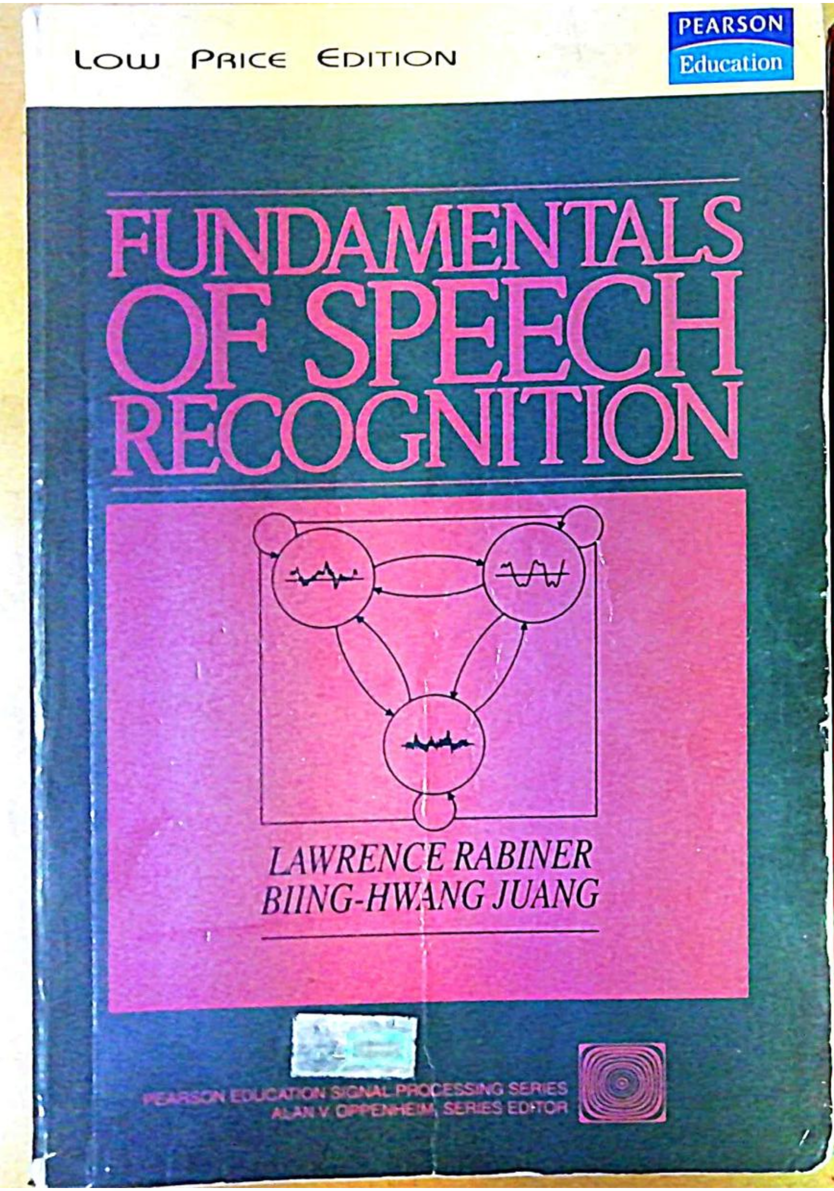
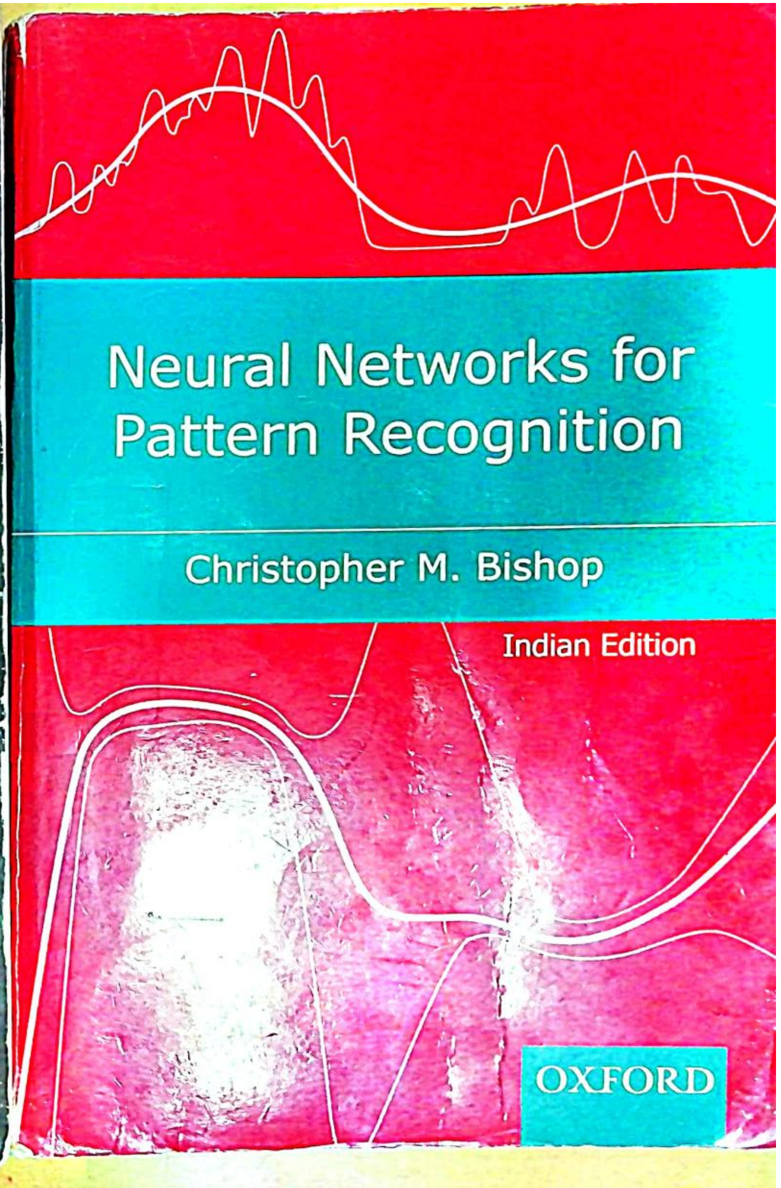
Statistical Pattern Recognition

Books



LOW PRICE EDITION
PEARSON
Education
FUNDAMENTALS OF SPEECH RECOGNITION
LAWRENCE RABINER
BIING-HWANG JUANG
PEARSON EDUCATION SIGNAL PROCESSING SERIES
ALAN V. OPPENHEIM, SERIES EDITOR

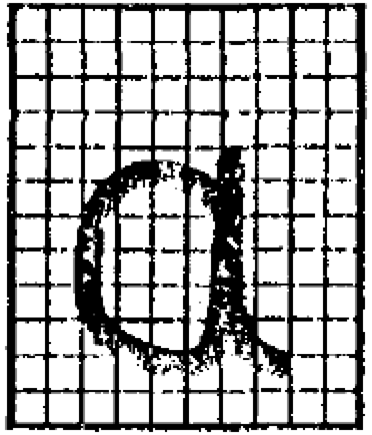


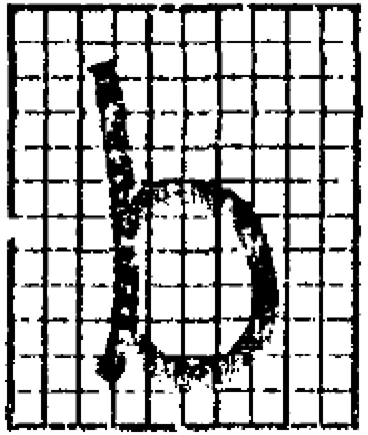
Neural Networks for Pattern Recognition
Christopher M. Bishop
Indian Edition
OXFORD

Overview

- Pattern Classification
- Preprocessing & Feature Extraction
- Curse of Dimensionality
- Classification & Regression
- Estimation of Mapping Function
- Generalization
- Model Complexity (Bias & Variance)
- Regularization
- Baye’s Theorem (Baye’s Rule)
- Inference & Decision
- Decision Boundaries
- Probability Density Estimation

Pattern Classification





Pattern classification (Ex: character Recognition)

$\boxed{a}$ $\boxed{b}$ => Images 256×256

Total possible Images = $2^{8\times256\times256} \approx 10^{158000}$

Feature => $\frac{\text{height}}{\text{width}} = \tilde{x}_1$

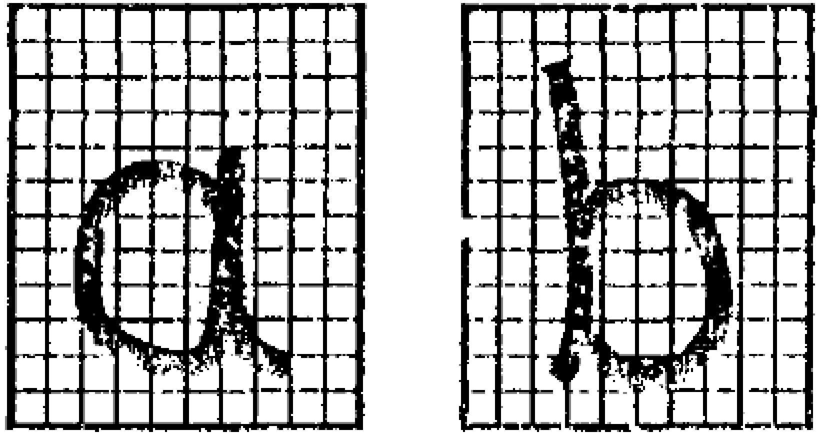
Preprocessing

Raw data (Pixels) ⟶ Features ($\frac{\text{height}}{\text{width}}$ $\tilde{x}_1$)

Overlapping of features => Misclassification

Point of Intersection of Two Histograms => Minimum Error in Classification

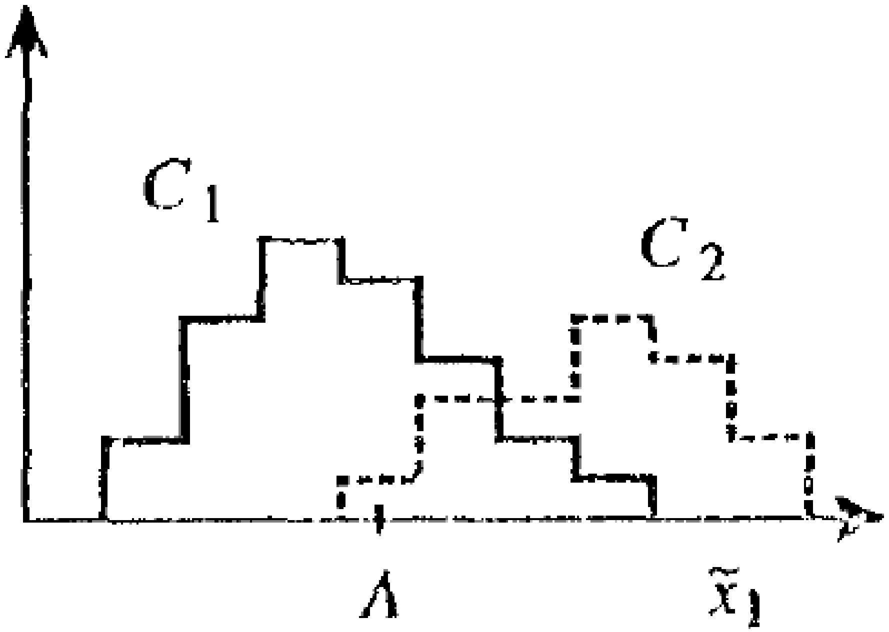
Pattern Classification (Cont..)



- Preprocessing
  - ✓ Raw data → Features
  - ✓ Pixel values → Height/width (x1)
  - ✓ Scale & Translation invariance

- Overlapping of features → Misclassification

- Point of intersection of two histograms → Minimum error in misclassification



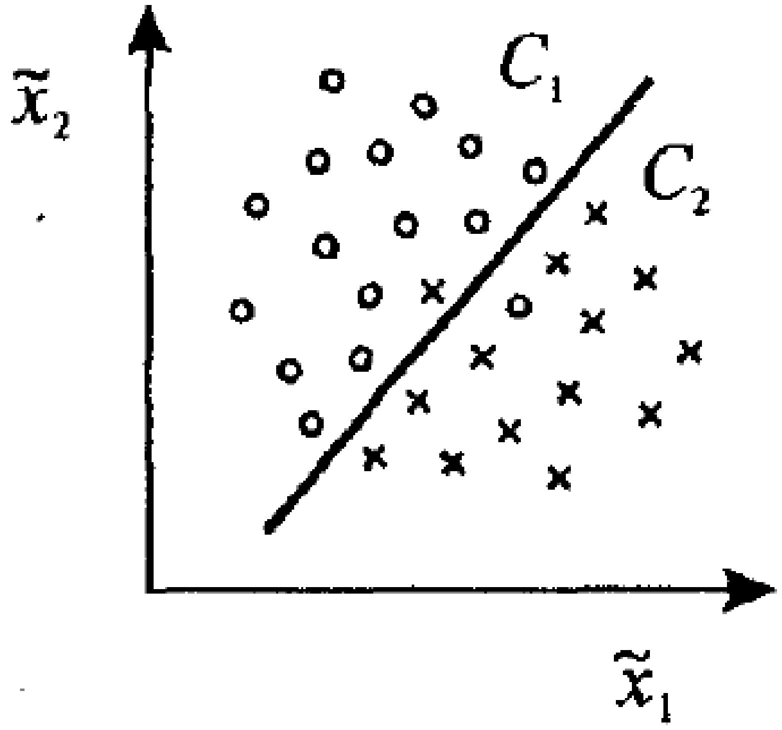
Pattern Classification (Cont..)

- Preprocessing
  - ✓ Raw data → Features
  - ✓ Pixel values → Height/width (x1)

- Overlapping of features → Misclassification

- Point of intersection of two histograms → Minimum error in misclassification

- Increase in number of features →Reduce misclassification



Curse of Dimensionality

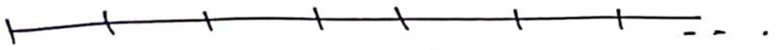
Curse of Dimensionality

d-features $\Rightarrow$ d-dimensional Feature Vector

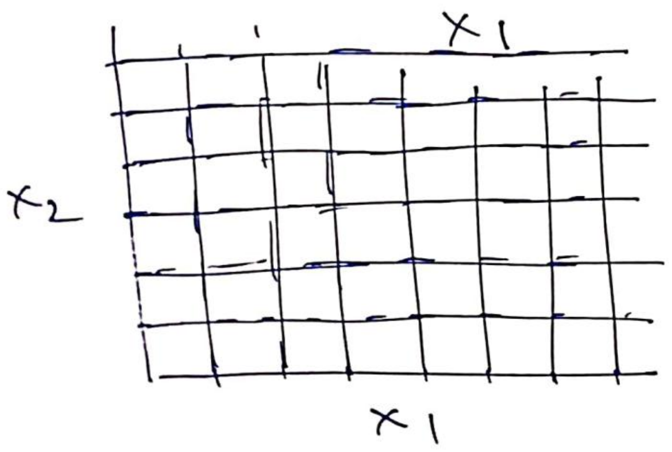
Each dimension (Feature) $\rightarrow$ M divisions

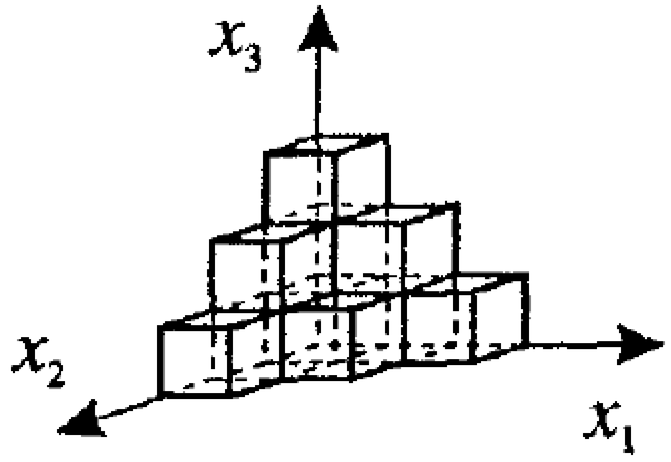
cells $\Rightarrow$ M^d

$d=1$ $\Rightarrow$ (M)



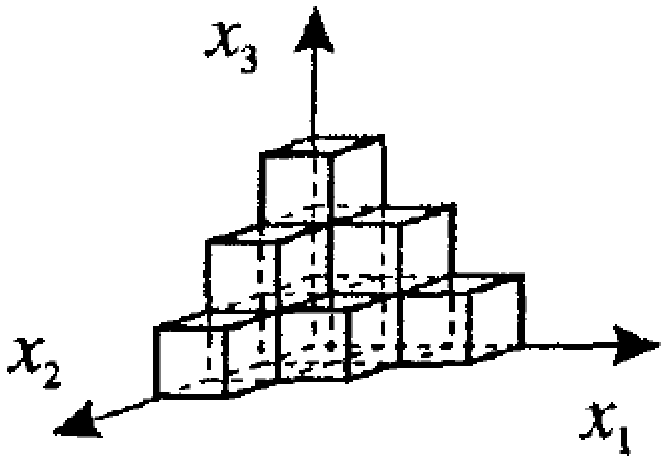
$d=2$ $\Rightarrow$ (M^2)





Curse of Dimensionality (Cont..)

- Concept of intrinsic Dimensionality
  - ✓ Input data points are correlated and restricted to subspace of lower dimensionality
- Interpolation
  - ✓ Output varies smoothly as a function of input variables
- Dimensionality Reduction
  - ✓ Limited & fixed size datasets



Classification & Regression

- ❑Function Approximation
  - ✓Classification (Probabilities of memberships of different classes)
  - ✓Regression (Function defined in terms of average over random quantity)

- ❑Classification : Discrete Classes
  - ✓Speech Recognition : 40 Classes (Phoneme as a class)
  - ✓Character Recognition : 26 Classes (Each alphabet as a class)

- ❑Regression : Continuous Output
  - ✓Stock price prediction
  - ✓Weather prediction

Estimation of Mapping Function

- Function : The underlying function is modeled using polynomial curve fitting

  $y(x) = w_0 + w_1\, x + w_2\, x^2 + \ \ldots\ldots + w_m\, x^m$

  y = f(x ; w)

- Input : $x_1, x_2, \ldots\ldots.. x_n$

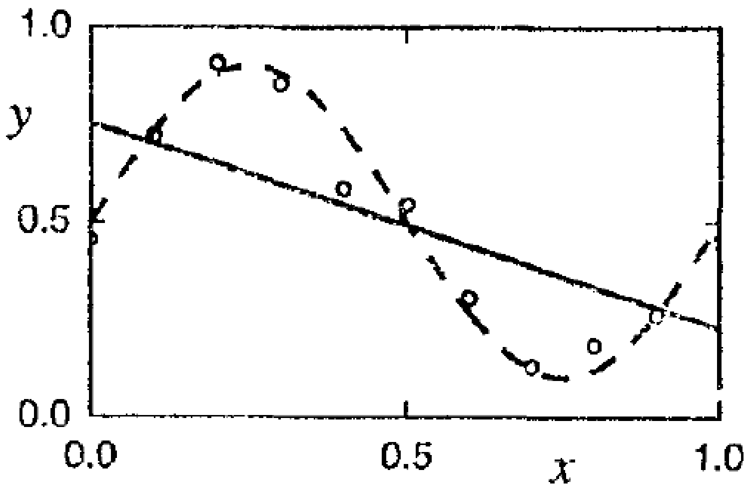
- Output (Target) : $t_1, t_2, \ldots\ldots.. t_n$

- h(x) = 0.5 + 0.4*sin(2πx)

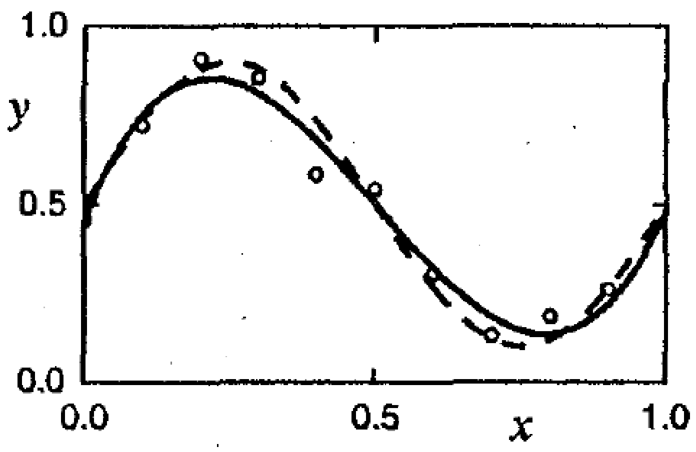
- Xn → h(x) + Random noise

- $E^{rms} = \sqrt{\sum_{n=1}^{N}(y(x_n; w^*) - t_n)^2}$

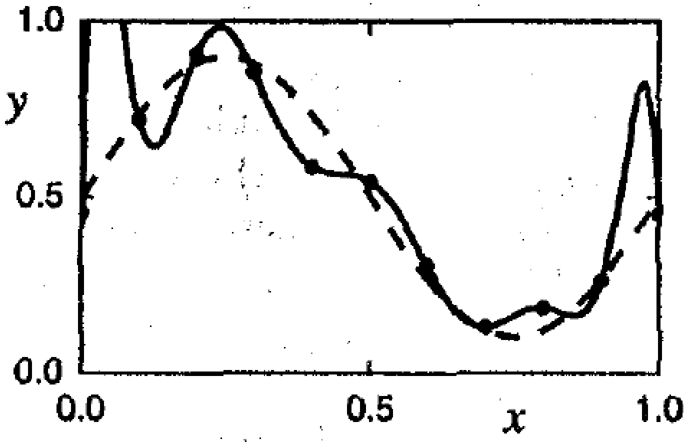
Estimation of Mapping Function



Simple Model (M=1)
Linear function



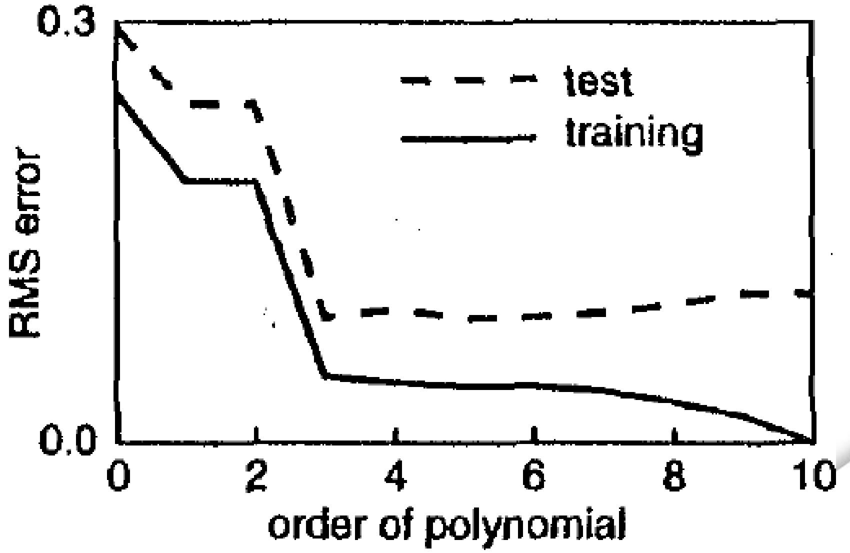
Optimal Model (M=3)
Cubic function



Complex Model (M=10)
10^{th} Order Non-linear function

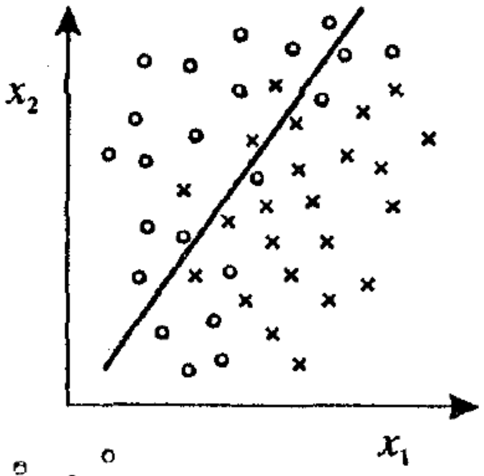
Estimation of Mapping Function (Cont..)

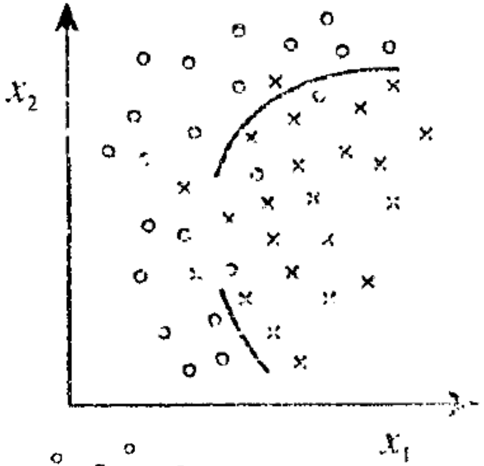
- Generalization
- Degrees of freedom → # free variables (w0, w1, w2, ….)
- Lower degrees of freedom → Higher Bias (under fitting)
- Higher degrees of freedom → Higher variance (over fitting, noisy)
- Generalization : Trade-off between Bias & Variance
- Good generalization : Low Bias & Low Variance

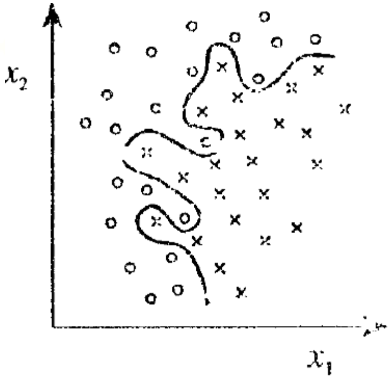


Model Complexity : Generalization & Regularization

- ❑ Best Generalization
  - ✓ Complexity of model is neither too small not too large
  - ✓ Optimal complexity
- ❑ Optimum Generalization → Controlling the effective complexity
  - ✓ Polynomial fit with M = 1 (simple model with poor fit to data)
  - ✓ M = 3 → Cubic polynomial (Optimal model with best generalization)
  - ✓ M =10 → Complex model over-fits the data
- ❑ Effective complexity → Adding penalty term to error function







Effective complexity => Adding Penalty to Error function

$\tilde{E} = E + \nu\Omega$; $\Omega \to$ Regularization Term

$\to$ depends on mapping fn.

$$\Omega = \frac{1}{2}\int\left(\frac{d^2y}{dx^2}\right)^2 dx$$

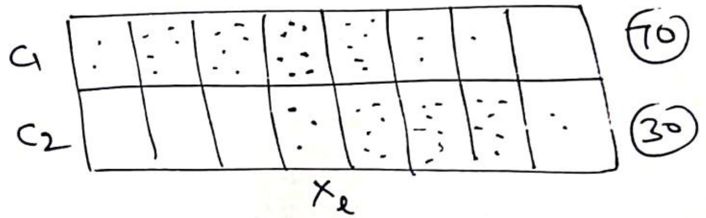
Large variance (Noise) => $\Omega \uparrow$

Baye's Theorem (Baye's Rule)

Baye's Theorem

$C_k \longrightarrow k^{th}$ class

$X \longrightarrow$ Data Vector



$P(C_k, x) \Rightarrow$ Joint probability of x & C_k

$$= P(C_k|x)\, P(x) = P(x|C_k)\, P(C_k)$$

$$P(C_1) = \frac{70}{100} \quad ; \quad P(C_2) = \frac{30}{100}$$

$$P(C_k, x_l) < \begin{matrix} P(C_1, x_l) - \frac{7}{100} \\ P(C_2, x_l) - \frac{3}{100} \end{matrix}$$

$$P(C_1, x_l) = P(x_l|C_1)\, P(C_1) = \frac{7}{70} \cdot \frac{70}{100} = \frac{7}{100}$$

$$P \cancel{(C_2, x_l)} = P(C_1|x_l)\, P(x_l) = \frac{7}{10} \cdot \frac{10}{100} = \frac{7}{100}$$

$$P(C_2, x_l) = P(C_2|x_l) \cdot P(x_l) = \frac{3}{10} \cdot \frac{10}{100} = \frac{3}{100}$$

$$= P(x_l|C_2)\, P(C_2) = \frac{3}{30} \cdot \frac{30}{100} = \frac{3}{100}$$

Baye's Theorem (Baye's Rule) Cont..

$$P(c_k, x_l) = P(c_k|x_l)\,P(x_l) = P(x_l|c_k)\,P(c_k)$$

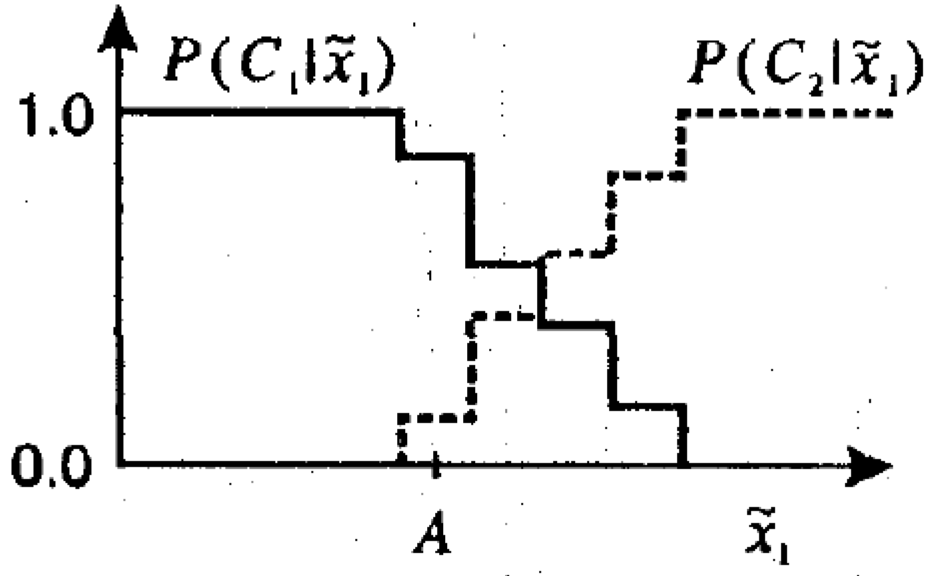
$$P(c_k|x_l) = \frac{P(x_l|c_k)\cdot P(c_k)}{P(x_l)}$$

$P(c_k|x_l) \longrightarrow$ Posterior probability

$P(x_l|c_k) \longrightarrow$ class conditional probability

$P(c_k) \longrightarrow$ Prior probability

$P(x_l) \longrightarrow$ Normalization factor (unconditional probability)



Baye's Theorem (Baye's Rule) Cont..

$P(c_k | x_l)$ — Posterior probability

$$\sum_{k=1}^{K} P(c_k | x_l) = 1$$

$$\sum_{k=1}^{K} \frac{P(x_l | c_k)\, P(c_k)}{P(x_l)} = 1$$

$$\Rightarrow \sum_{k=1}^{K} P(x_l | c_k)\, P(c_k) = P(x_l)$$

<u>Two-class problem</u>

$$P(c_1 | x_l) + P(c_2 | x_l) = 1$$

$$P(x_l) = P(x_l | c_1)\, P(c_1) + P(x_l | c_2)\, P(c_2)$$

Inference & Decision

Inference $\longrightarrow$ Estimation of posterior probabilities

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

$P(x|C_k)$ => Prob density estimation of a specific class

$P(C_k)$ => Prior probability estimated from general statistics

$P(x)$ => Normalization factor

Decision Rule => Minimum Probability of Misclassification

Baye's Rule => Class label assign to Max posterior prob.

Label $= C_k \Rightarrow P(C_k|x) > P(C_j|x) \;\; \forall\; j \neq k$

$$\text{Posterior Prob} = \frac{\text{Likelihood} \times \text{Prior prob}}{\text{Normalized factor}}$$

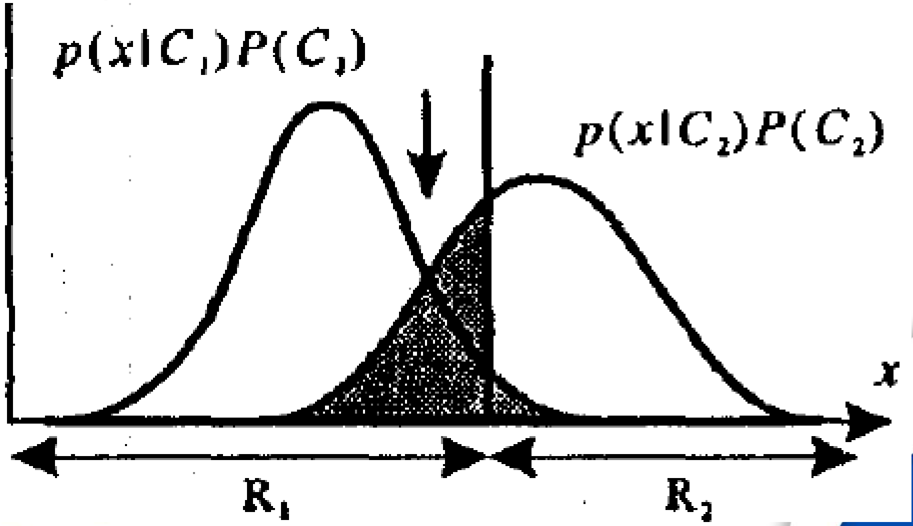
Decision Boundaries

Decision towards class C_k $\quad p(C_k|x) > p(C_j|x) \quad \forall j \neq k$

$$P_e = p(x \in R_2, C_1) + p(x \in R_1, C_2)$$

$$= p(x \in R_2 | C_1)\, p(C_1) + p(x \in R_1 | C_2)\, p(C_2)$$

$$= \int_{R_2} p(x|C_1)\, p(C_1)\, dx + \int_{R_1} p(x|C_2)\, p(C_2)\, dx$$



$$P_{correct} = \sum_k p(x \in R_k, C_k)$$

$$= \sum_k p(x \in R_k | C_k)\, p(C_k)$$

$$= \sum_k \int_{R_k} p(x|R_k)\, p(C_k)\, dx$$