Pattern Classification (II)

杜俊
jundu@ustc.edu.cn

Review

- Probability & Statistics
  - Bayes' theorem
  - Random variables: discrete vs. continuous
  - Probability distribution: PDF and CDF
  - Statistics: mean, variance, moment
  - Parameter estimation: MLE

- Information Theory
  - Entropy, mutual information, information channel, KL divergence

- Function Optimization
  - Constrained/unconstrained optimization

- Linear Algebra
  - Matrix manipulation

Outline

- Pattern Classification Problems
  - Inference and decision

- Bayesian Decision Theory
  - How to make the optimal decision?
  - Maximum *a posterior* (MAP) decision rule

- Generative Models
  - Joint distribution of observation and label sequences
  - Model estimation: MLE, Bayesian learning, discriminative training

- Discriminative Models
  - Model the posterior probability directly (discriminant function)
  - Logistic regression, support vector machine, neural network

Bayesian Decision Theory (I)

- Bayesian decision theory is a fundamental statistical approach to all pattern classification problems

- Pattern classification problem is posed in probabilistic terms
  - Observation X is viewed as random variables (vectors,…)
  - Class id C (C_1, C_2, … , C_N) is treated as a discrete random variable
  - All info about X and C can be obtained via joint distribution

$$p(X,\mathrm{C}) = \mathrm{P}(\mathrm{C}) \cdot p(X \mid \mathrm{C})$$

- Bayesian decision theory leads to the optimal classification with $p(X,\mathrm{C})$
  - Optimal → guarantee minimum average classification error
  - The minimum classification error is called the Bayes error

Bayesian Decision Theory (II)

- Prior probabilities of each class $\mathrm{P(C)}$
  - How likely any pattern from class C before observing any features
  - Prior knowledge from previous experience

$$\sum_{i=1}^{N} \mathrm{P(C}_i) = 1$$

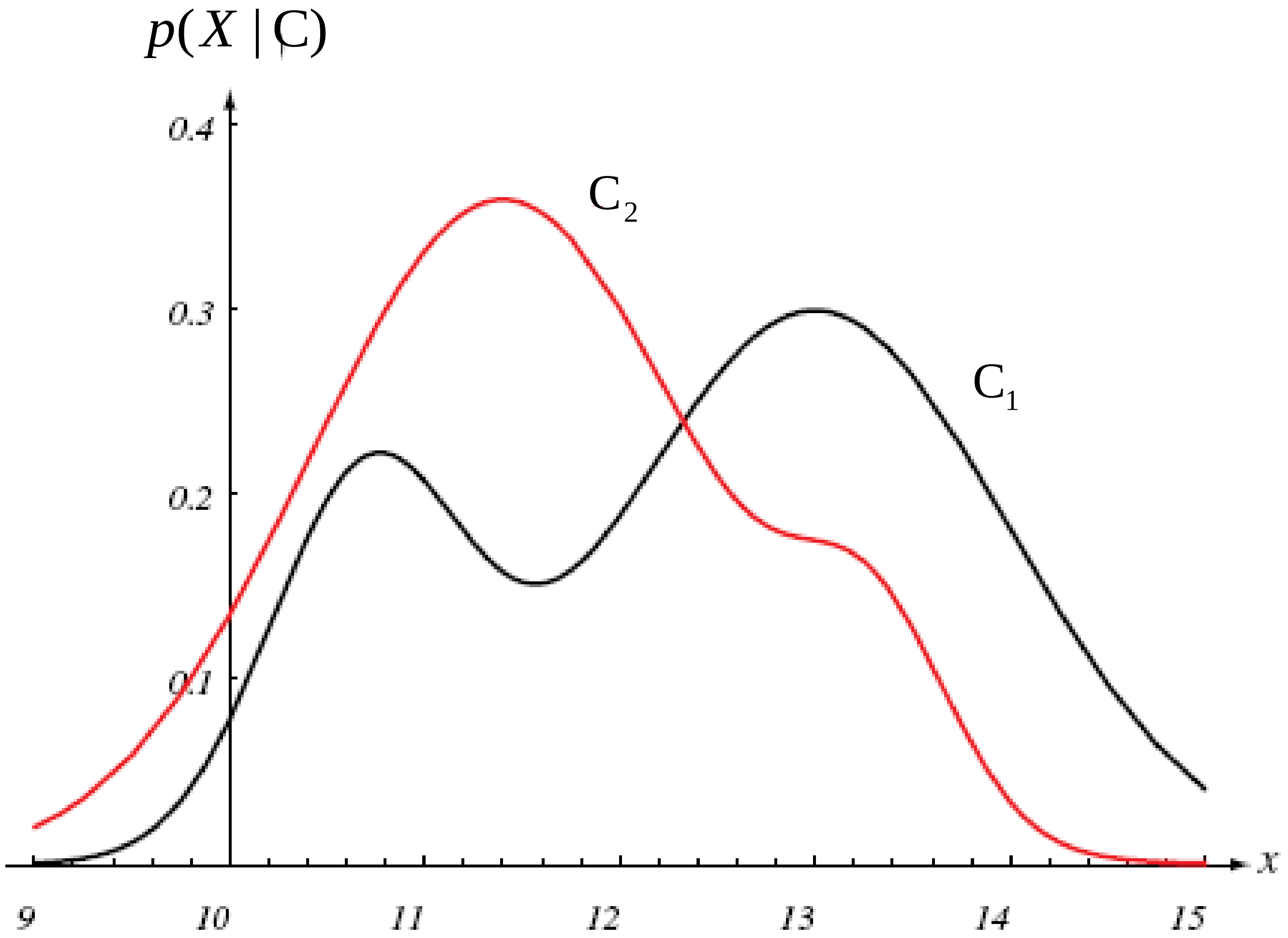
- Class-conditional probability of observed feature $p(X \mid \mathrm{C})$
  - How the feature X distributes for all patterns belonging to class C
  - If X is continuous, $p(X \mid \mathrm{C})$ is a PDF

$$\int_X p(X \mid \mathrm{C}_i) \cdot \mathrm{d}X = 1$$

  - If X is discrete, $p(X \mid \mathrm{C})$ is a PMF

$$\sum_X p(X \mid \mathrm{C}_i) = 1$$

Examples of Class Conditional Probability



Bayes Decision Rule (I)

- If not observe any feature of an incoming unknown pattern *P*, classify it based on prior knowledge only
  - Roughly guess it as the class with largest prior probability

$$C_P = \arg\max_{C} P(C)$$

- If observe some features *X* of the unknown patter *P*, we can convert the prior probability $P(C)$ into a posterior probability based on the Bayes' theorem:

$$\boxed{posterior = \frac{prior \times likelihood}{evidence}}$$

Bayes Decision Rule (II)

Prior   Likelihood

$$\underbrace{p(C_i \mid X)}_{\text{Posterior}} = \frac{\overbrace{P(C_i)}^{\text{Prior}} \cdot \overbrace{p(X \mid C_i)}^{\text{Likelihood}}}{\underbrace{p(X)}_{\text{Evidence}}} = \frac{P(C_i) \cdot p(X \mid C_i)}{\sum_i P(C_i) \cdot p(X \mid C_i)}$$

Posterior

Evidence

Bayes Decision Rule (III)

- Intuitively, we can classify an unknown pattern into the class with the largest posterior probabilities, resulting in the *maximum a posterior (MAP) decision rule*, also called *Bayes decision rule*

$$\mathrm{C}_P = \arg\max_{\mathrm{C}_i} p(\mathrm{C}_i \mid X) = \arg\max_{\mathrm{C}_i} P(\mathrm{C}_i) \cdot p(X \mid \mathrm{C}_i)$$

The MAP Decision Rule is Optimal (I)

- How well the MAP decision rule behaves??

- Optimality: assume we have complete knowledge $p(X, \mathrm{C})$, the MAP decision rule is optimal to classify patterns, which means it will achieve the lowest average classification error rate.

- Proof of optimality of the MAP rule:
  - Given a pattern *P*, if its true class id is C_i, but we classify it as C_p, then the classification error is counted as

$$l(\mathrm{C}_P \mid \mathrm{C}_i) = \begin{cases} 0 & (\mathrm{C}_P = \mathrm{C}_i) \\ 1 & (\mathrm{C}_P \neq \mathrm{C}_i) \end{cases}$$

which is also known as *0-1 loss function.*

The MAP Decision Rule is Optimal (II)

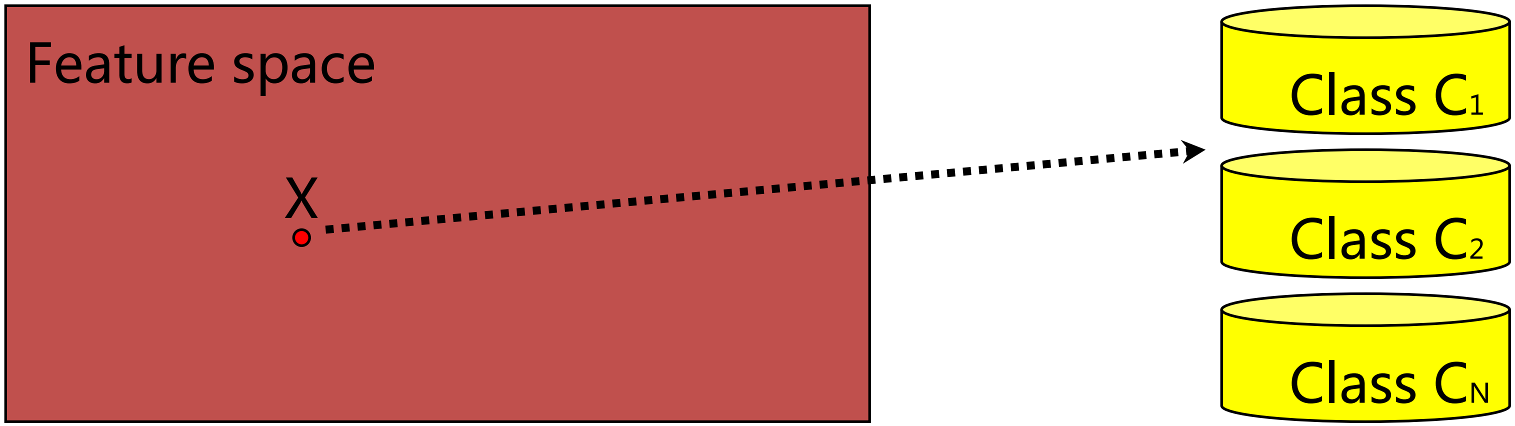
- The expected (average) classification error

$$R(C_P \mid X) = \sum_{i=1}^{N} l(C_P \mid C_i) \cdot p(C_i \mid X) = \sum_{C_i \neq C_P} p(C_i \mid X) = 1 - p(C_P \mid X)$$

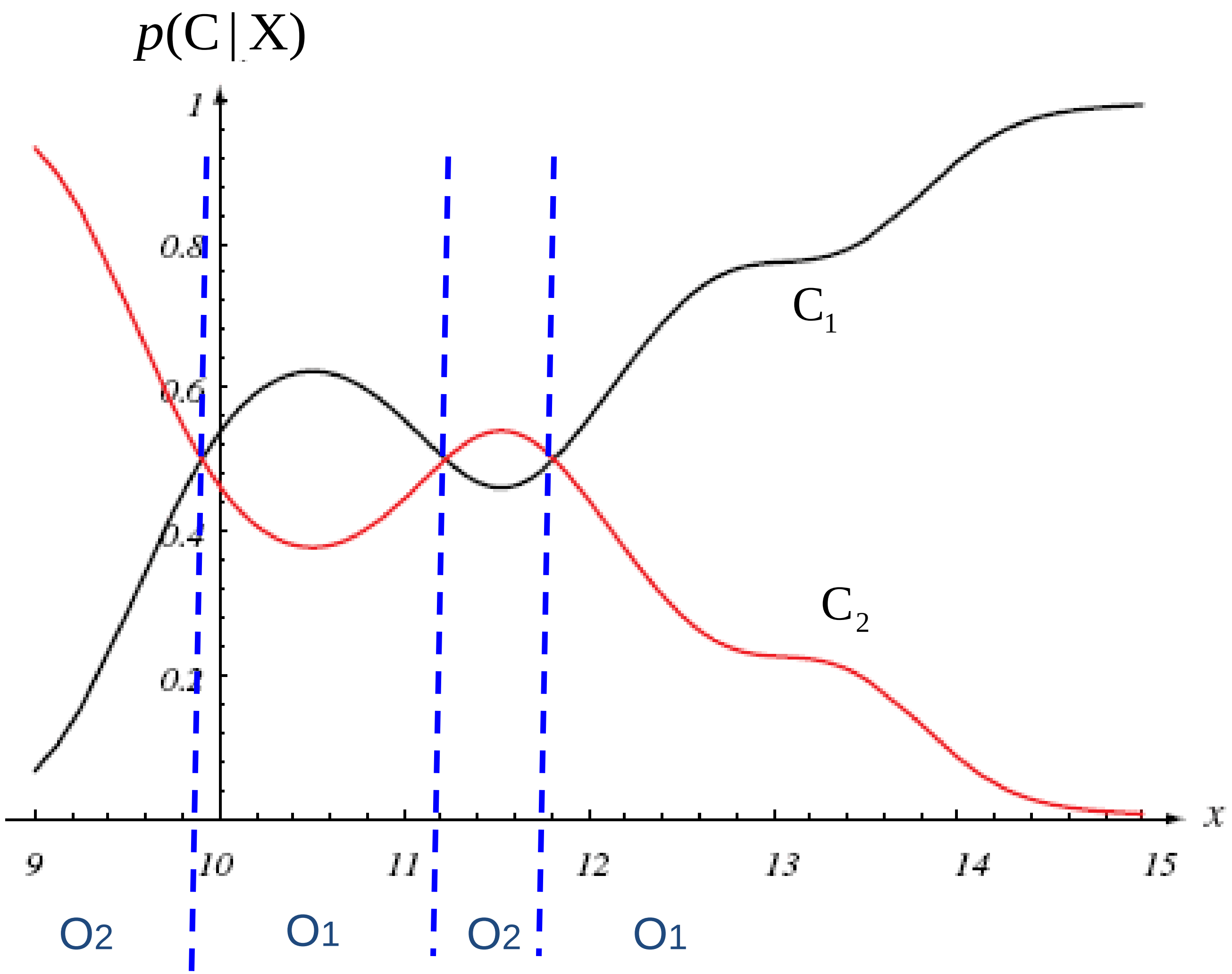
- The optimal classification is to minimize $R(C_P \mid X)$
  - ➔ maximize $p(C_P \mid X)$
  - ➔ the MAP decision rule is optimal

The MAP Decision Rule

- A general decision rule is a mapping function: $X \rightarrow C$
- A decision rule will partition the entire feature space of X into N different regions, $O_1, O_2, \ldots, O_N$. Each region O_i could consist of many contiguous areas.
- If X is located in the region O_i, we classify it as class C_i.
- The MAP decision rule is optimal among all possible decision rules in terms of minimizing average classification errors conditional on that we have complete knowledge about the underlying problem.



Example


$p(\mathrm{C}|\mathrm{X})$
1
0.8
0.6
0.4
0.2
C_1
C_2
9
10
11
12
13
14
15
x
O2
O1
O2
O1

Classification Error Probability

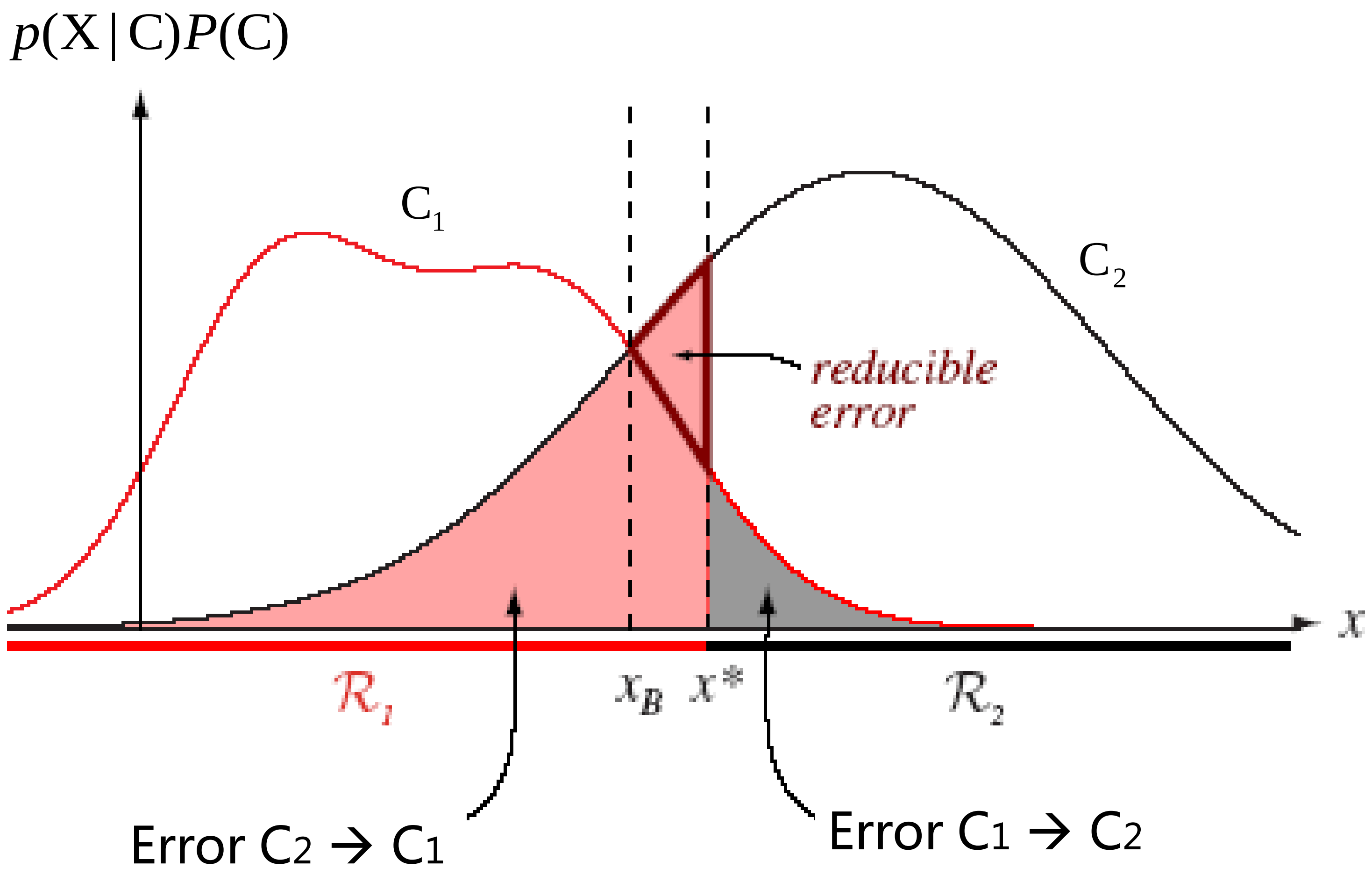
- Assume N-class problem, any a decision rule partitions the feature space into N regions, O_1, O_2, … , O_N.
- $\Pr(X \in O_i, C_j)$ denotes the probability of the observation X with true class id C_j in the region O_i.
- The overall classification error probability of the decision rule is:

$$\Pr(error) = 1 - \Pr(correct) = 1 - \sum_{i=1}^{N} \Pr(X \in O_i, C_i)$$

$$= 1 - \sum_{i=1}^{N} \Pr(X \in O_i \mid C_i) \cdot P(C_i)$$

$$= 1 - \sum_{i=1}^{N} \int_{O_i} p(X \mid C_i) \cdot P(C_i) \, \mathrm{d}X$$

Example


$p(X|C)P(C)$
C_1
C_2
reducible error
x
$\mathcal{R}_1$
x_B
x^*
$\mathcal{R}_2$
Error $C_2 \rightarrow C_1$
Error $C_1 \rightarrow C_2$

Bayes Error

- Bayes error: error probability of the Bayes (MAP) decision rule.

- Since Bayes decision rule guarantees the minimum error, the Bayes error is the lower bound of all possible error probabilities.

- It is difficult to calculate the Bayes error, even for the very simple cases because of discontinuous nature of the decision regions in the integral, especially in high dimensions.

- Some approximation methods to estimate an upper bound.
  - Chernoff bound
  - Bhattacharyya bound

- Evaluate on an independent test set.

Example: X is Discrete (I)

- A simple case (Binomial model): 2-class (C_1, C_2), feature vector is d-dimensional vector, whose components are binary-valued and conditionally independent.

$$X = (x_1, x_2, \cdots, x_d)^t \qquad x_i = 0,1\,(1 \le i \le d)$$

$$p_i = \Pr(x_i = 1 \mid C_1) \qquad q_i = \Pr(x_i = 1 \mid C_2)$$

$$p(X \mid C_1) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i}$$

$$p(X \mid C_2) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}$$

Example: X is Discrete (II)

- The MAP decision rule:

$$\text{classify to } C_1 \text{ if } P(C_1)\cdot p(X\mid C_1) \ge P(C_2)\cdot p(X\mid C_2), \text{ otherwise } C_2$$

Equivalently, we have the decision function:

$$g(X)=\sum_{i=1}^{d}\left[x_i\ln\frac{p_i}{q_i}+(1-x_i)\ln\frac{1-p_i}{1-q_i}\right]+\ln\frac{P(C_1)}{P(C_2)}=\sum_{i=1}^{d}\lambda_i x_i+\lambda_0$$

$$\lambda_i=\ln\frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad \lambda_0=\sum_{i=1}^{d}\ln\frac{1-p_i}{1-q_i}+\ln\frac{P(C_1)}{P(C_2)}$$

If $g(X)\ge 0$, classify to C_1, otherwise C_2.

Example: X is Continuous

- Gaussian model: 2-class (C_1, C_2), the feature vector is a scalar which is real-valued

$$P(x|C_1) = N(x;\mu_1,\sigma_1^2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-(x-\mu_1)^2/2\sigma_1^2}$$

$$P(x|C_2) = N(x;\mu_2,\sigma_2^2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-(x-\mu_2)^2/2\sigma_2^2}$$

- The MAP decision rule:

$$\text{classify to } C_1 \text{ if } P(C_1)\cdot p(x|C_1) \geq P(C_2)\cdot p(x|C_2), \text{ otherwise } C_2$$

.

Missing Features/Data (I)

- If we know the full probability structure of a problem, we can construct the optimal Bayes decision rule.
- In some practical situations, for some patterns, we can't observe the full feature vector described in the probability structure. Only partial information of the feature vector is observed, but some components are missing.
- How to classify such corrupted inputs to obtain minimum average error?
- Let the full feature vector $X=[X_g, X_b]$, X_g represents the observed or good features, X_b represents the missing or bad ones.
- In this case, the optimal decision rule is constructed as follows:

$$\mathrm{C}_P = \arg\max_{\mathrm{C}_i} \ p(\mathrm{C}_i \mid X_{\mathrm{g}})$$

Missing Features/Data (II)

$$p(\mathrm{C}_i \mid X_{\mathrm{g}}) = \frac{p(\mathrm{C}_i, X_{\mathrm{g}})}{p(X_{\mathrm{g}})} = \frac{\int p(\mathrm{C}_i, X_{\mathrm{g}}, X_{\mathrm{b}})\,\mathrm{d}X_{\mathrm{b}}}{p(X_{\mathrm{g}})}$$

$$(1) = \frac{\int p(\mathrm{C}_i \mid X_{\mathrm{g}}, X_{\mathrm{b}}) \cdot p(X_{\mathrm{g}}, X_{\mathrm{b}})\mathrm{d}X_{\mathrm{b}}}{p(X_{\mathrm{g}})} = \frac{\int p(\mathrm{C}_i \mid X) \cdot p(X)\,\mathrm{d}X_{\mathrm{b}}}{\int p(X)\,\mathrm{d}X_{\mathrm{b}}}$$

$$(2) = \frac{\int P(\mathrm{C}_i) \cdot p(X_{\mathrm{g}}, X_{\mathrm{b}} \mid \mathrm{C}_i)\,\mathrm{d}X_{\mathrm{b}}}{\sum_{\mathrm{C}_i} \int P(\mathrm{C}_i) \cdot p(X_{\mathrm{g}}, X_{\mathrm{b}} \mid \mathrm{C}_i)\,\mathrm{d}X_{\mathrm{b}}} = \frac{\int P(\mathrm{C}_i) \cdot p(X \mid \mathrm{C}_i)\,\mathrm{d}X_{\mathrm{b}}}{\sum_{\mathrm{C}_i} \int P(\mathrm{C}_i) \cdot p(X \mid \mathrm{C}_i)\,\mathrm{d}X_{\mathrm{b}}}$$

Practical Issue

- The optimal Bayes decision rule is not feasible in practice.
  - In any practical problem, we can not have a complete knowledge about the problem.
  - E.g., the class-conditional probability are always unavailable and extremely hard to estimate.

- However, possible to collect a set of sample data for each class in question.
  - The sample data are always far from enough to estimate a reliable PDF by using sample data themselves ONLY.

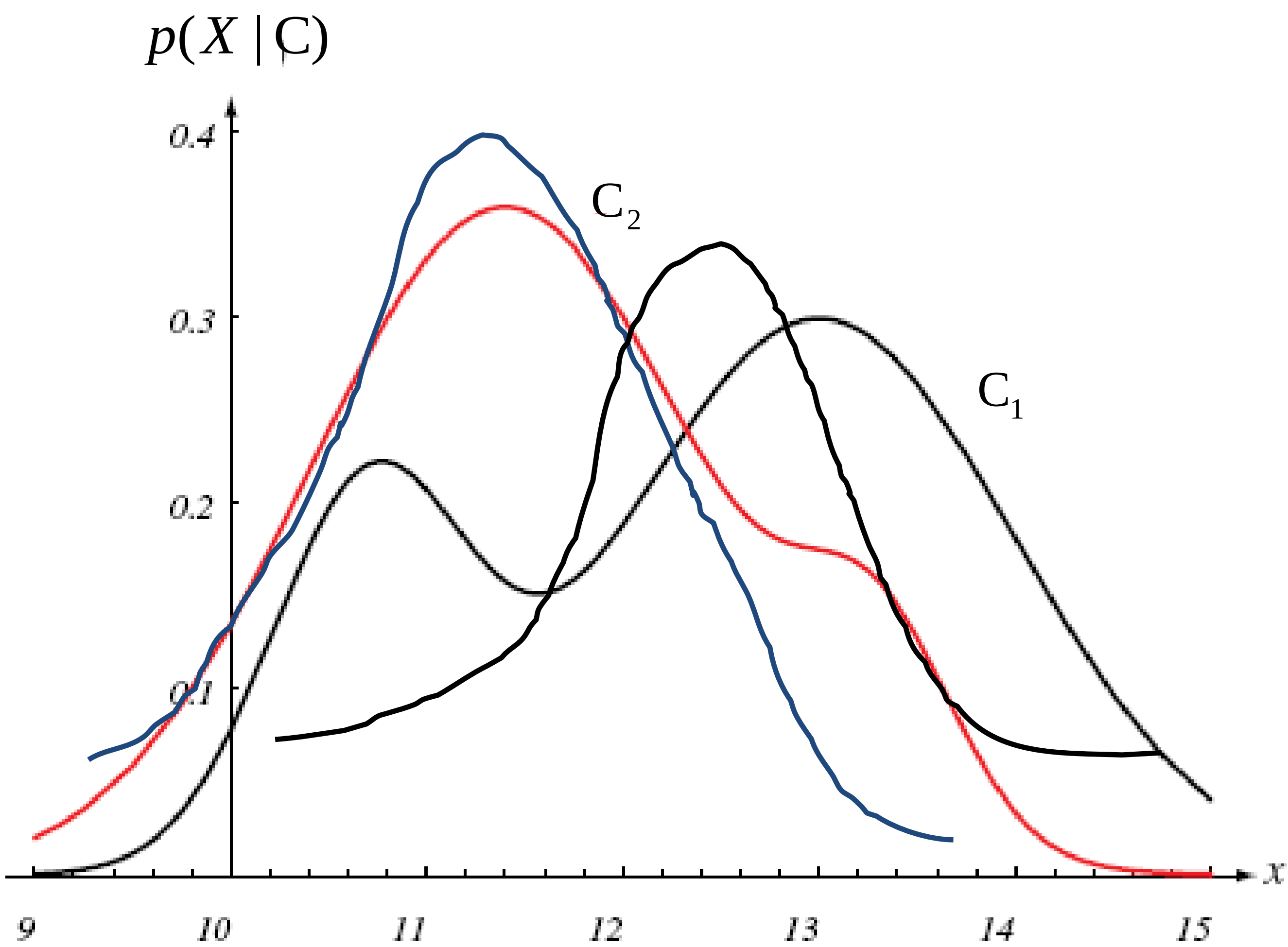
- Question: How to build a reasonable classifier based on a limited set of sample data, instead of the true PDF?

Statistical Data Modeling

- For any real problem, the true PDFs are always unknown

- Statistical data modeling: based on the available sample data set, choose a proper statistical model to fit into the available data set.
  - Data modeling stage: once the statistical model is selected, its function form becomes known except a set of model parameters associated with the model are unknown to us.
  - Learning (training) stage: the unknown parameters can be estimated by fitting the model into the data set based on certain estimation criterion.
  - Decision (test) stage: the estimated PDFs are plugged into the optimal Bayes decision rule in place of the real PDFs， so called plug-in MAP decision rule
    - Not optimal but performs reasonably well in practice

Data Modeling Example



Plug-in MAP Decision Rule

- Once the statistical models are estimated, they are treated as if they were true distributions of the data, and plug into the form of the optimal Bayes (MAP) decision rule in place of the unknown true PDFs.

- The plug-in MAP decision rule:

$$\mathrm{C}_P = \arg\max_{\mathrm{C}_i} p(\mathrm{C}_i \mid X) = \arg\max_{\mathrm{C}_i} P(\mathrm{C}_i) \cdot p(X \mid \mathrm{C}_i)$$

$$\approx \arg\max_{\mathrm{C}_i} \overline{P}_{\Gamma_i}(\mathrm{C}_i) \cdot \overline{p}_{\Lambda_i}(X \mid \mathrm{C}_i)$$