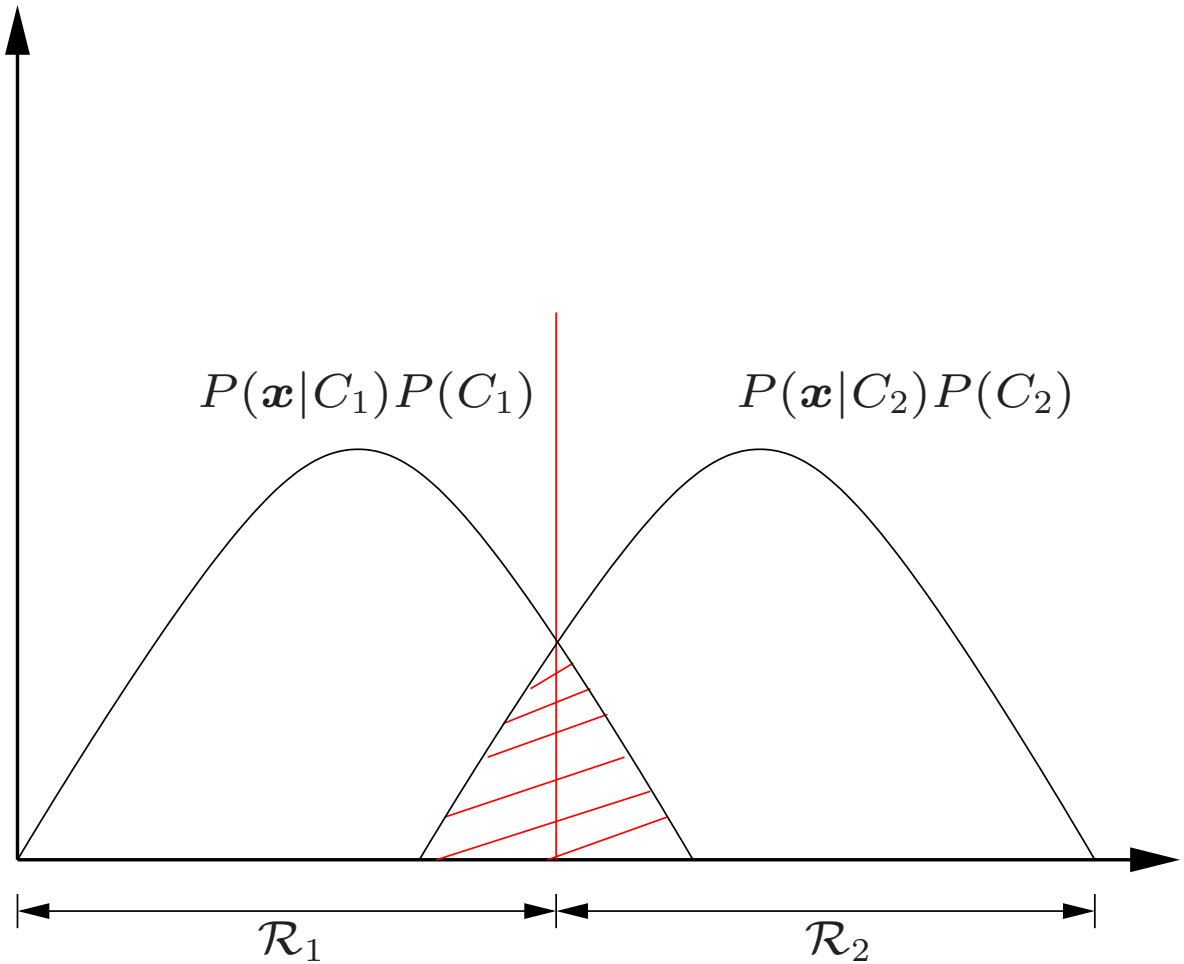
$$P(C_k|\boldsymbol{x}) = \frac{\overbrace{P(\boldsymbol{x}|C_k)}^{\text{class-conditional density}} \overbrace{P(C_k)}^{\text{prior}}}{\underbrace{\sum_j P(\boldsymbol{x}|C_j)P(C_j)}_{\text{normalization factor}}}.$$

In practice, we model the class-conditional density $P(\boldsymbol{x}|C_k)$ by a parameterized form.

Decision Boundary: Case 1



Decision Boundary: Case 2



Decision Boundaries

- Decision boundaries are boundaries between decision regions.
- The probability of misclassification in the binary classification problem, is given by

$$\begin{aligned} P(\text{error}) &= P(\boldsymbol{x} \in \mathcal{R}_2, C_1) + P(\boldsymbol{x} \in \mathcal{R}_1, C_2) \\ &= P(\boldsymbol{x} \in \mathcal{R}_2 | C_1) P(C_1) + P(\boldsymbol{x} \in \mathcal{R}_1 | C_2) P(C_2) \\ &= \int_{\mathcal{R}_2} P(\boldsymbol{x}|C_1) P(C_1) d\boldsymbol{x} + \int_{\mathcal{R}_1} P(\boldsymbol{x}|C_2) P(C_2). \end{aligned}$$

- One can observe that if $P(\boldsymbol{x}|C_1)P(C_1) > P(\boldsymbol{x}|C_2)P(C_2)$, we should choose the regions $\mathcal{R}_1$ and $\mathcal{R}_2$ such that $\boldsymbol{x}$ is in $\mathcal{R}_1$ since this give a smaller contribution to the error.

Alternatively, we consider

$$\begin{aligned} P(\text{correct}) &= \sum_{k=1}^{K} P(\boldsymbol{x} \in \mathcal{R}_k, C_k) \\ &= \sum_{k=1}^{K} P(\boldsymbol{x} \in \mathcal{R}_k | C_k) P(C_k) \\ &= \sum_{k=1}^{K} \int_{\mathcal{R}_k} P(\boldsymbol{x} | C_k) P(C_k). \end{aligned}$$

This probability is maximized by choosing $\{\mathcal{R}_k\}$ such that $\boldsymbol{x}$ is assigned to the class for which the integrand is a maximum.

Discriminant Functions

- Decide C_k if $f_k(\boldsymbol{x}) > f_j(\boldsymbol{x})\ \forall j \neq k$.
- Choose $f_k(\boldsymbol{x}) = P(C_k|\boldsymbol{x}) = P(\boldsymbol{x}|C_k)P(C_k)$ where $P(\boldsymbol{x})$ is dropped.
- If $f_k(\boldsymbol{x})$ is a discriminant function then

$$af_k(\boldsymbol{x}) \quad \forall a > 0$$
$$f_k(\boldsymbol{x}) + b$$
$$g(f_k(\boldsymbol{x})) \quad g \text{ is a monotonically increasing function}$$

  are also eligible discriminant functions.
- Then, $f_k(\boldsymbol{x}) = \log\left[P(\boldsymbol{x}|C_k)P(C_k)\right] = \log P(\boldsymbol{x}|C_k) + \log P(C_k)$ is also a discriminant function.

Discriminant Functions for Normal Density

Consider discriminant functions

$$f_i(\boldsymbol{x}) = \log P(\boldsymbol{x}|C_i) + \log P(C_i).$$

Assume $P(\boldsymbol{x}|C_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
Then, discriminant functions have the form

$$\begin{aligned} f_i(\boldsymbol{x}) &= \log\left[\frac{1}{(2\pi)^{\frac{m}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\}\right] + \log P(C_i) \\ &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{m}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log P(C_i). \end{aligned}$$

Case 1: $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$

- Features are statistically independent with the same variance σ^2.
- Hyperspherical cluster.
- $|\boldsymbol{\Sigma}_i| = \sigma^{2m}$, $\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2}\boldsymbol{I}$.
- $f_i(\boldsymbol{x}) = -\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 + \log P(C_i)$.
- If $P(C_i)$ are the same for all classes, then the discriminant functions become

$$f_i(\boldsymbol{x}) = -\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2,$$

which is a minimum distance classifier.

Case 1 Leads to Linear Discriminant Functions

In case 1, the linear discriminant function $f_i(\boldsymbol{x})$ can be rewritten as

$$f_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0},$$

where

$$\begin{aligned} \boldsymbol{w}_i &= \frac{1}{\sigma^2}\boldsymbol{\mu}_i, \\ w_{i0} &= -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \log P(C_i). \end{aligned}$$

Decision boundaries are hyperplanes defined by $f_i(\boldsymbol{x}) = f_j(\boldsymbol{x})$.

Case 2: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

In such a case, discriminant functions are given by

$$f_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) + \log P(C_i).$$

If $P(C_i)$ are the same for all classes, the discriminant function is simply based on the Mahalanobis distance, $(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$.
Case 2 also leads to linear discriminant functions which have the form

$$f_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0},$$

where

$$\begin{aligned} \boldsymbol{w}_i &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i, \\ w_{i0} &= -\frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \log P(C_i). \end{aligned}$$

Case 3: Arbitrary $\boldsymbol{\Sigma}_i$

In such a case, discriminant functions have the form

$$f_i(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{W}_i \boldsymbol{x} + \boldsymbol{w}_i^T + w_{i0},$$

where

$$\begin{aligned} \boldsymbol{W}_i &= -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}, \\ \boldsymbol{w}_i &= \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i, \\ w_{i0} &= -\frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log P(C_i). \end{aligned}$$

This case leads to a quadratic discriminant function.
The decision boundaries are hyperquadrics. They can assume any of the general forms such as pairs of hyperplanes, hyperspheres, hyperellipsoids, and hyperparaboloids.

Loss Function and Expected Loss

Suppose that we are given a set of training data, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$.

- Loss function, $l(f(\boldsymbol{x}), y)$, quantifies the loss or cost associating with the prediction $f(\boldsymbol{x})$ when the data $\boldsymbol{x}$ were actually labeled with y.
- Expected loss is defined as

$$\begin{aligned} L(f(\boldsymbol{x}), y) &= \mathbb{E}_{p(y|\boldsymbol{x})}\left[l(f(\boldsymbol{x}), y)\right] \\ &= \int l(f(\boldsymbol{x}), y) p(y|\boldsymbol{x}) dy. \end{aligned}$$

0-1 Loss

The 0-1 binary loss function is of the form:

$$l(f(\boldsymbol{x}), y) = 1 - \delta_{f(\boldsymbol{x}),y} = \begin{cases} 0, & \text{if } f(\boldsymbol{x}) = y \\ 1, & \text{otherwise.} \end{cases}$$

It makes most sense when the hypothesis space is discrete.

The expected loss is given by

$$\begin{aligned} L(f(\boldsymbol{x}), y) &= \sum_y l(f(\boldsymbol{x}), y) p(y|\boldsymbol{x}) \\ &= \sum_y (1 - \delta_{f(\boldsymbol{x}),y}) p(y|\boldsymbol{x}) \\ &= \sum_y p(y|\boldsymbol{x}) - \sum_y \delta_{f(\boldsymbol{x}),y} p(y|\boldsymbol{x}) \\ &= 1 - p(f(\boldsymbol{x})|\boldsymbol{x}). \end{aligned}$$

The expected loss is minimized when $f(\boldsymbol{x})$ is chosen to be the maximum of the posterior distribution $p(y|\boldsymbol{x})$, i.e., MAP estimate.

Squared Loss

The squared error loss function is of the form:

$$\boxed{l(f(\boldsymbol{x}), y) = (y - f(\boldsymbol{x}))^2.}$$

It is most appropriate when y lives in a continuous space with a well-defined metric.

The expected loss is given by

$$\begin{aligned}
L(f(\boldsymbol{x}), y) &= \int l(f(\boldsymbol{x}), y) p(y|\boldsymbol{x}) dy \\
&= \int (y - f(\boldsymbol{x}))^2 p(y|\boldsymbol{x}) dy \\
&= \int y^2 p(y|\boldsymbol{x}) dy + f(\boldsymbol{x})^2 \int p(y|\boldsymbol{x}) dy - 2f(\boldsymbol{x}) \int y\, p(y|\boldsymbol{x}) dy \\
&= f(\boldsymbol{x})^2 - 2f(\boldsymbol{x}) \mathbb{E}[y|\boldsymbol{x}] + \mathbb{E}[y^2|\boldsymbol{x}] \\
&= (f(\boldsymbol{x}) - \mathbb{E}[y|\boldsymbol{x}])^2 + \mathbb{E}\left[(y - \mathbb{E}[y|\boldsymbol{x}])^2 \,\middle|\, \boldsymbol{x} \right],
\end{aligned}$$

which is minimized when $f(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$.