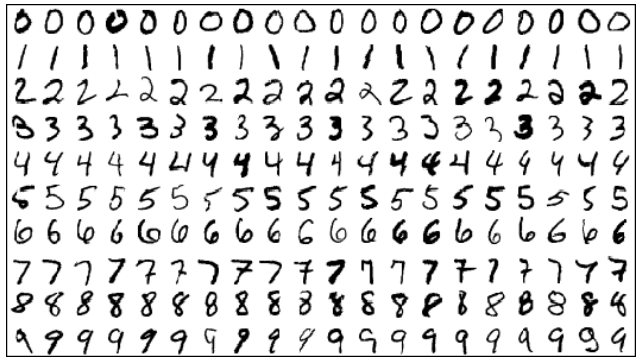
Pattern Recognition : Bayesian Classifier

Hyun-Min Park

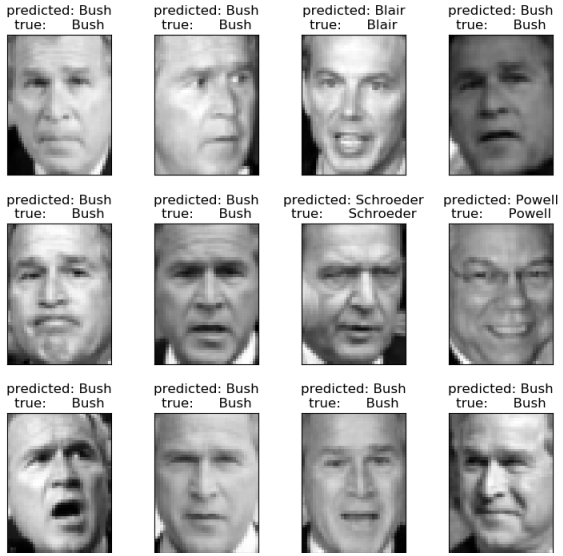
A.I Lab. Dept. of Computer Science and Engineering
Dongguk University. Seoul. Korea

Introduction

- Pattern Recognition, Machine Learning, Data Mining, Knowledge Discovery in Databases
- Pattern recognition is a branch of machine learning.
- Focuses on the recognition of patterns and regularities in data.
- It refers to grouping(recognizing) a given set of data into several groups(classes) according to specific criteria based on input values.



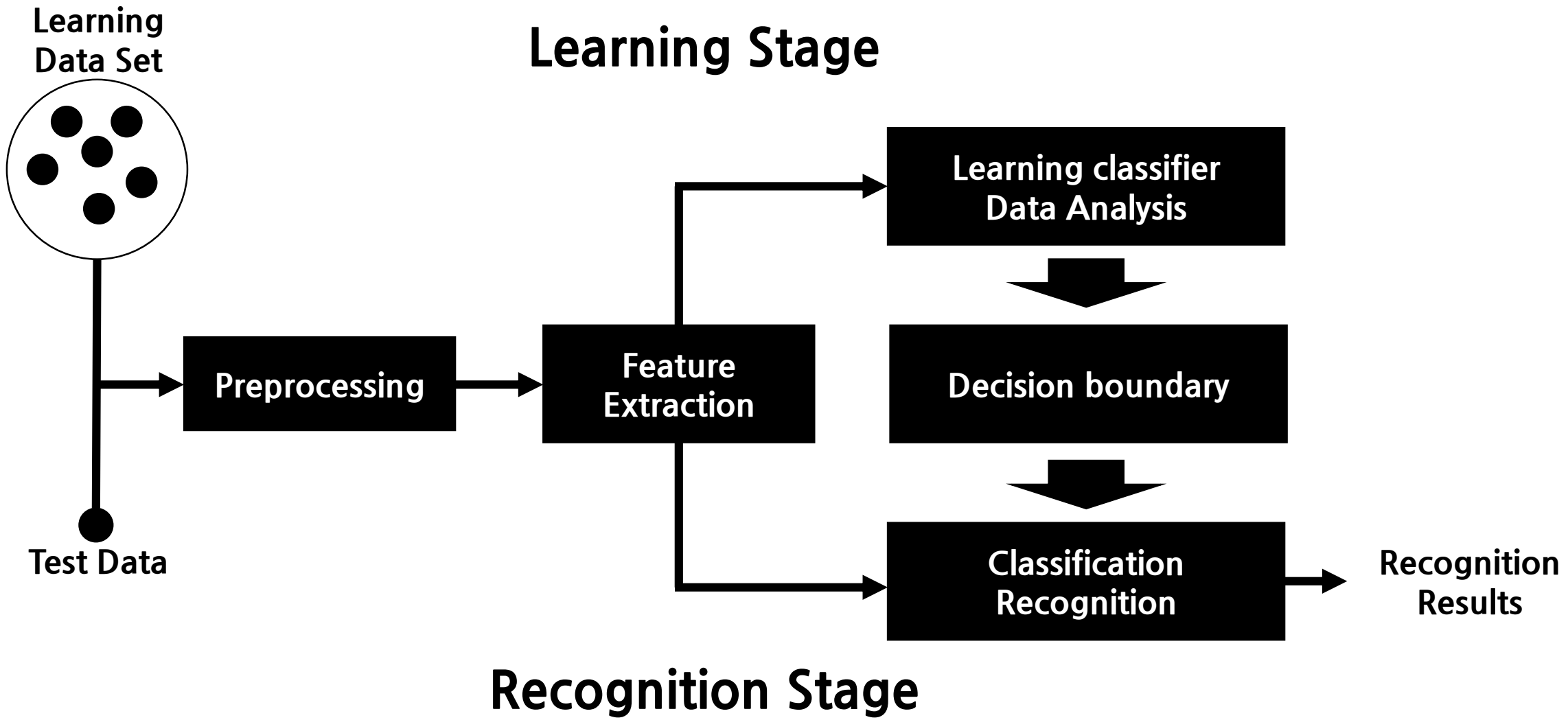
Digit Recognition



Face Recognition

Expression Recognition

Process


Learning
Data Set
Learning Stage
Learning classifier
Data Analysis
Preprocessing
Feature
Extraction
Decision boundary
Test Data
Classification
Recognition
Recognition
Results
Recognition Stage

Statistical Approach

- There is a population of bases for the data we observe in real life, and the data we currently see are those sampled from it.

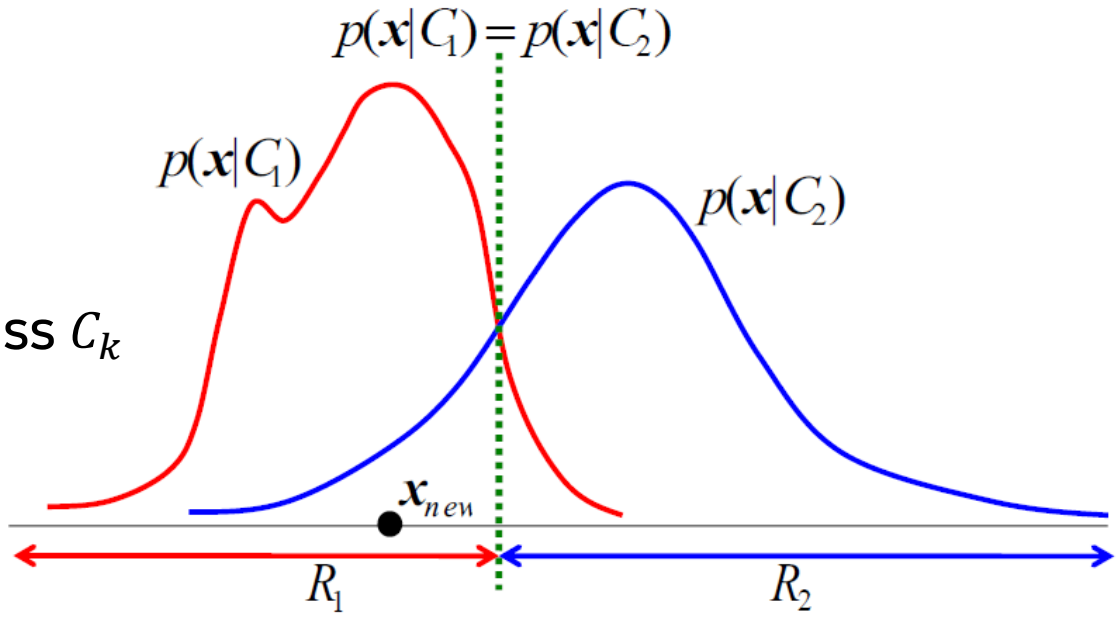
- Therefore, analyzing the present data is a process of establishing a probability distribution model for the population and estimating it using data.

- x , $p(x)$, m

Pattern Recognition

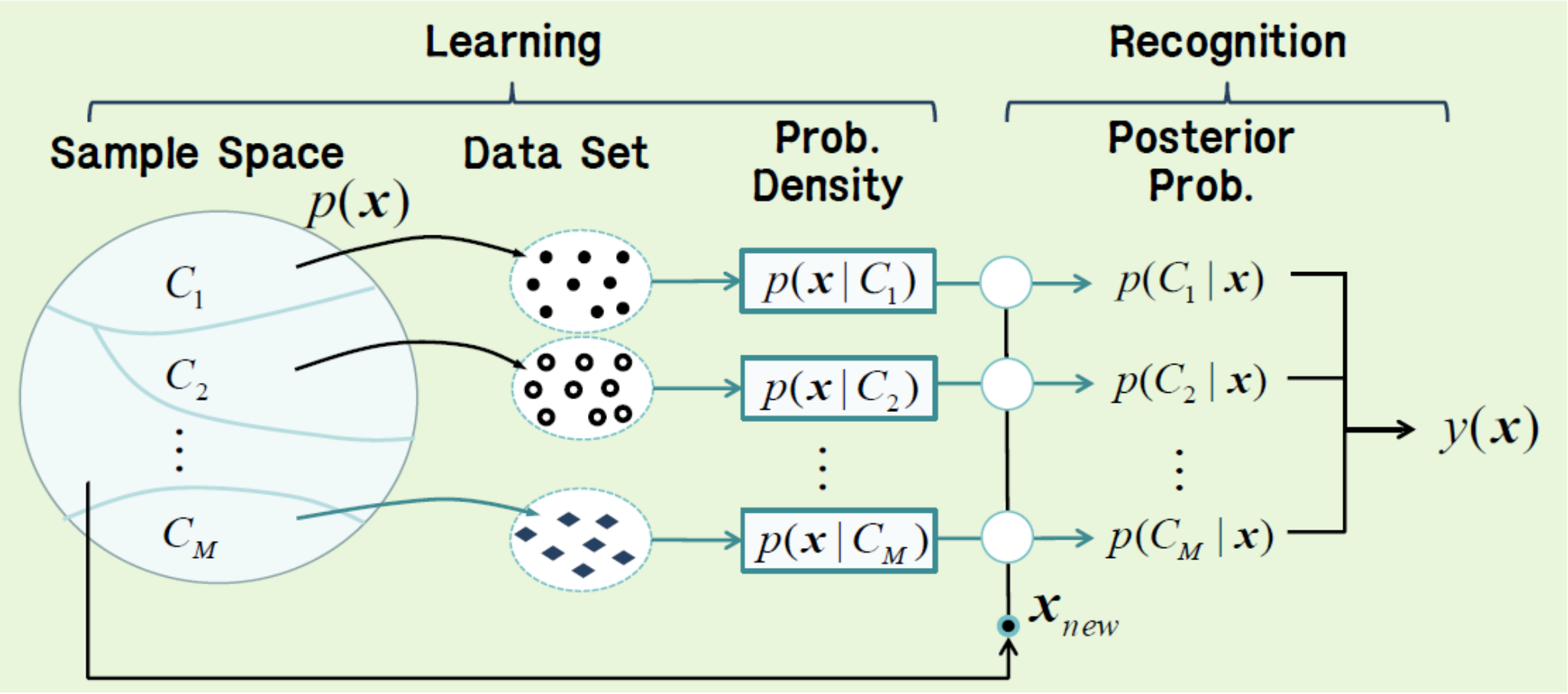
- $x_{new} \rightarrow P(C_k|x_{new}) \rightarrow x_{new} \in C_i$

- x_{new} : Test Data
- C_i : i^{th} Class
- $p(x|C_k)$: Conditional Probability of data x for class C_k
- $P(C_k|x_{new})$: Posterior Probability
- $P(C_k)$: Prior Probability



Pattern Recognition

- $x_{new} \rightarrow P(C_k|x_{new}) \rightarrow x_{new} \in C_i$



- Bayes' theorem
  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$,
  - Where A and B are events and $P(B) \neq 0$,
  - $P(A|B)$ is a conditional probability: the likelihood of event A occurring given that B is true.
  - $P(B|A)$ is also a conditional probability: the likelihood of event B occurring given that A is true.
  - $P(A)$ and $P(B)$ are the probabilities of observing A and B independently of each other; this is known as marginal probability.

$$P(C_k|x) = \frac{p(x|C_k)P(C_k)}{p(x)}$$

How can we find $P(C_k|x_{new})$?

- Generative approach
  - Ex)Bayesian, K-NN
  - Find $P(C_k|x_{new})$ after the estimation of $p(x|C_k)$
  - Disadvantage : Error accumulation
  - Parametric vs. Non-Parametric density estimation

- Discriminative approach
  - Ex)LDA, SVM
  - Find $P(C_k|x_{new})$ directly
  - Disadvantage : Cannot use Probability density

Bayesian Classifier

- $C^{Bayes}(x) = argmax\{P(Y = r|X = x)\}, r \in \{1,2, \dots, K\}$.

- A classifier is a rule that assigns to an observation X=x a guess or estimate of what the unobserved label Y=r actually was.

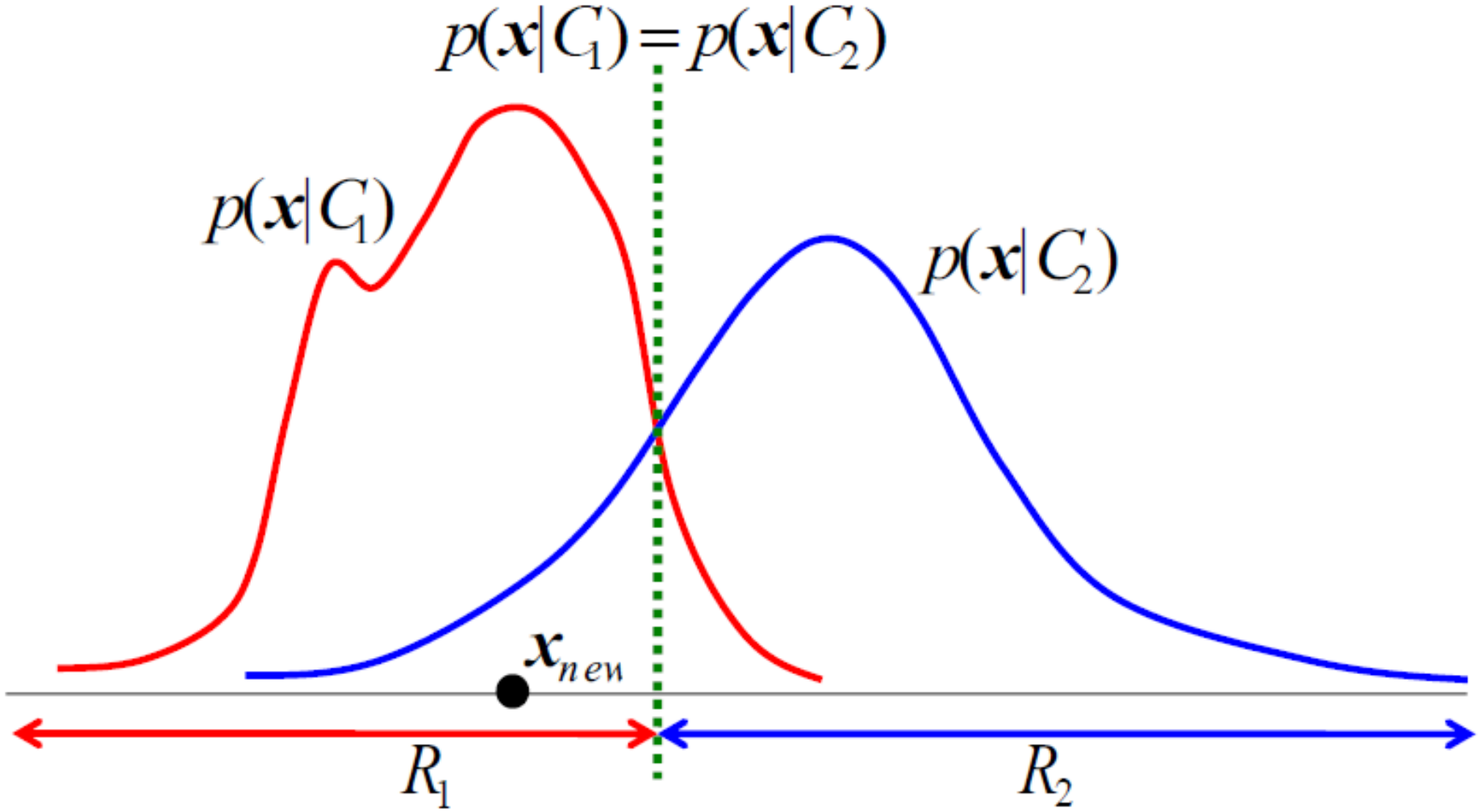
Binary classification

- aka. Two class problem : $x \in C_1$? , $x \in C_2$?
- Which class has the max. posterior prob. Btw $P(C_1|x)$ and $P(C_2|x)$?
- $y(x)$: class mapping function
- $g(x) = P(C_1|x) - P(C_2|x) = 0$
- Using $P(C_k|x) = \frac{p(x|C_k)P(C_k)}{p(x)}$,
- $g(x) = \frac{p(x|C_1)P(C_1)}{p(x)} - \frac{p(x|C_2)P(C_2)}{p(x)} = 0$
- Dividing by $\frac{p(x|C_2)P(C_1)}{p(x)}$,

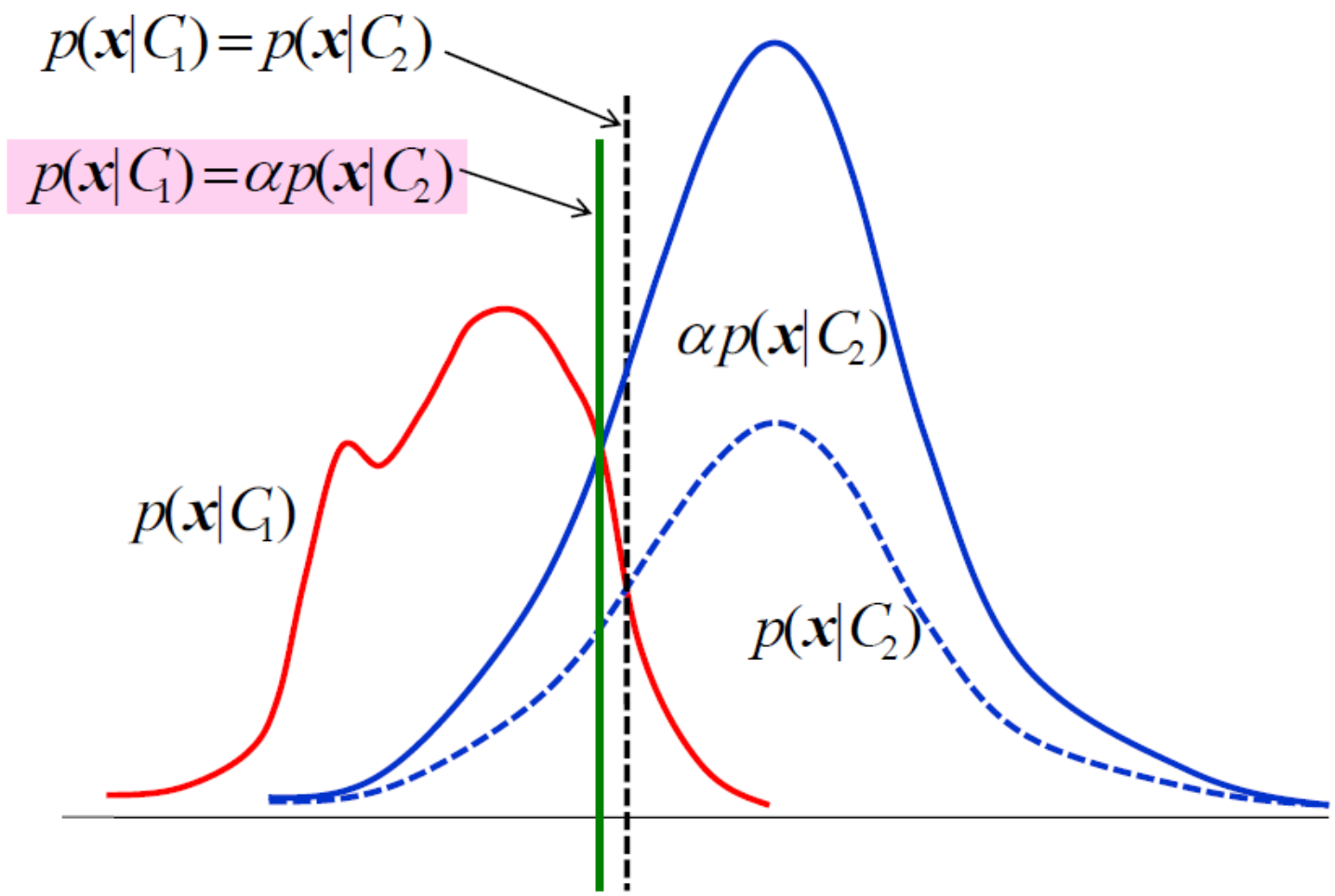
Likelihood Ratio Test (LRT)

- $g_{LRT}(x) = \frac{p(x|C_1)}{p(x|C_2)} - \frac{P(C_2)}{P(C_1)} = 0$
- $\frac{p(x|C_1)}{p(x|C_2)}$ = Likelihood ratio
- $y(x) = \begin{cases} 1 & if\ g_{LRT}(x) > 0 \\ -1 & otherwise \end{cases}$
- Simplify, $P(C_1) = P(C_2)$
- $y(x) = \begin{cases} 1 & if\ p(x|C_1) > p(x|C_2) \\ -1 & otherwise \end{cases}$

- If $P(C_1) = P(C_2)$

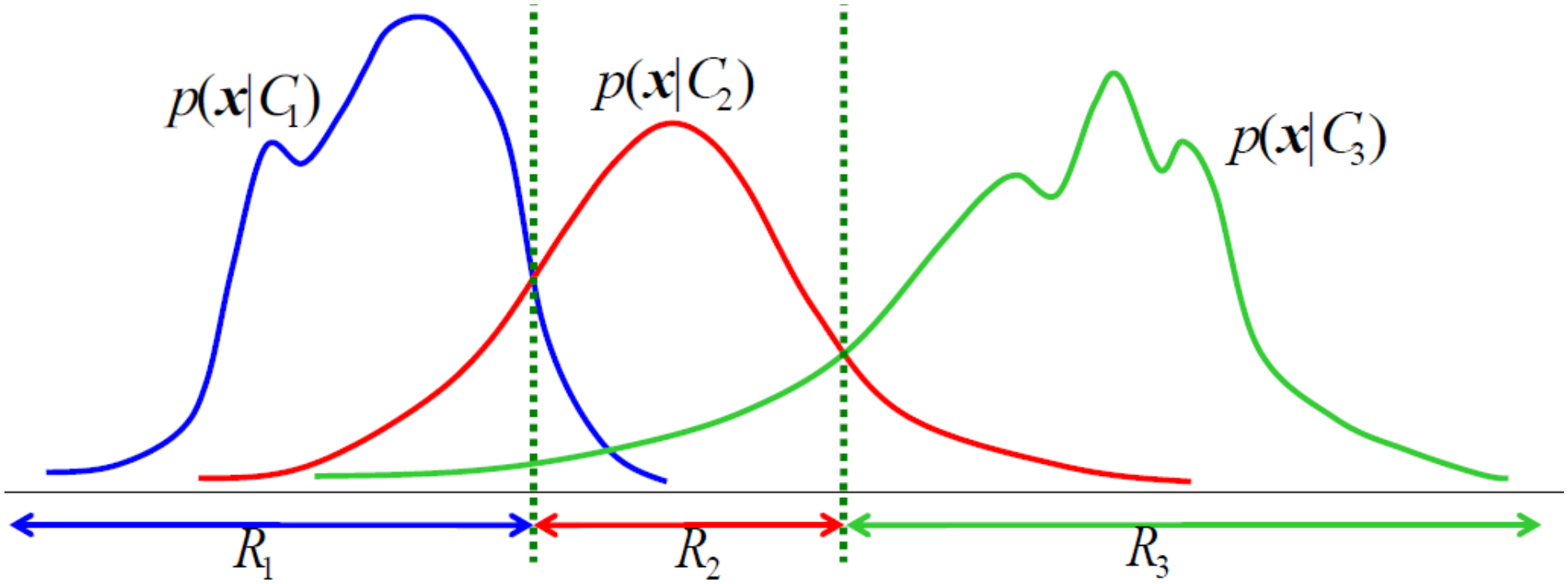

$p(\mathbf{x}|C_1) = p(\mathbf{x}|C_2)$
$p(\mathbf{x}|C_1)$
$p(\mathbf{x}|C_2)$
$\mathbf{x}_{new}$
R_1
R_2

- If $P(C_1) \neq P(C_2), P(C_1) = \alpha P(C_2), \alpha > 1$



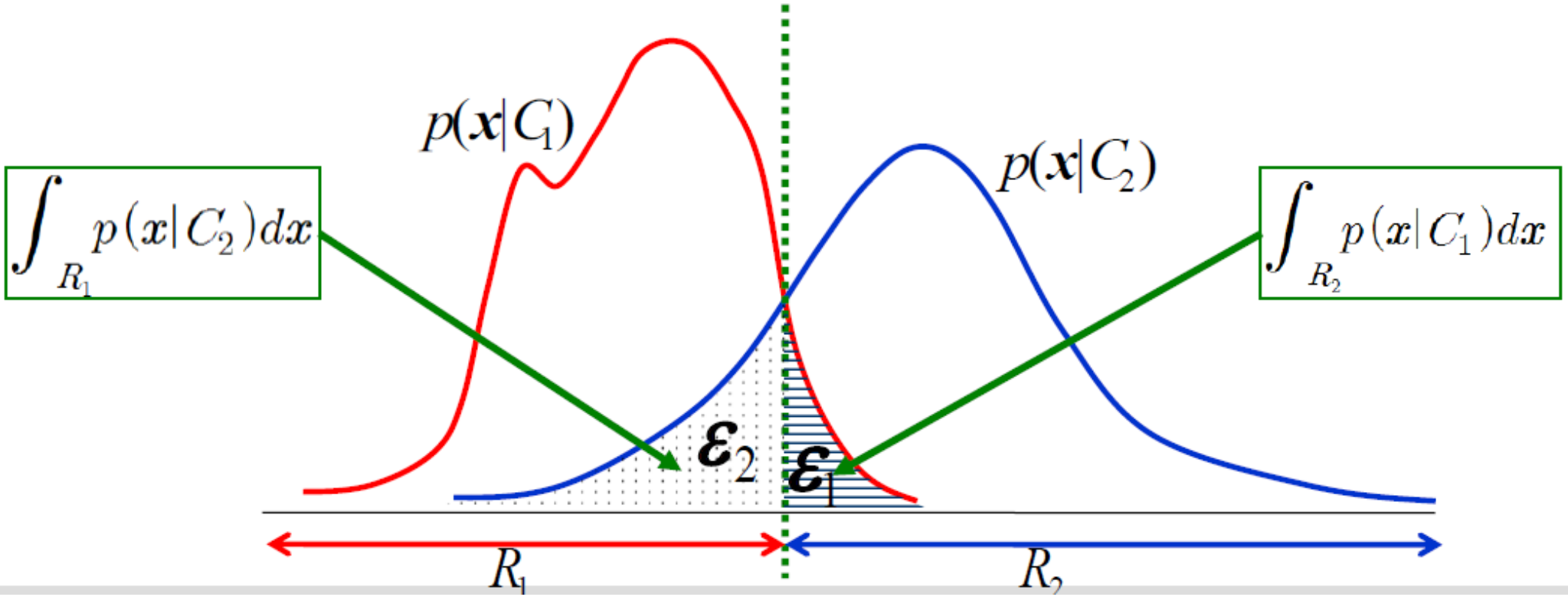
Bayesian Classifier for Multiple Class

- For each class $C_i \rightarrow g_i(x) = p(x|C_i)P(C_i)$
- Class label $y(x)$ is $\rightarrow y(x) = argmax_i\{g_i(x)\}$
- If $P(C_1) = P(C_2) = P(C_3)$



Probability of Error for Two Class

- $P_{err} = Prob(x \in \Re_2, x \in C_1) + Prob(x \in \Re_1, x \in C_2)$
- $P_{err} = P(C_1) \int_{\Re_2} p(x|C_1)dx + P(C_2) \int_{\Re_1} p(x|C_2)dx$
- $P_{err} = P(C_1)\varepsilon_1 + P(C_2)\varepsilon_2$



- $P_{err} = P(C_1) \int_{\Re_2} p(x|C_1)dx + P(C_2) \int_{\Re_1} p(x|C_2)dx$
- $P_{err} = P(C_1)\{1 - \int_{\Re_1} p(x|C_1)dx\} + P(C_2) \int_{\Re_1} p(x|C_2)dx$
- $P_{err} = P(C_1) + \int_{\Re_1} p(x|C_2)P(C_2) - p(x|C_1)p(C_1)dx$

- The minimum error can be achieved when
  - $p(x|C_2)P(C_2) = p(x|C_1)P(C_1)$

- The result of LRT
  - $g_{LRT}(x) = \frac{p(x|C_1)}{p(x|C_2)} - \frac{p(C_2)}{p(C_1)} = 0$

Conclusion

- Pattern Recognition = To find $P(C_k|x_{new})$.
- Generative approach
  - $\rightarrow Estimate\ p(x|C_k)$
  - Parametric
    - Presume $p(C_k)$, estimate from parameter
    - Bayesian
  - Non-parametric
    - $p(x) = \frac{1}{V} \times \frac{K}{N} = \frac{1}{N^n} \times \frac{K}{N}$
    - Fix V
      - Kernel Density Estimation
        - Parzen Window
        - Gaussian Window
    - Fix K
      - KNN
- Discriminative
  - $\rightarrow Find\ Decision\ Boundary\ directly$
  - LDA, SVM

References

- Pattern Recognition and Machine Learning, Christopher M. Bishop
- 패턴인식과 기계학습, 박혜영, 이관용 저
- https://en.wikipedia.org/wiki/Pattern_recognition
- https://en.wikipedia.org/wiki/Bayes_classifier
- http://www.aistudy.com/math/likelihood.htm
- http://norman3.github.io/prml/
- What is the difference between data mining, machine learning and pattern recognition?
- 베이지안 추론 - brunch
- http://artint.info/html/ArtInt_181.html
- http://www.cs.cornell.edu/courses/cs4780/2017sp/lectures/lecturenote05.html
- https://psi.engr.tamu.edu/courses/ - Lecture Notes, Pattern recognition
- Pattern Classification, R. O. Duda, P. E. Hart, D. G. Stork

Thank You !