Bayes Classification

Zhiyao Duan
Associate Professor of ECE and CS
University of Rochester

Bayes Formula

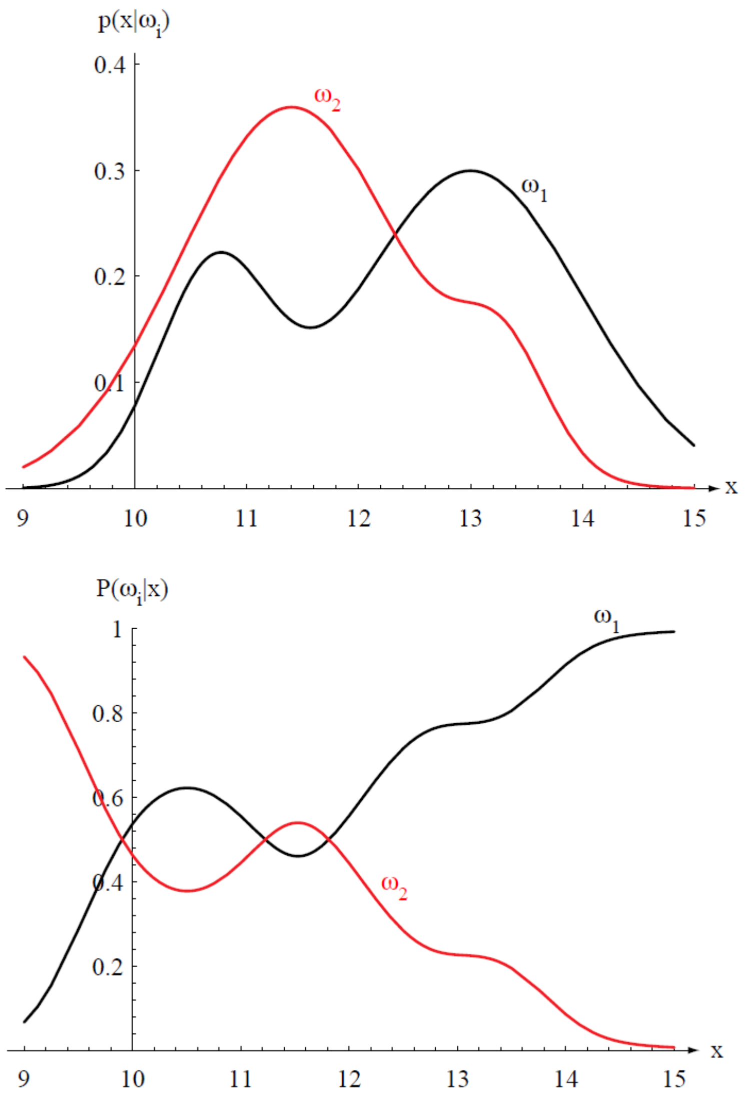
- Let $\boldsymbol{x}$ and y be random variables for the feature vector and target class, respectively

- Prior probability of class: $P(y)$
  - Probability of class before observing the feature vector
- Class-conditional probability, also called likelihood of class: $p(\boldsymbol{x}|y)$
  - Probability of the feature vector for each particular class
- Posterior probability: $P(y|\boldsymbol{x})$
  - Probability of class after observing the feature vector

- Bayes formula: $P(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)P(y)}{p(\boldsymbol{x})}$, or $posterior = \frac{likelihood \times prior}{evidence}$
  - $p(\boldsymbol{x}) = \sum_y p(\boldsymbol{x}|y)P(y)$

Example

- Let x be the lightness reading of a fish, y be the type of the fish (either sea bass: $y = \omega_1$, or salmon: $y = \omega_2$)
- Prior probability
  - $P(y = \omega_1) = \frac{2}{3}; P(y = \omega_2) = \frac{1}{3}$
- Class-conditional probability (likelihood)
  - $p(x|y = \omega_1)$ and $p(x|y = \omega_2)$
- Posterior probability
  - $p(y = \omega_1|x) = \frac{p(x|y = \omega_1)P(y=\omega_1)}{p(x)}$
  - $p(y = \omega_2|x) = \frac{p(x|y = \omega_2)P(y=\omega_2)}{p(x)}$
  - $p(x) = p(x|y = \omega_1)P(y = \omega_1) + p(x|y = \omega_2)P(y = \omega_2)$



(Figs. 2.1 and 2.2 in Duda, Hart & Stork, 2001)

Classification Error

- Error for each x

$$P(error|x) = \begin{cases} P(\omega_1|x) & if\ we\ decide\ y = \omega_2 \\ P(\omega_2|x) & if\ we\ decide\ y = \omega_1 \end{cases}$$

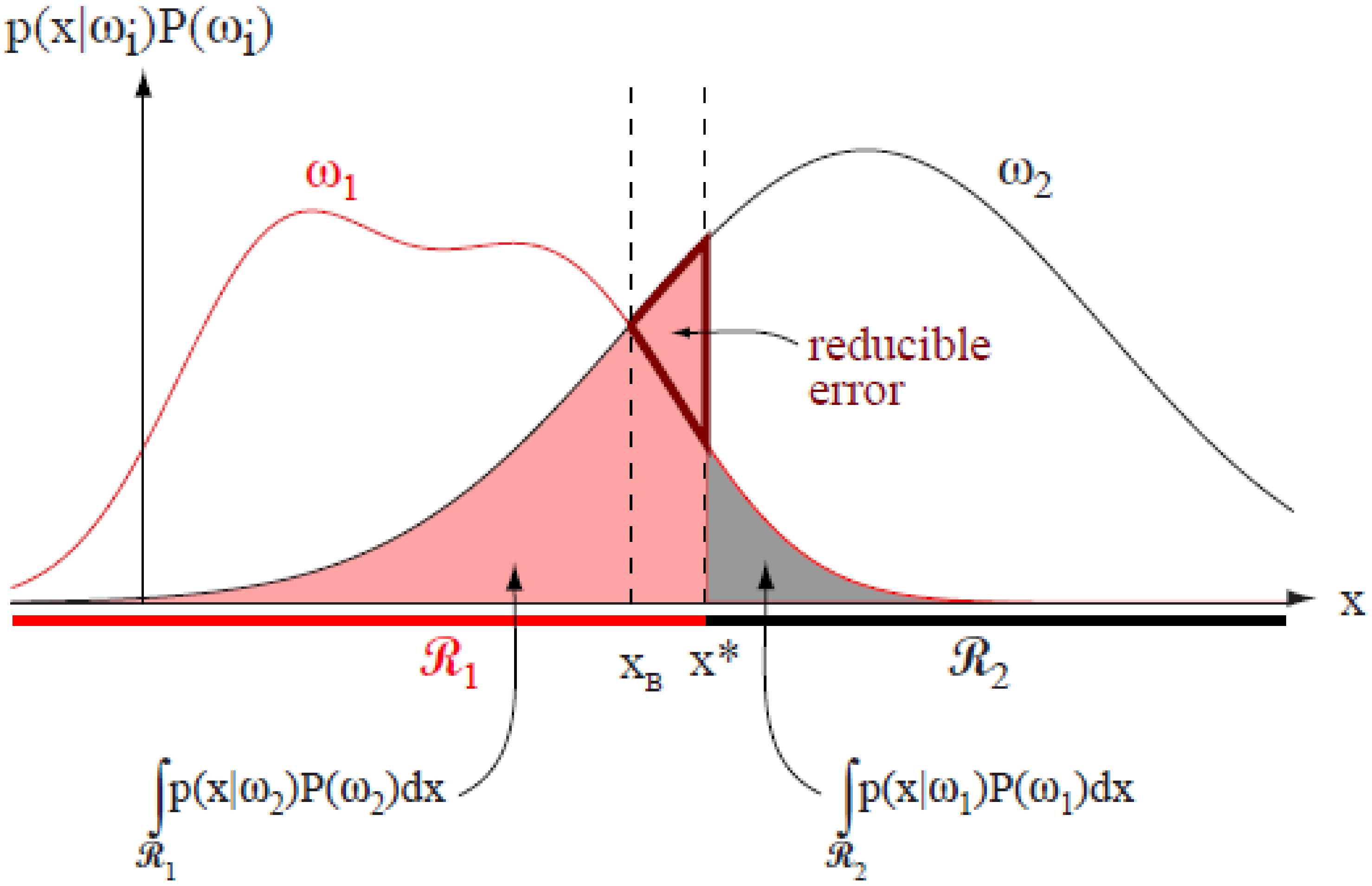
  - Bayes decision rule: decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2
  - Equivalently: decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2
  - Bayes error: $P(error|x) = \min\{P(\omega_1|x), P(\omega_2|x)\}$

- Average probability of error

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)dx$$

- Bayes decision rule minimizes the probability of error, i.e., optimal classifier!

Error Components



Generalize Classification Error to Risk

- Suppose we observe $\boldsymbol{x}$ and then take an action α_i
- Suppose action α_i incurs a loss $\lambda(\alpha_i|\omega_j)$ when the true class label is ω_j
- We define conditional risk as the expected loss for taking action α_i given $\boldsymbol{x}$

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{C} \lambda(\alpha_i|\omega_j) P(\omega_j|\boldsymbol{x})$$

- Let $\alpha(x)$ be a general decision rule, then
- Overall risk is

$$R = \int R(\alpha(\boldsymbol{x})|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

- Bayes decision rule: choose $\alpha(\boldsymbol{x}) = \alpha_i$ that minimizes $R(\alpha(\boldsymbol{x})|\boldsymbol{x})$

What is key to Bayes classification/decision?

- Posterior probability!

- It requires class-conditional probability (likelihood) and prior probability

$$P(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)P(y)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|y)P(y)}{\sum_y p(\boldsymbol{x}|y)P(y)}$$

- How to estimate prior probability?
  - Typically a one-dimensional discrete probability function

- How to estimate class-conditional probability?
  - Often a high-dimensional probability density function

Probability Density Estimation

- Non-parametric methods
  - Parzen-window: set a window around x and count the number of data points in the window
  - K-nearest-neighbor: find the volume of the K-nearest-neighborhood
- Parametric methods
  - Represent probability density with a parametric function, e.g., a Gaussian Mixture Model (GMM), and optimize the parameters to maximize the likelihood

Summary

- Bayes classifier classifiers data points to classes with the highest posterior probability

- It is optimal in the sense that minimizes classification error

- It requires a good estimate of the class-conditional probability distribution, which is often a difficult task

- Probability density estimation methods
  - Non-parametric methods
  - Parametric methods