Supervised Machine Learning: Pattern Classification

Bayes Classifier

Classification using Reference Template Methods

- For a test example, a distance measure is computed with the reference template of each class
- The class of the reference template with least distance is assigned to the test pattern
- When mean vector and covariance matrix is used as reference template for each class, Mahalanobis distance is used
- Mahalanobis distance gives the notion that distance measure is computed between a test example and the distribution (density) of a class
  - Distribution (density) of class: All the training examples are drawn from that distribution
  - Density here is normal (Gaussian) density
- In other way, we are interested to estimate probability of class, $P(C_i \mid \mathbf{x})$
  - Given the test example $\mathbf{x}$, what is the probability that it belongs to i^{th} class (C_i)
- Solution: Bayes classifier

Bayes Classifier

- Let $C_1, C_2, \ldots, C_i, \ldots, C_M$ be the M classes
  - Each class has N_i number of training examples
- Given: a test example $\mathbf{x}$
- To Compute:
  - Probability of class, $P(C_i \mid \mathbf{x})$
- Bayes decision rule:

Posterior Probability of a class

Prior

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i)P(C_i)}{P(\mathbf{x})}$$

  - Prior: Prior information of a class
    - Example: Human data – Each person is represented using height and weight
    - Assume that data is collected from primary school
      - Adult: Teachers and staff
      - Child: Students
    - What is the prior information about persons in primary school?
      - Probability of Child is more than Adult
    - If the human data is collected irrespective of any location
      - Prior probabilities of Adult and Child are same

Bayes Classifier

- Let C_1, C_2, …, C_i, …, C_M be the M classes
  - Each class has N_i number of training examples
- Given: a test example $\mathbf{x}$
- To Compute:
  - Probability of class, $P(C_i \mid \mathbf{x})$
- Bayes decision rule:

Posterior Probability of a class **Likelihood** **Prior**

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i) P(C_i)}{P(\mathbf{x})}$$

Total probability

  - Prior: Prior information of a class $P(C_i) = \frac{N_i}{N}$
    - where, N is total number of training examples
  - Likelihood of a class: Given the training data of a class (C_i), what is the likelihood that $\mathbf{x}$ is coming that class
    - It follows the distribution of the data of a class
  - Total probability: Evidence/probability that $\mathbf{x}$ exists

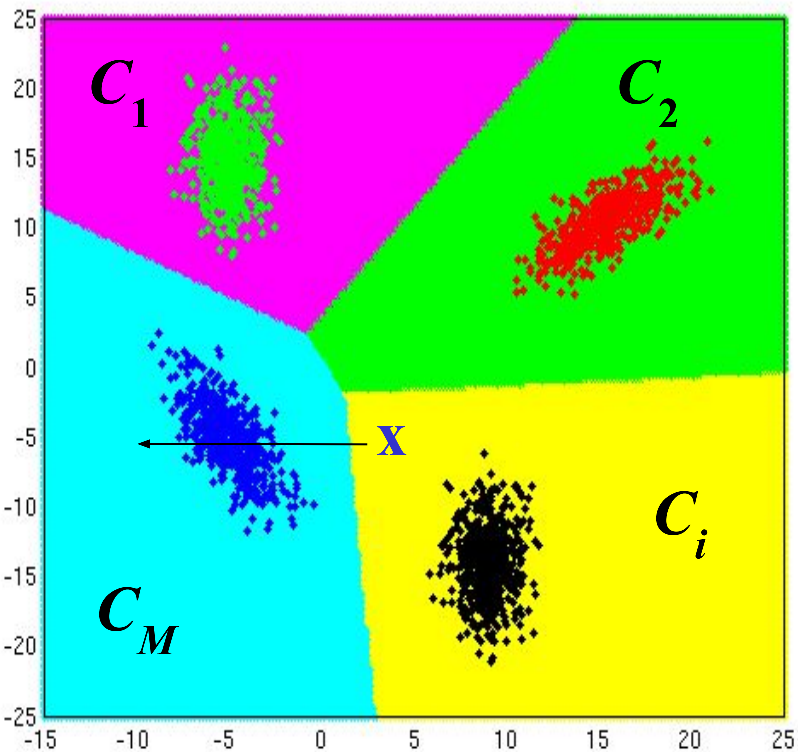
$$p(\mathbf{x}) = \sum_{i=1}^{M} p(\mathbf{x} \mid C_i) P(C_i)$$

    - Out of all the samples, what is the probability of the sample we are looking at

---

Class label for $\mathbf{x}$ = $\arg\max_i P(C_i \mid \mathbf{x}) \qquad i = 1, 2, \ldots, M$

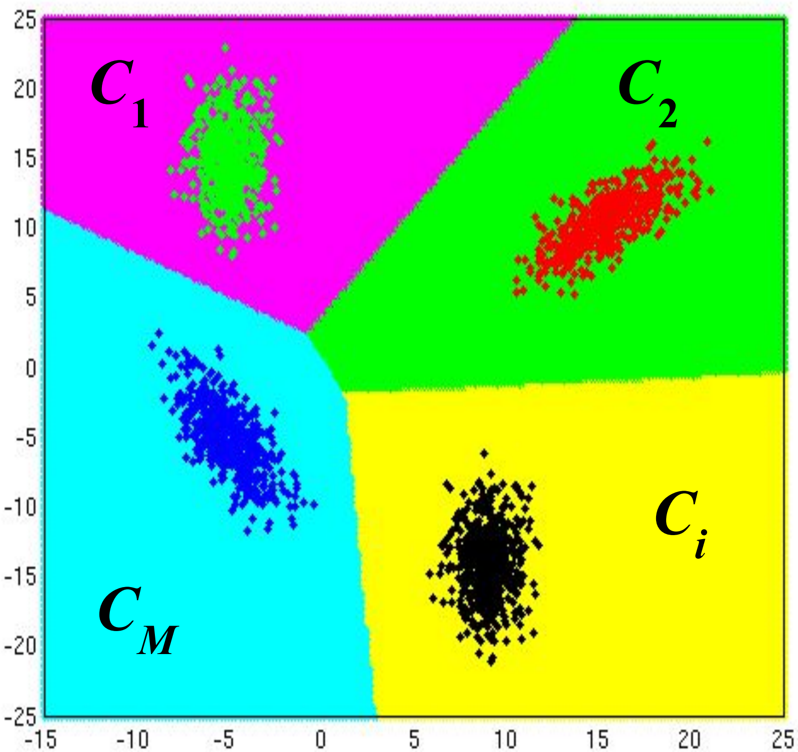
Probability Theory and Bayes Rule



- $P(A)$: Probability of an event A
- The sample space is partitioned into C_1, C_2, …, C_i, …, C_M where each partitions are disjoint
  - Example:
    - Data space is sample space
    - Each class is my partitions
- Let $\mathbf{x}$ be an event defined in sample space
  - Example: A finite data points (training data) are the event $\mathbf{x}$
- $P(\mathbf{x})$: Total probability i.e. joint probability of $\mathbf{x}$ and C_i, $P(\mathbf{x}, C_i)$, for all i

$$P(\mathbf{x}) = \sum_{i=1}^{M} p\left(\mathbf{x}, C_i\right)$$

Probability Theory and Bayes Rule



- Conditional probability:

$$p(\mathbf{x} \mid C_i) = \frac{p(\mathbf{x}, C_i)}{P(C_i)} \qquad (1)$$

- Rewriting (1)

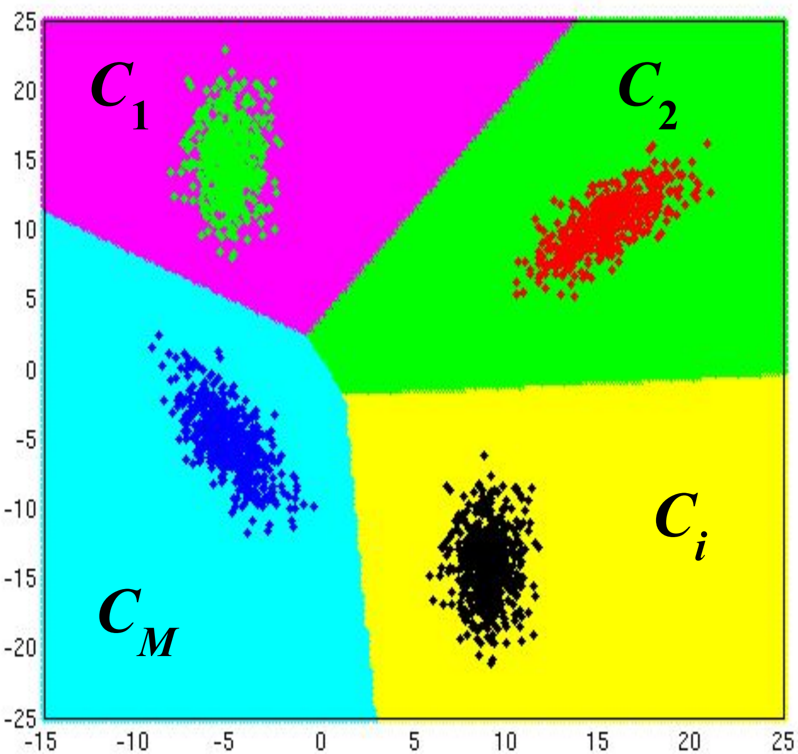
$$p(\mathbf{x}, C_i) = p(\mathbf{x} \mid C_i)P(C_i) \qquad (3)$$

- $P(\mathbf{x})$: Total probability i.e. joint probability of $\mathbf{x}$ and C_i, $P(\mathbf{x}, C_i)$, for all i

$$P(\mathbf{x}) = \sum_{i=1}^{M} p(\mathbf{x}, C_i) = \sum_{i=1}^{M} p(\mathbf{x} \mid C_i)P(C_i)$$

- $P(\mathbf{x})$ is marginal probability – probability of $\mathbf{x}$ is obtained by marginalising over all the events C_i, where $i=1,2,\ldots,M$

Probability Theory and Bayes Rule



$$p(\mathbf{x} \mid C_i) = \frac{p(\mathbf{x}, C_i)}{P(C_i)} \qquad (1)$$

$$p(\mathbf{x}, C_i) = P(C_i \mid \mathbf{x}) P(\mathbf{x}) \qquad (4)$$

$$p(\mathbf{x}, C_i) = p(\mathbf{x} \mid C_i) P(C_i) \qquad (3)$$

$$p(\mathbf{x}, C_i) = P(C_i \mid \mathbf{x}) P(\mathbf{x}) \qquad (4)$$

- From (3) and (4): $p(\mathbf{x}, C_i) = P(C_i \mid \mathbf{x}) P(\mathbf{x}) \qquad (4)$

- Bayes decision rule:

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i) P(C_i)}{P(\mathbf{x})}$$

Bayes Classifier

- Let $C_1, C_2, \ldots, C_i, \ldots, C_M$ be the M classes
  - Each class has N_i number of training examples
- Given: a test example $\mathbf{x}$
- To Compute:
  - Probability of class, $P(C_i \mid \mathbf{x})$
- Bayes decision rule:

Posterior Probability of a class

Likelihood

Prior

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i) P(C_i)}{P(\mathbf{x})}$$

Evidence

  - Likelihood of a class (Class conditional density) follows the distribution of the data of a class
  - Computation of likelihood of a class (class conditional density) depends on the
    - distribution of the data (i.e. data follows some distribution) and
    - the parameters of that distribution
- Bayes decision rule can be given as $P(\boldsymbol{\theta}_i \mid \mathbf{x}) = \dfrac{p(\mathbf{x} \mid \boldsymbol{\theta}_i) P(C_i)}{P(\mathbf{x})}$
  - $\boldsymbol{\theta}_i$ is the parameters of the distribution of class C_i *estimated from training data* of that class

Parameter Estimation from Training Data: Maximum Likelihood (ML) Method

- Given: Training data for a class C_i: having N_i samples

$$\mathcal{D}_i = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_{Ni}\}, \quad \mathbf{x}_n \in \mathbb{R}^d$$

- Data of a class C_i is sampled from a distribution, which is defined by parameter vector: $\boldsymbol{\theta}_i = [\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}]^\mathsf{T}$ of that distribution
  - Data of a class C_i is now represented by parameter vector, $\boldsymbol{\theta}_i$
- Unknown: $\boldsymbol{\theta}_i$
- Likelihood of training data (Total data likelihood) for a given $\boldsymbol{\theta}_i$:

$$p(\mathcal{D}_i \mid \boldsymbol{\theta}_i) = \prod_{n=1}^{N_i} p(\mathbf{x}_n \mid \boldsymbol{\theta}_i)$$

- Log likelihood: $\mathcal{L}(\boldsymbol{\theta}_i) = \ln p(\mathcal{D}_i \mid \boldsymbol{\theta}_i) = \sum_{n=1}^{N_i} \ln p(\mathbf{x}_n \mid \boldsymbol{\theta}_i)$
  - Advantage of applying monotonous increasing function, $ln(.)$:
    - Likelihood of an example is very small value. Product of small values lead to 0. It converts product of likelihoods into sum of likelihoods
    - Simplifies computation for certain forms of distribution

Parameter Estimation from Training Data:
Maximum Likelihood (ML) Method

Parameter Estimation from Training Data: Maximum Likelihood (ML) Method

- Given: Training data for a class C_i: having N_i samples

$$\mathcal{D}_i = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_{Ni}\}, \quad \mathbf{x}_n \in \mathsf{R}^d$$

- Data of a class C_i is sampled from a distribution, which is defined by parameter vector: $\boldsymbol{\theta}_i = [\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK}]^{\mathsf{T}}$ of that distribution
  - Data of a class C_i is now represented by parameter vector, $\boldsymbol{\theta}_i$
- Unknown: $\boldsymbol{\theta}_i$
- Likelihood of training data (Total data likelihood) for a given $\boldsymbol{\theta}_i$:

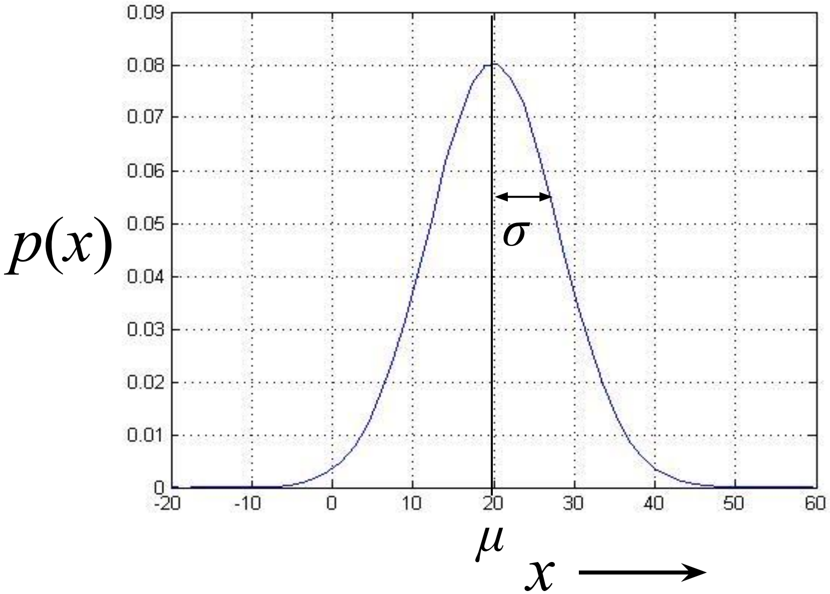
$$p(\mathsf{D}_i \mid \boldsymbol{\theta}_i) = \prod_{n=1}^{N_i} p(\mathbf{x}_n \mid \boldsymbol{\theta}_i)$$

- Log likelihood: $\mathsf{L}(\boldsymbol{\theta}_i) = \ln p(\mathsf{D}_i \mid \boldsymbol{\theta}_i) = \sum_{n=1}^{N_i} \ln p(\mathbf{x}_n \mid \boldsymbol{\theta}_i)$
- Choose the parameters for which the total data likelihood (log likelihood) is maximum:

$$\boldsymbol{\theta}_{i_{\mathrm{ML}}} = \arg\max_{\boldsymbol{\theta}_i} \mathsf{L}(\boldsymbol{\theta}_i)$$

Probability Distribution

- Data of a class is represented by a probability distribution
- For a class whose data is considered to be forming a single cluster, it can be represented by a normal or Gaussian distribution
- Gaussian distribution is a unimodal distribution
  - Single mode or single peak
- Univariate Gaussian distribution:
  - Univariate data means 1-dimensional data



$$p(x) = \mathrm{N}(x \mid \mu, \sigma)$$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

- μ is the mean
- σ^2 is the variance

Probability Distribution

- Data of a class is represented by a probability distribution
- For a class whose data is considered to be forming a single cluster, it can be represented by a normal or Gaussian distribution
- Gaussian distribution is a unimodal distribution
  - Single mode or single peak
- Multivariate Gaussian distribution:
  - Multivariate data means d-dimensional data
  - *Bivariate Gaussian distribution*
    - Bivariate data means 2-dimensional data

Multivariate Gaussian Distribution

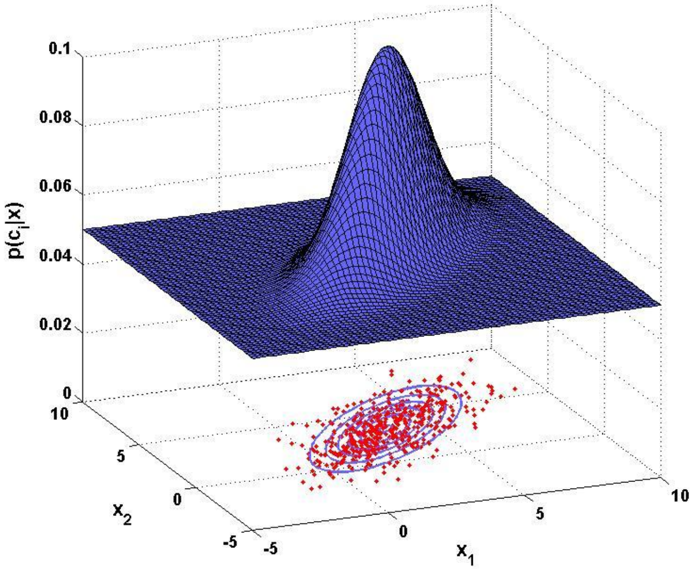
- Data in *d*-dimensional space

$$p(\mathbf{x}) = N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$= \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} \underbrace{(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}_{\text{Mahalanobis distance}} \right)$$

  - $\boldsymbol{\mu}$ is the mean vector
  - $\boldsymbol{\Sigma}$ is the covariance matrix

- Bivariate Gaussian distribution: d=2



$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} E[x_1] \\ E[x_2] \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} E\left[(x_1-\mu_1)^2\right] & E\left[(x_1-\mu_1)(x_2-\mu_2)\right] \\ E\left[(x_2-\mu_2)(x_1-\mu_1)\right] & E\left[(x_2-\mu_2)^2\right] \end{bmatrix}$$

Maximum Likelihood (ML) Method for Parameter Estimation of Multivariate Gaussian Distribution

- Given: Training data for a class C_i having N_i samples

$$\mathcal{D}_i = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_{Ni}\}, \mathbf{x}_n \in \mathrm{R}^d$$

- Data of a class C_i is coming from Gaussian distribution
  - Training data of a class C_i is represented by parameter vector: $[\boldsymbol{\mu}_i\ \boldsymbol{\Sigma}_i]^\mathsf{T}$, of Gaussian distribution
- Unknown: $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$
- Likelihood of training data (Total data likelihood) for a given $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$: $p(\mathrm{D} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \prod_{n=1}^{N_i} p(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
- Log likelihood: $\mathrm{L}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \ln p(\mathrm{D}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \sum_{n=1}^{N_i} \ln p(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
- Choose the parameters for which the total data likelihood (log likelihood) is maximum:

$$\boldsymbol{\mu}_{i_{\mathrm{ML}}}, \boldsymbol{\Sigma}_{i_{\mathrm{ML}}} = \arg\max_{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i} \mathrm{L}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

ML Method for Parameter Estimation of Multivariate Gaussian Distribution

- Parameters of Gaussian distribution of class C_i : $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$
- Likelihood for a single example, $\mathbf{x}_n$:

$$p(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_i)^\mathsf{T} \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_i)\right)$$

- Log likelihood for total training data of class C_i , $\mathcal{D}_i = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$:

$$\mathsf{L}(\boldsymbol{\mu}_\mathsf{i}, \boldsymbol{\Sigma}_\mathsf{i}) = \ln p(\mathsf{D}_\mathsf{i} \mid \boldsymbol{\mu}_\mathsf{i}, \boldsymbol{\Sigma}_\mathsf{i}) = \ln \prod_{n=1}^{N_i} p(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \sum_{n=1}^{N_i} \ln p(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$= \sum_{n=1}^{N_i} -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{d}{2}\ln 2\pi - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_i)^\mathsf{T} \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_i)$$

- Setting the derivatives of $\mathcal{L}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ w.r.t. $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ to zero, we get:

$$\frac{\partial \mathsf{L}(\boldsymbol{\mu}_\mathsf{i}, \boldsymbol{\Sigma}_\mathsf{i})}{\partial \boldsymbol{\mu}_\mathsf{i}} = \mathbf{0} \qquad \frac{\partial \mathsf{L}(\boldsymbol{\mu}_\mathsf{i}, \boldsymbol{\Sigma}_\mathsf{i})}{\partial \boldsymbol{\Sigma}_\mathsf{i}} = \mathbf{0}$$

ML Method for Parameter Estimation of Multivariate Gaussian Distribution

- Parameters of Gaussian distribution of class C_i : $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$
- Likelihood for a single example, $\mathbf{x}_n$:

$$p(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_i)^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_i) \right)$$

- Log likelihood for total training data of class C_i , $\mathcal{D}_i = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$:

$$\mathsf{L}(\boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_\mathrm{i}) = \ln p(\mathsf{D}_\mathrm{i} \mid \boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_\mathrm{i}) = \ln \prod_{n=1}^{N_i} p(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \sum_{n=1}^{N_i} \ln p(\mathbf{x}_n \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$= \sum_{n=1}^{N_i} -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{d}{2}\ln 2\pi - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_i)^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_i)$$

- Setting the derivatives of $\mathcal{L}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ w.r.t. $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ to zero, we get:

$$\boldsymbol{\mu}_{i_{\mathrm{ML}}} = \frac{1}{N_i}\sum_{n=1}^{N_i} \mathbf{x}_n \qquad \boldsymbol{\Sigma}_{i_{\mathrm{ML}}} = \frac{1}{N_i}\sum_{n=1}^{N_i} (\mathbf{x}_n - \boldsymbol{\mu}_{i_{\mathrm{ML}}})(\mathbf{x}_n - \boldsymbol{\mu}_{i_{\mathrm{ML}}})^{\mathsf{T}}$$

Bayes Classifier with Unimodal Gaussian Density – Training Process

- Let C_1, C_2, …, C_i, …, C_M be the M classes
- Let $\mathcal{D}_1$, $\mathcal{D}_2$, …, $\mathcal{D}_i$, …, $\mathcal{D}_M$ be the training data for M classes
- Let each class having N_i number of training examples
- Estimate the parameters
  - $\boldsymbol{\theta}_1 = [\boldsymbol{\mu}_1\ \boldsymbol{\Sigma}_1]^\top$,
  - $\boldsymbol{\theta}_2 = [\boldsymbol{\mu}_2\ \boldsymbol{\Sigma}_2]^\top$,
  - …,
  - $\boldsymbol{\theta}_i = [\boldsymbol{\mu}_i\ \boldsymbol{\Sigma}_i]^\top$,
  - …,
  - $\boldsymbol{\theta}_M = [\boldsymbol{\mu}_M\ \boldsymbol{\Sigma}_M]^\top$ for each of the classes
- Number of parameters to be estimated for each class is dependent on dimensionality of the data space d
  - Number of parameters for each class: $d + (d(d+1))/2$

Bayes Classifier with Unimodal Gaussian Density – Training Process

- Let C_1, C_2, …, C_i, …, C_M be the M classes
- Let $\mathcal{D}_1$, $\mathcal{D}_2$, …, $\mathcal{D}_i$, …, $\mathcal{D}_M$ be the training data for M classes
- Compute sample mean vector and sample covariance matrix from training data of class 1, $\boldsymbol{\theta}_1 = [\boldsymbol{\mu}_1\ \boldsymbol{\Sigma}_1]^\mathsf{T}$
- Compute sample mean vector and sample covariance matrix from training data of class 2, $\boldsymbol{\theta}_2 = [\boldsymbol{\mu}_2\ \boldsymbol{\Sigma}_2]^\mathsf{T}$,
- …,
- Compute sample mean vector and sample covariance matrix from training data of class M, $\boldsymbol{\theta}_M = [\boldsymbol{\mu}_M\ \boldsymbol{\Sigma}_M]^\mathsf{T}$

Bayes Classifier with Unimodal Gaussian Density: Classification

- For a test example $\mathbf{x}$:
  - likelihood of $\mathbf{x}$ generated from each of the classes $p(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and class posterior probability $P(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i|\mathbf{x})$ is computed

$$P(\boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_\mathrm{i} \mid \mathbf{x}) = \frac{p\left(\mathbf{x}\middle|\boldsymbol{\mu}_\mathrm{i}, \boldsymbol{\Sigma}_\mathrm{i}\right) P(C_i)}{P(\mathbf{x})}$$

Bayes Classifier with Unimodal Gaussian Density: Classification

- For a test example $\mathbf{x}$:
  - likelihood of $\mathbf{x}$ generated from each of the classes $p(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and class posterior probability $P(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i|\mathbf{x})$ is computed

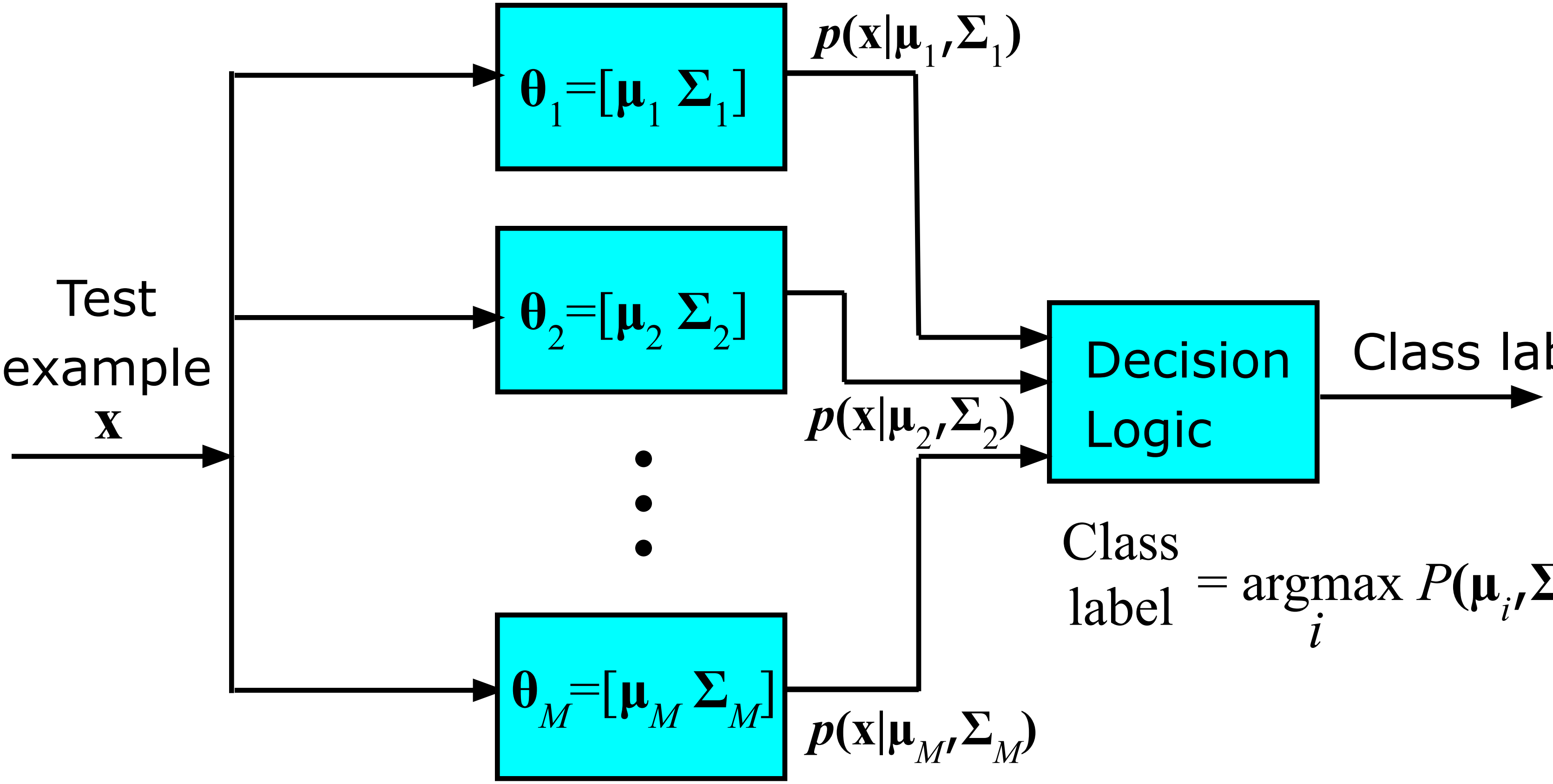
$$P(\boldsymbol{\mu}_\text{i}, \boldsymbol{\Sigma}_\text{i} \mid \mathbf{x}) = \frac{p\left(\mathbf{x}\middle|\boldsymbol{\mu}_\text{i}, \boldsymbol{\Sigma}_\text{i}\right)P(C_i)}{\sum_{i=1}^{M} p\left(\mathbf{x}\middle|\boldsymbol{\mu}_\text{i}, \boldsymbol{\Sigma}_\text{i}\right)P(C_i)}$$

  - Assign the label of class for which $P(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i|\mathbf{x})$ is maximum

$$\text{Class label} = \arg\max_{i} P(\boldsymbol{\mu}_\text{i}, \boldsymbol{\Sigma}_\text{i} \mid \mathbf{x})$$

Bayes Classifier with Unimodal Gaussian Density: Classification

- For a test example $\mathbf{x}$:
  - likelihood of $\mathbf{x}$ generated from each of the classes $p(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ or class posterior probability $P(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i|\mathbf{x})$ is computed
  - Assign the label of class for which $P(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i|\mathbf{x})$ is maximum



$$\text{Class label} = \underset{i}{\operatorname{argmax}}\, P(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i|\mathbf{x})$$

Illustration of Bayes Classifier with Unimodal Gaussian Density : Adult(1)-Child(0) Classification

Height	Weight	Class
90	21.5	0
95	23.67	0
100	32.45	0
116	38.21	0
98	28.43	0
108	36.32	0
104	27.38	0
112	39.28	0
121	35.8	0
92	23.56	0
152	46.8	1
178	78.9	1
163	67.45	1
173	82.9	1
154	52.6	1
168	66.2	1
183	90	1
172	82	1
156	45.3	1
161	59	1

- Training Phase:
  - Compute sample mean vector and sample covariance matrix from training data of class 0 (Child)

$$\boldsymbol{\mu}_0 = [103.60 \quad 30.66]$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 109.38 & 61.35 \\ 61.35 & 43.54 \end{pmatrix}$$

  - Prior probability for class 0 (Child):

$$P(C_0) = 10/20 = 0.5$$

Illustration of Bayes Classifier with Unimodal Gaussian Density : Adult(1)-Child(0) Classification

Height	Weight	Class
90	21.5	0
95	23.67	0
100	32.45	0
116	38.21	0
98	28.43	0
108	36.32	0
104	27.38	0
112	39.28	0
121	35.8	0
92	23.56	0
152	46.8	1
178	78.9	1
163	67.45	1
173	82.9	1
154	52.6	1
168	66.2	1
183	90	1
172	82	1
156	45.3	1
161	59	1

- Training Phase:
  - Compute sample mean vector and sample covariance matrix from training data of class 0 (Child)

$$\boldsymbol{\mu}_0 = [103.60 \quad 30.66]$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 109.38 & 61.35 \\ 61.35 & 43.54 \end{pmatrix}$$

  - Prior probability for class 0 (Child):

$$P(C_0) = 10/20 = 0.5$$

  - Compute sample mean vector and sample covariance matrix from training data of class 1 (Adult)

$$\boldsymbol{\mu}_1 = [166.00 \quad 67.12]$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 110.67 & 160.53 \\ 160.53 & 255.49 \end{pmatrix}$$

  - Prior probability for class 1 (Adult):

$$P(C_1) = 10/20 = 0.5$$

Illustration of Bayes Classifier with Unimodal Gaussian Density : Adult(1)-Child(0) Classification

$\boldsymbol{\mu}_0 = [103.60 \quad 30.66]$

$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 109.38 & 61.35 \\ 61.35 & 43.54 \end{pmatrix}$

Prior: $P(C_0) = 0.5$

Class
0

$\boldsymbol{\mu}_1 = [166.00 \quad 67.12]$

$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 110.67 & 160.53 \\ 160.53 & 255.49 \end{pmatrix}$

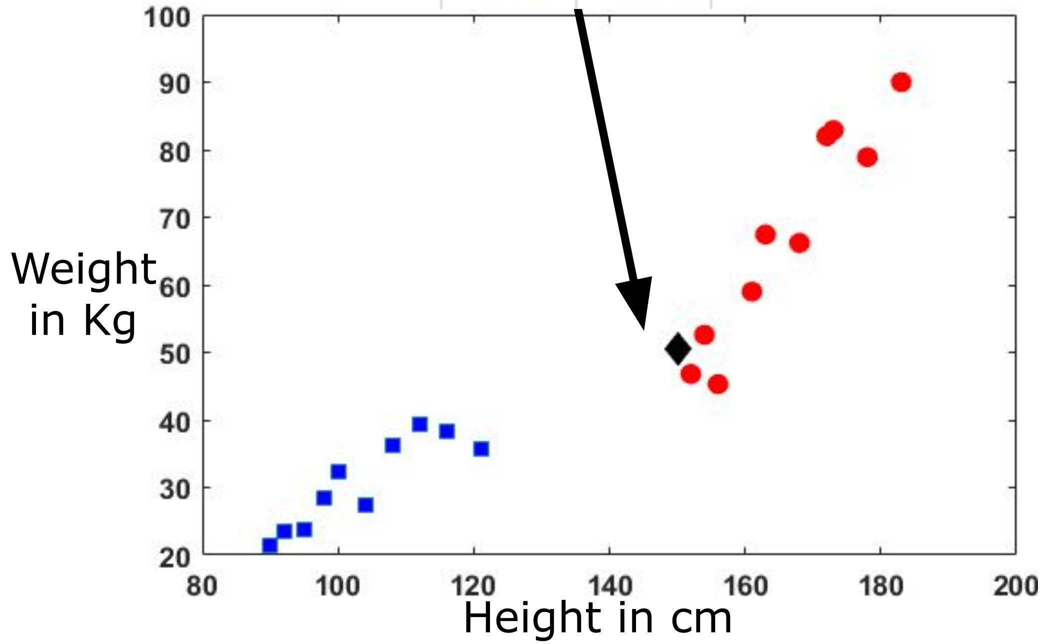
Prior: $P(C_1) = 0.5$

Class
1

- Test Phase - Classification:

Test Example, $\mathbf{x}$:

150	50.6



$$p(\mathbf{x}, C_i) = P(C_i \mid \mathbf{x})P(\mathbf{x}) \qquad (4)$$

Illustration of Bayes Classifier with Unimodal Gaussian Density : Adult(1)-Child(0) Classification

$\boldsymbol{\mu}_0 = [103.60 \;\; 30.66]$

$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 109.38 & 61.35 \\ 61.35 & 43.54 \end{pmatrix}$

$Prior: P(C_0) = 0.5$

Class
0

$\boldsymbol{\mu}_1 = [166.00 \;\; 67.12]$

$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 110.67 & 160.53 \\ 160.53 & 255.49 \end{pmatrix}$

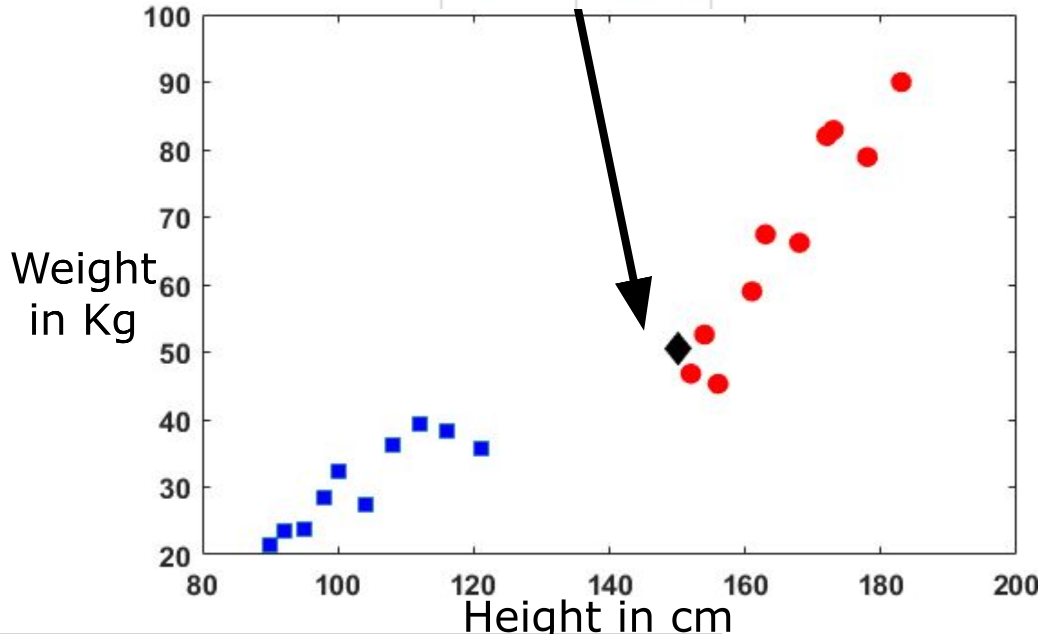
$Prior: P(C_1) = 0.5$

Class
1

- Test Phase - Classification:

Test Example, $\mathbf{x}$:

150	50.6



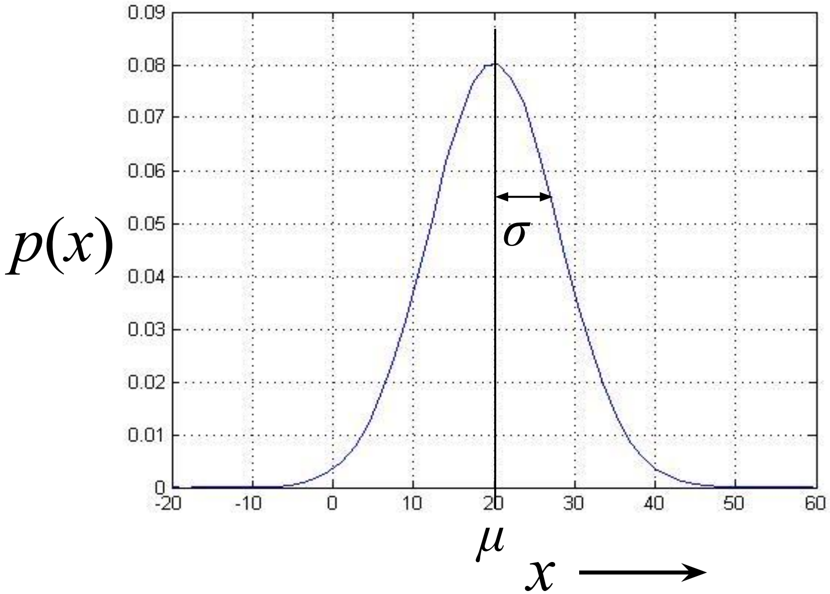
$$p(\mathbf{x}, C_i) = P(C_i \mid \mathbf{x}) P(\mathbf{x}) \qquad (4)$$

- Compute a posterior probability for class 0 (Child): $p(\mathbf{x}, C_i) = P(C_i \mid \mathbf{x}) P(\mathbf{x}) \qquad (4)$
- Compute a posterior probability for class 1 (Adult): $p(\mathbf{x}, C_i) = P(C_i \mid \mathbf{x}) P(\mathbf{x}) \qquad (4)$

Class label of $\mathbf{x}$ = Adult

Summary: Bayes Classifier with Unimodal Gaussian Density

- The relation between examples and class can be captured in a statistical model
  - Bayes classifier
  - Data is represented by any distribution
  - If the underlying distribution (density) of data is known, then Bayes classifier is the minimum error classifier
- Statistical model:
  - Example distribution: Unimodal Gaussian density
    - Univariate



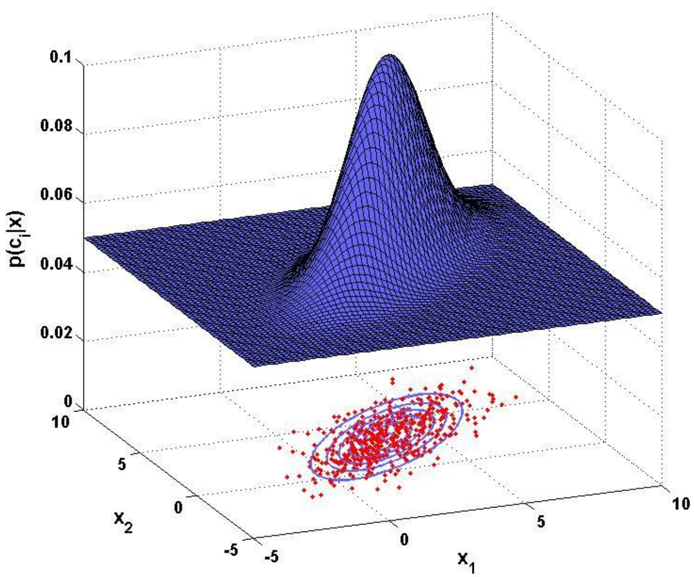
$$p(x) = \mathrm{N}(x \mid \mu, \sigma)$$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

- μ is the mean
- σ^2 is the variance

Summary: Bayes Classifier with Unimodal Gaussian Density

- The relation between examples and class can be captured in a statistical model
  - Bayes classifier
  - Data is represented by any distribution
- Statistical model:
  - Example distribution: Unimodal Gaussian density
    - Univariate
    - Multivariate (*Bivariate when the dimension is 2*)



$$p(\mathbf{x}) = N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

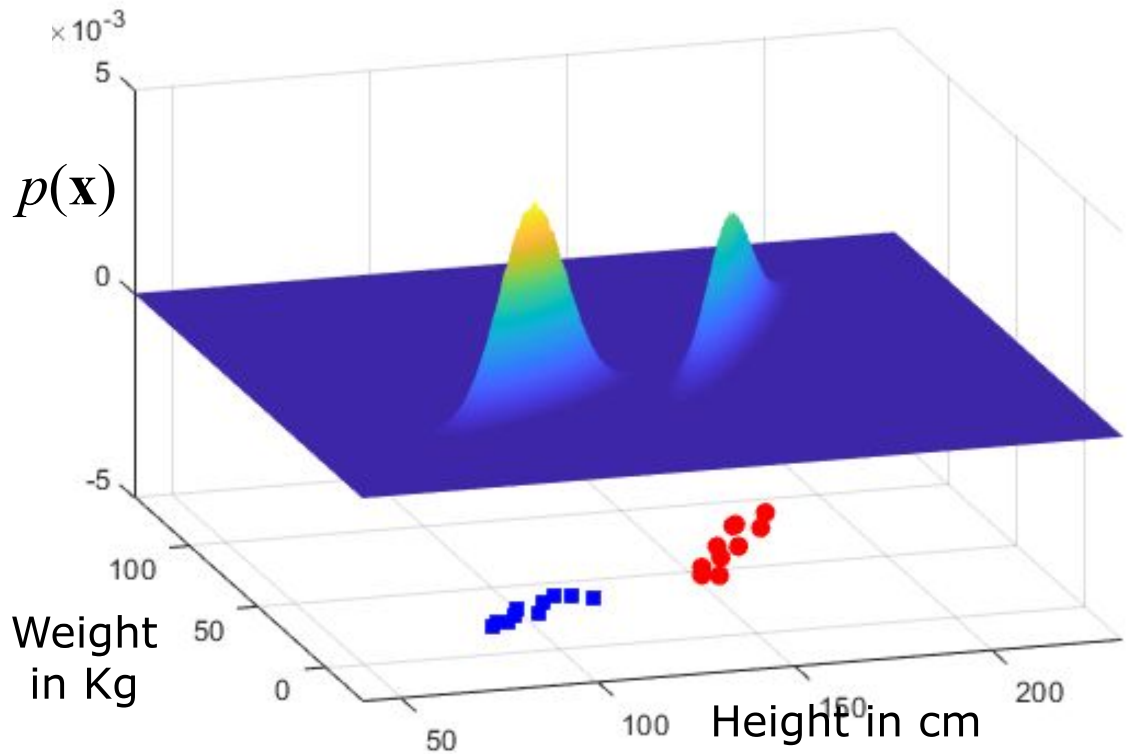
$$= \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

- $\boldsymbol{\mu}$ is the mean vector
- $\boldsymbol{\Sigma}$ is the covariance matrix

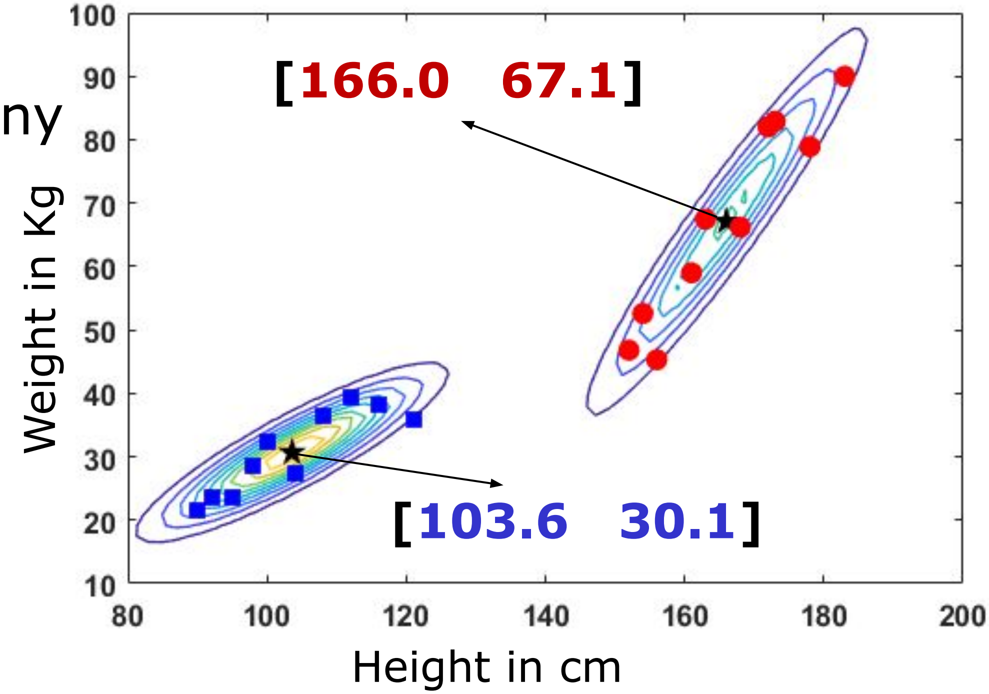
Summary: Bayes Classifier with Unimodal Gaussian Density

- The relation between examples and class can be captured in a statistical model
  - Bayes classifier
  - Data is represented by any distribution
- Statistical model:
  - Example distribution: Unimodal Gaussian density
    - Univariate
    - Multivariate



Summary: Bayes Classifier with Unimodal Gaussian Density

- The relation between examples and class can be captured in a statistical model
  - Bayes classifier
  - Data is represented by any
- Statistical model:
  - Example distribution: Unimodal Gaussian density
    - Univariate
    - Multivariate



- The real world data need not be unimodal
  - The shape of the density can be arbitrary
  - Bayes classifier?
- Multimodal density function

Text Books

1. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Third Edition, Morgan Kaufmann Publishers, 2011.

2. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 2009.

3. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.