Maximum-Likelihood & Bayesian Parameter Estimation (Sections 3.1-3.5)

Mingmin Chi

SCS Fudan University, Shanghai, China

- Data availability in a Bayesian framework
  - We could design an optimal classifier if we knew:
    - $P(\omega_i)$ (priors)
    - $P(x|\omega_i)$ (class-conditional densities)

    Unfortunately, we rarely have this complete information!

- Data availability in a Bayesian framework
  - We could design an optimal classifier if we knew:
    - $P(\omega_i)$ (priors)
    - $P(x|\omega_i)$ (class-conditional densities)

  Unfortunately, we rarely have this complete information!
- Design a classifier from a training data set
  - No problem with prior estimation
  - Samples are often too small for class-conditional estimation (large dimension of feature space!)

- Data availability in a Bayesian framework
  - We could design an optimal classifier if we knew:
    - $P(\omega_i)$ (priors)
    - $P(x|\omega_i)$ (class-conditional densities)

    Unfortunately, we rarely have this complete information!
- Design a classifier from a training data set
  - No problem with prior estimation
  - Samples are often too small for class-conditional estimation (large dimension of feature space!)
- If we know the number of parameters in advance, our general knowledge about the problem can by reduced by parameterizing the conditional density, e.g.,

- Data availability in a Bayesian framework
  - We could design an optimal classifier if we knew:
    - $P(\omega_i)$ (priors)
    - $P(x|\omega_i)$ (class-conditional densities)

    Unfortunately, we rarely have this complete information!
- Design a classifier from a training data set
  - No problem with prior estimation
  - Samples are often too small for class-conditional estimation (large dimension of feature space!)
- If we know the number of parameters in advance, our general knowledge about the problem can by reduced by parameterizing the conditional density, e.g.,
  - assuming that $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$
  - estimating an unknown function $p(\mathbf{x}|\omega_i) \Rightarrow$

- Data availability in a Bayesian framework
  - We could design an optimal classifier if we knew:
    - $P(\omega_i)$ (priors)
    - $P(x|\omega_i)$ (class-conditional densities)

  Unfortunately, we rarely have this complete information!
- Design a classifier from a training data set
  - No problem with prior estimation
  - Samples are often too small for class-conditional estimation (large dimension of feature space!)
- If we know the number of parameters in advance, our general knowledge about the problem can by reduced by parameterizing the conditional density, e.g.,
  - assuming that $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$
  - estimating an unknown function $p(\mathbf{x}|\omega_i) \Rightarrow$ estimating the parameters $\boldsymbol{\mu}_i$ and Σ_i

Ways of Parameter Estimation

- The problem of parameter estimation is a classical one in statistics and it can be approached in several ways
- Two common and reasonable procedures:

Ways of Parameter Estimation

- The problem of parameter estimation is a classical one in statistics and it can be approached in several ways
- Two common and reasonable procedures:
  - Maximum-Likelihood (ML) estimation

Ways of Parameter Estimation

- The problem of parameter estimation is a classical one in statistics and it can be approached in several ways
- Two common and reasonable procedures:
  - Maximum-Likelihood (ML) estimation
    - Parameters in ML estimation are fixed but unknown

Ways of Parameter Estimation

- The problem of parameter estimation is a classical one in statistics and it can be approached in several ways
- Two common and reasonable procedures:
  - Maximum-Likelihood (ML) estimation
    - Parameters in ML estimation are fixed but unknown
    - Best parameters are obtained by maximizing the probability of obtaining the samples available

Ways of Parameter Estimation

- The problem of parameter estimation is a classical one in statistics and it can be approached in several ways
- Two common and reasonable procedures:
  - Maximum-Likelihood (ML) estimation
    - Parameters in ML estimation are fixed but unknown
    - Best parameters are obtained by maximizing the probability of obtaining the samples available
  - Bayesian learning

Ways of Parameter Estimation

- The problem of parameter estimation is a classical one in statistics and it can be approached in several ways
- Two common and reasonable procedures:
  - Maximum-Likelihood (ML) estimation
    - Parameters in ML estimation are fixed but unknown
    - Best parameters are obtained by maximizing the probability of obtaining the samples available
  - Bayesian learning
    - Parameters are random variables having some known distribution
    - Training Data convert the known distribution to a posterior density $P(\theta|\mathbf{x}, y)$ and sharpen $P(\hat{\theta}|\mathbf{x}, y)$
- In either approach, we use posterior density for our classification rule

Supervised and Unsupervised Learning

Common properties

- Samples $\mathbf{x}$ are assumed to be obtained by

Supervised and Unsupervised Learning

Common properties

- Samples $\mathbf{x}$ are assumed to be obtained by
  - selecting a state of nature ω_i with probability $P(\omega_i)$

Supervised and Unsupervised Learning

Common properties

- Samples **x** are assumed to be obtained by
  - selecting a state of nature ω_i with probability $P(\omega_i)$
  - independently selecting **x** according to the probability law $p(\mathbf{x}|\omega_i)$

Supervised and Unsupervised Learning

Common properties

- Samples $\mathbf{x}$ are assumed to be obtained by
  - selecting a state of nature ω_i with probability $P(\omega_i)$
  - independently selecting $\mathbf{x}$ according to the probability law $p(\mathbf{x}|\omega_i)$

Distinction

- For supervised learning, we know the state of nature (class label) for each training sample
- For unsupervised learning, we donot know the state of nature (class label) for all training samples

1. Introduction

2. Maximum-Likelihood Estimation
   - General Principle
   - Gaussian Case: Unknown Mean μ
   - Gaussian Case: Unknown μ and Σ
   - Bias

3. Bayesian Estimation
   - Parameter Distribution
   - Gaussian Case

4. General Theory
   - MLE vs. Bayes estimates

Advantages

- Has good convergence properties as the sample size increases
- Simpler than any other alternative techniques

1. Introduction

2. Maximum-Likelihood Estimation
   - General Principle
   - Gaussian Case: Unknown Mean μ
   - Gaussian Case: Unknown μ and Σ
   - Bias

3. Bayesian Estimation
   - Parameter Distribution
   - Gaussian Case

4. General Theory
   - MLE vs. Bayes estimates

Setting

- Suppose that we separate a collection of samples $\mathcal{D}$ according to class, so that we have c datasets, $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_c$
- The samples in $\mathcal{D}_i$ have been drawn independently according to the probability law $p(\mathbf{x}|\omega_i)$

Setting

- Suppose that we separate a collection of samples $\mathcal{D}$ according to class, so that we have c datasets, $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_c$
- The samples in $\mathcal{D}_i$ have been drawn independently according to the probability law $p(\mathbf{x}|\omega_i)$
- We say such samples are i.i.d. - independent and identically distributed random variables

Parameter Vector $\boldsymbol{\theta}_i$ for $p(\mathbf{x}|\omega_i)$

- Assume that $p(\mathbf{x}|\omega_i)$ has a known parametric form, determined uniquely by the value of a parameter vector $\boldsymbol{\theta}_i$
- E.g., $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$, where

$$\boldsymbol{\theta}_i =$$

Parameter Vector $\boldsymbol{\theta}_i$ for $p(\mathbf{x}|\omega_i)$

- Assume that $p(\mathbf{x}|\omega_i)$ has a known parametric form, determined uniquely by the value of a parameter vector $\boldsymbol{\theta}_i$
- E.g., $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$, where

$$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$$

Parameter Vector $\boldsymbol{\theta}_i$ for $p(\mathbf{x}|\omega_i)$

- Assume that $p(\mathbf{x}|\omega_i)$ has a known parametric form, determined uniquely by the value of a parameter vector $\boldsymbol{\theta}_i$
- E.g., $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$, where

$$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$$

- To show the dependence of $p(\mathbf{x}|\omega_i)$ on $\boldsymbol{\theta}_i$

$$p(\mathbf{x}|\omega_i) \propto$$

Parameter Vector $\boldsymbol{\theta}_i$ for $p(\mathbf{x}|\omega_i)$

- Assume that $p(\mathbf{x}|\omega_i)$ has a known parametric form, determined uniquely by the value of a parameter vector $\boldsymbol{\theta}_i$
- E.g., $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$, where

$$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$$

- To show the dependence of $p(\mathbf{x}|\omega_i)$ on $\boldsymbol{\theta}_i$

$$p(\mathbf{x}|\omega_i) \propto p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$$

- The target of ML estimation is to

Parameter Vector $\boldsymbol{\theta}_i$ for $p(\mathbf{x}|\omega_i)$

- Assume that $p(\mathbf{x}|\omega_i)$ has a known parametric form, determined uniquely by the value of a parameter vector $\boldsymbol{\theta}_i$
- E.g., $p(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$, where

$$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$$

- To show the dependence of $p(\mathbf{x}|\omega_i)$ on $\boldsymbol{\theta}_i$

$$p(\mathbf{x}|\omega_i) \propto p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$$

- The target of ML estimation is to use the information provided by the training samples to obtain good estimates for the unknown parameter vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_c$ for each category

Data Independent Assumption

- To simplify treatment of the problem, we shall assume that if $i \neq j$, the samples in $\mathcal{D}_i$ give no information about $\boldsymbol{\theta}_j$

Data Independent Assumption

- To simplify treatment of the problem, we shall assume that if $i \neq j$, the samples in $\mathcal{D}_i$ give no information about $\boldsymbol{\theta}_j$

  $$\Downarrow$$

  the parameters for the different classes are functionally independent

Data Independent Assumption

- To simplify treatment of the problem, we shall assume that if $i \neq j$, the samples in $\mathcal{D}_i$ give no information about $\boldsymbol{\theta}_j$

$$\Downarrow$$

  the parameters for the different classes are functionally independent
- We can work with each category separately and simplify our notation by deleting indications of class distinctions

Use a set $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_c\}$ of training samples drawn independently from the probability density $p(\mathbf{x}|\boldsymbol{\theta})$ to estimate the unknown parameter vector $\boldsymbol{\theta}$

Likelihood of Parameter Vector $\boldsymbol{\theta}$

Suppose $\mathcal{D} = (\mathbf{x}_i)_{i=1}^n$

Samples drawn independently

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

which is called the likelihood of $\boldsymbol{\theta}$ with respect to the set of samples, if viewed as a function of $\boldsymbol{\theta}$

- The maximum-likelihood estimate of $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}$ that maximizes $p(\mathcal{D}|\boldsymbol{\theta})$

Likelihood of Parameter Vector $\boldsymbol{\theta}$

Suppose $\mathcal{D} = (\mathbf{x}_i)_{i=1}^{n}$

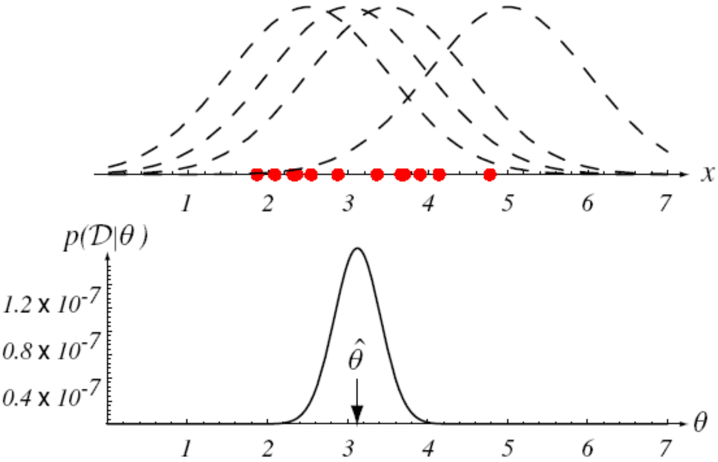
Samples drawn independently

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

which is called the likelihood of $\boldsymbol{\theta}$ with respect to the set of samples, if viewed as a function of $\boldsymbol{\theta}$

- The maximum-likelihood estimate of $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}$ that maximizes $p(\mathcal{D}|\boldsymbol{\theta})$
- It is the value of $\boldsymbol{\theta}$ that agrees with or supports the actually observed training samples

$p(\mathcal{D}|\hat{\theta}), \theta = \mu$



Optimal Estimation (1)

- Let $\boldsymbol{\theta} = [\theta_1 \ ... \ \theta_d]^\top$ and let $\nabla_{\boldsymbol{\theta}}$ be the gradient operator

$$\nabla_{\boldsymbol{\theta}} =$$

Optimal Estimation (1)

- Let $\boldsymbol{\theta} = [\theta_1 \ \ldots \ \theta_d]^\top$ and let $\nabla_{\boldsymbol{\theta}}$ be the gradient operator

$$\nabla_{\boldsymbol{\theta}} = \left[\frac{\partial}{\partial\theta_1} \ldots \frac{\partial}{\partial\theta_d}\right]^\top$$

- Define $l(\boldsymbol{\theta})$ as the log-likelihood function

$$l(\boldsymbol{\theta}) \equiv \ln p(\mathcal{D}|\boldsymbol{\theta})$$

- New problem statement: to determine the argument $\boldsymbol{\theta}$ that maximizes the log-likelihood

$$\hat{\boldsymbol{\theta}} =$$

Optimal Estimation (1)

- Let $\boldsymbol{\theta} = [\theta_1 \ ... \ \theta_d]^\top$ and let $\nabla_{\boldsymbol{\theta}}$ be the gradient operator

$$\nabla_{\boldsymbol{\theta}} = \left[\frac{\partial}{\partial\theta_1} \cdots \frac{\partial}{\partial\theta_d}\right]^\top$$

- Define $l(\boldsymbol{\theta})$ as the log-likelihood function

$$l(\boldsymbol{\theta}) \equiv \ln p(\mathcal{D}|\boldsymbol{\theta})$$

- New problem statement: to determine the argument $\boldsymbol{\theta}$ that maximizes the log-likelihood

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

Optimal Estimation (2)

- The dependence on the dataset $\mathcal{D}$ is implicit
- since $l(\boldsymbol{\theta}) \equiv \ln p(\mathcal{D}|\boldsymbol{\theta})$,

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\boldsymbol{\theta})$$

and

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_i|\boldsymbol{\theta})$$

- A set of necessary conditions for the maximum likelihood estimate for $\boldsymbol{\theta}$ can be obtained from the set of d equations

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = 0$$

Optimal Estimation (3)

- Representing a true global maximum, or a local maximum

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = 0$$

- second derivatives to check the global minimum
-

$$\lim_{n \to \infty} \hat{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}$$

MAP

- $p(\theta)$, the prior probability of different parameter values
- maximum a posteriori (MAP)

$$\max_{\theta} l(\theta)p(\theta)$$

- MAP = MLE + uniform or “flat" prior
- drawback of MAP

Sample mean

- Samples are drawn from a multivariate normal population, i.e., $p(\mathbf{x}_i|\omega) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$
- For simplicity, consider only the mean is unknown, i.e., $\boldsymbol{\theta} =$

Sample mean

- Samples are drawn from a multivariate normal population, i.e., $p(\mathbf{x}_i|\omega) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$
- For simplicity, consider only the mean is unknown, i.e., $\boldsymbol{\theta} = \boldsymbol{\mu}$

$$
\begin{aligned}
\ln p(\mathbf{x}_i|\boldsymbol{\theta}) &= \ln p(\mathbf{x}_i|\boldsymbol{\mu}) \\
&= -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) - \frac{1}{2}\ln[(2\pi)^d|\Sigma|]
\end{aligned}
$$

and

$$
\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_i|\boldsymbol{\theta}) = \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})
$$

- The ML estimate for $\boldsymbol{\mu}$ must satisfy

$$
\sum_{i=1}^{n} \Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) = 0 \Rightarrow
$$

Sample mean

- Samples are drawn from a multivariate normal population, i.e., $p(\mathbf{x}_i|\omega) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$
- For simplicity, consider only the mean is unknown, i.e., $\boldsymbol{\theta} = \boldsymbol{\mu}$

$$\begin{aligned} \ln p(\mathbf{x}_i|\boldsymbol{\theta}) &= \ln p(\mathbf{x}_i|\boldsymbol{\mu}) \\ &= -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) - \frac{1}{2}\ln[(2\pi)^d|\Sigma|] \end{aligned}$$

and

$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_i|\boldsymbol{\theta}) = \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

- The ML estimate for $\boldsymbol{\mu}$ must satisfy

$$\sum_{i=1}^{n} \Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) = 0 \Rightarrow \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

Univariate Gaussian

- $\boldsymbol{\theta} = (\theta_1,\ \theta_2)\ =\ (\mu,\ \sigma^2)$

Univariate Gaussian

- $\boldsymbol{\theta} = (\theta_1,\ \theta_2)\ =\ (\mu,\ \sigma^2)$
- The log-likelihood of a single point is

$$l(\boldsymbol{\theta}) = \ln p(x_i|\boldsymbol{\theta}) \quad = \quad -\frac{1}{2\theta_2}(x_i - \theta_1)^2 - \frac{1}{2}\ln 2\pi\theta_2$$

and its derivative is

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

Univariate Gaussian

- $\boldsymbol{\theta} = (\theta_1,\ \theta_2)\ =\ (\mu,\ \sigma^2)$
- The log-likelihood of a single point is

$$l(\boldsymbol{\theta}) = \ln p(x_i|\boldsymbol{\theta}) \quad = \quad -\frac{1}{2\theta_2}(x_i - \theta_1)^2 - \frac{1}{2}\ln 2\pi\theta_2$$

and its derivative is

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \ = \ \begin{bmatrix} \nabla_{\theta_1} l(\boldsymbol{\theta}) \\ \nabla_{\theta_2} l(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \frac{1}{\theta_2}(x_i - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_i-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

to the full log-likelihood

Univariate Gaussian

- $\boldsymbol{\theta} = (\theta_1,\ \theta_2)\ =\ (\mu,\ \sigma^2)$
- The log-likelihood of a single point is

$$l(\boldsymbol{\theta}) = \ln p(x_i|\boldsymbol{\theta}) \quad = \quad -\frac{1}{2\theta_2}(x_i - \theta_1)^2 - \frac{1}{2}\ln 2\pi\theta_2$$

and its derivative is

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \ = \ \begin{bmatrix} \nabla_{\theta_1} l(\boldsymbol{\theta}) \\ \nabla_{\theta_2} l(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \frac{1}{\theta_2}(x_i - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_i-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

to the full log-likelihood

$$\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(x_i|\boldsymbol{\theta}) \ = \ 0 \ \Rightarrow \begin{cases} \sum_{i=1}^{n} \frac{1}{\hat{\theta}_2}(x_i - \hat{\theta}_1) = 0 \\ \sum_{i=1}^{n} -\frac{1}{2\hat{\theta}_2} + \frac{(x_i-\hat{\theta}_1)^2}{2\hat{\theta}_2^2} = 0 \end{cases}$$

Univariate normal case

- we can obtain the following maximum-likelihood estimates for μ and σ^2

$$\begin{cases} \hat{\mu} = \hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \end{cases}$$

Multivariate normal case

- With similar analysis, we can obtain the following maximum-likelihood estimates for μ and Σ^2

$$\begin{cases} \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \\ \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \end{cases}$$

1. Introduction
2. Maximum-Likelihood Estimation
   - General Principle
   - Gaussian Case: Unknown Mean μ
   - Gaussian Case: Unknown μ and Σ
   - Bias
3. Bayesian Estimation
   - Parameter Distribution
   - Gaussian Case
4. General Theory
   - MLE vs. Bayes estimates

Mean

- Bias of an estimator $\hat{\theta}$:

$$\mathcal{E}_{\boldsymbol{\theta}}\left[(f(\hat{\boldsymbol{\theta}};\mathcal{D})-f(\boldsymbol{\theta};\mathcal{D}))^2\right]$$

- ML estimate $\hat{\mu}$ of the mean $\boldsymbol{\mu}$ is unbiased

$$\mathcal{E}[\hat{\boldsymbol{\mu}}]=\mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]=\boldsymbol{\mu}$$

Variance

- ML estimate for the variance σ^2 is biased

$$\mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\right]=\frac{n-1}{n}\sigma^2\neq\sigma^2$$

  which is asymptotically unbiased

- An elementary unbiased estimator for $\hat{\Sigma}$ is given by

$$\underbrace{\mathbf{C}=\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top}_{\text{Sample covariance matrix}}$$

  which is absolutely unbiased

Class-Conditional Densities

- In MLE, $\boldsymbol{\theta}$ was supposed fixed
- In BE, $\boldsymbol{\theta}$ is a random variable
- The computation of posterior probabilities $P(\omega_i|\mathbf{x})$ lies at the heart of Bayesian classification
- if prior probabilities and class densities are unknown, how to compute a posteriori?
- Goal: Given the sample set $\mathcal{D}$, compute $P(\omega_i|\mathbf{x}, \mathcal{D})$
- Bayes formula can be written

$$P(\omega_i|\mathbf{x}, \mathcal{D}) =$$

Class-Conditional Densities

- In MLE, $\boldsymbol{\theta}$ was supposed fixed
- In BE, $\boldsymbol{\theta}$ is a random variable
- The computation of posterior probabilities $P(\omega_i|\mathbf{x})$ lies at the heart of Bayesian classification
- if prior probabilities and class densities are unknown, how to compute a posteriori?
- Goal: Given the sample set $\mathcal{D}$, compute $P(\omega_i|\mathbf{x},\ \mathcal{D})$
- Bayes formula can be written

$$
\begin{aligned}
P(\omega_i|\mathbf{x},\ \mathcal{D}) &= \frac{p(\mathbf{x}|\omega_i, \mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \mathcal{D})P(\omega_j|\mathcal{D})} \\
&\propto
\end{aligned}
$$

Class-Conditional Densities

- In MLE, $\boldsymbol{\theta}$ was supposed fixed
- In BE, $\boldsymbol{\theta}$ is a random variable
- The computation of posterior probabilities $P(\omega_i|\mathbf{x})$ lies at the heart of Bayesian classification
- if prior probabilities and class densities are unknown, how to compute a posteriori?
- Goal: Given the sample set $\mathcal{D}$, compute $P(\omega_i|\mathbf{x},\ \mathcal{D})$
- Bayes formula can be written

$$
\begin{aligned}
P(\omega_i|\mathbf{x},\ \mathcal{D}) &= \frac{p(\mathbf{x}|\omega_i,\mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j,\mathcal{D})P(\omega_j|\mathcal{D})} \\
&\propto \frac{p(\mathbf{x}|\omega_i,\mathcal{D}_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j,\mathcal{D}_j)P(\omega_j)}
\end{aligned}
$$

Central problem of Bayesian learning

- Desired probability density $p(\mathbf{x})$ is unknown, but has a known parametric form $p(\mathbf{x}|\boldsymbol{\theta})$
- The only thing assumed unknown is the value of a parameter vector $\boldsymbol{\theta}$

- known prior density $p(\boldsymbol{\theta})$
- $\mathcal{D}$ converts the prior to a posterior density $p(\boldsymbol{\theta}|\mathcal{D})$

Learning a probability density function $\Rightarrow$ estimating a parameter vector

$p(\mathbf{x})$,

$p(\mathbf{x})$, $p(\mathbf{x}|\mathcal{D})$,

$p(\mathbf{x})$, $p(\mathbf{x}|\mathcal{D})$, $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$

$p(\mathbf{x}), p(\mathbf{x}|\mathcal{D}), p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$

$p(\mathbf{x})$, $p(\mathbf{x}|\mathcal{D})$, $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$

$$
\begin{aligned}
p(\mathbf{x}|\mathcal{D}) &= \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\
&= \int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}
\end{aligned}
$$

$p(\mathbf{x})$, $p(\mathbf{x}|\mathcal{D})$, $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$

$$
\begin{aligned}
p(\mathbf{x}|\mathcal{D}) &= \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\
&= \int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\
&= \int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}
\end{aligned}
$$

$p(\mathbf{x})$, $p(\mathbf{x}|\mathcal{D})$, $p(\mathbf{x},\boldsymbol{\theta}|\mathcal{D})$

$$\begin{aligned} p(\mathbf{x}|\mathcal{D}) &= \int p(\mathbf{x},\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \\ &= \int p(\mathbf{x}|\boldsymbol{\theta},\mathcal{D})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \\ &= \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \end{aligned}$$

- which links the desired class-conditional density $p(\mathbf{x}|\mathcal{D})$ to the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$ for the unknown parameter
- which can be performed numerically, e.g., by Monte-Carlo simulation
- if $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply about some value $\hat{\theta}$, $p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\theta})$

Use the BE to calculate the *a posteriori* density $p(\boldsymbol{\theta}|\mathcal{D})$ and the desired probability density $p(\mathbf{x}|\mathcal{D})$ for the case where $p(\mathbf{x}|\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Univariate Case: $p(\mu|\mathcal{D})$ - Prior $p(\mu)$

Consider the case where μ is the only unknown parameter. For the simplicity, we treat first the univariate case, i.e.,

$$p(x|\mu) \sim \mathcal{N}(\mu, \sigma^2)$$

where the only unknown quantity is the mean μ

- Assume that prior knowledge we might have about μ can be expressed by a known prior density $p(\mu)$
- Assume that

$$p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

where μ_0 and σ_0^2 are known

Univariate Case: $p(\mu|\mathcal{D})$

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)d\mu}$$

Univariate Case: $p(\mu|\mathcal{D})$

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)d\mu} \\
&= \alpha \prod_{i=1}^{n} p(\mathbf{x}_i|\mu)p(\mu)
\end{aligned}
$$

where α is a normalization factor that depends on $\mathcal{D}$ but is independent of μ

$p(x|\mu) \sim \mathcal{N}(\mu, \sigma^2)$ and $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$

$$
p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]
$$

Bayesian Learning

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

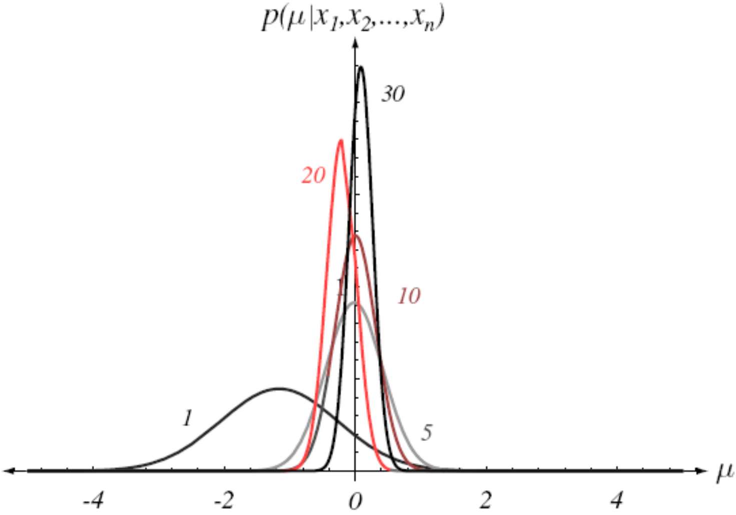
$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n\right) + \left(\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0\right)$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

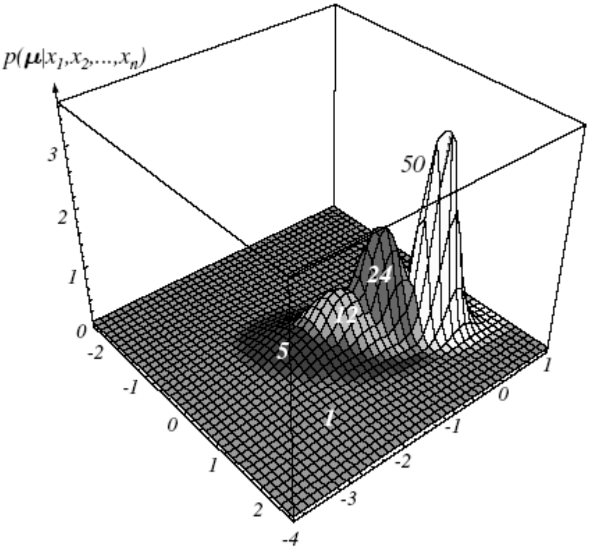
- μ_n represents our best guess for μ after observing n samples
- σ_n^2 measures our uncertainty about this guess
- As n increases, $p(\mu|\mathcal{D})$ becomes more and more sharply peaked, approaching a Dirac delta function as n approaches infinity

This is known as Bayesian learning

Bayesian Learning: Visualization in 1-D



Bayesian Learning: Visualization in 2-D



Univariate Case: $p(x|\mathcal{D})$

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu)d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]f(\sigma,\sigma_n)
\end{aligned}
$$

Finally, we can obtain

$$p(x|\mathcal{D}) \sim \mathcal{N}(\mu_n, \sigma^2+\sigma_n^2)$$

Therefore: $P(x|\omega_i, \mathcal{D}_i)$ together with $P(\omega_i)$ and using Bayes formula, we obtain the Bayesian classification rule:

$$\max_{\omega_i}\left[P(\omega_i|x,\mathcal{D}_i)\right] = \max_{\omega_i}\left[P(x|\omega_i,\mathcal{D}_i)\cdot P(\omega_i)\right]$$

Univariate Case: $p(x|\mathcal{D})$

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu)d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma, \sigma_n)
\end{aligned}
$$

Finally, we can obtain

$$
p(x|\mathcal{D}) \sim \mathcal{N}(\mu_n, \sigma^2 + \sigma_n^2)
$$

Therefore: $P(x|\omega_i, \mathcal{D}_i)$ together with $P(\omega_i)$ and using Bayes formula, we obtain the Bayesian classification rule:

$$
\max_{\omega_i}\left[P(\omega_i|x, \mathcal{D}_i)\right] = \max_{\omega_i}\left[P(x|\omega_i, \mathcal{D}_i) \cdot P(\omega_i)\right]
$$

Conjugate prior

definition

- for a given probability distribution $p(\mathbf{x}|\boldsymbol{\theta})$, we can seek a prior $p(\boldsymbol{\theta})$ that is conjugate to the likelihood function, so that the posterior distribution has the same functional form as the prior
- e.g., Gaussian case, for unknown $\boldsymbol{\theta} = \boldsymbol{\mu}$

$$
\begin{aligned}
p(x_i|\mu) &\sim \mathcal{N}(\mu, \sigma^2) \\
p(\mu) &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\
p(\mu|\mathcal{D}) &\sim \mathcal{N}(\mu_n, \sigma_n^2)
\end{aligned}
$$

pdf	conjugate prior
Multinomial	Dirichlet
Binomial	Beta

Bernoulli Case

- consider $P(\mathbf{x}|\boldsymbol{\theta}) = \text{Bern}(\mathbf{x}|\boldsymbol{\mu}) = \boldsymbol{\mu}^{\mathbf{x}}(1-\boldsymbol{\mu})^{1-\mathbf{x}}$, where $\boldsymbol{\theta} = \boldsymbol{\mu}$ is the unknown parameter
- a conjugate prior distribution $p(\boldsymbol{\mu}) = \text{Beta}(\alpha, \beta)$ where α, β are both known
- $p(\boldsymbol{\mu}|\mathcal{D}) = \text{Beta}\left(\alpha + \sum_{i=1}^{n} \mathbf{x}_i \beta + n - \sum_{i=1}^{n} \mathbf{x}_i\right)$

- So far, we see how the BE can be used to obtain the desired density $p(\mathbf{x}|\mathcal{D})$ in a special case-uni/multi- variate Gaussian
- This approach can be generalized to apply to any situation in which unknown density can be parameterized

Basic Assumption

- Form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known, but the value of the parameter vector $\boldsymbol{\theta}$ is not known exactly
- Our initial knowledge about $\boldsymbol{\theta}$ is assumed to be contained in a known prior density $p(\boldsymbol{\theta})$
- The rest of our knowledge about $\boldsymbol{\theta}$ is contained in a set $\mathcal{D}$ of n random variables $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ that follows $P(\mathbf{x})$

The basic problem is to compute the posterior density $P(\boldsymbol{\theta}|\mathcal{D})$ to derive $P(\mathbf{x}|\mathcal{D})$

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

Problems

- Computation difficulties
- convergence of $p(\mathbf{x}|\mathcal{D}) \rightarrow p(\mathbf{x})$?

Recursive Bayes learning

Problems

- Provided $\mathcal{D}^n = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, if $n > 1$

$$p(\mathcal{D}^n|\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})p(\mathcal{D}^{n-1}|\boldsymbol{\theta})$$
$$\rightarrow p(\boldsymbol{\theta}|\mathcal{D}^n) = \frac{p(\mathbf{x}_n|\boldsymbol{\theta} p(\boldsymbol{\theta}|\mathcal{D}^{n-1})}{\int p(\mathbf{x}_n|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}^{n-1})d\boldsymbol{\theta}}$$

- $p(\boldsymbol{\theta}|\mathcal{D}^0) = p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{x}_1)$, $p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2), \cdots$
- To estimate $p(\boldsymbol{\theta}|\mathcal{D}^n)$, all the training data in $\mathcal{D}^{n-1}$ should be kept
- sufficient statistics, where distributions can be represented using only a few parameters

Recursive Bayes learning

Problems

- Provided $\mathcal{D}^n = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, if $n > 1$

$$
\begin{aligned}
p(\mathcal{D}^n|\boldsymbol{\theta}) &= p(\mathbf{x}_n|\boldsymbol{\theta})p(\mathcal{D}^{n-1}|\boldsymbol{\theta}) \\
\rightarrow p(\boldsymbol{\theta}|\mathcal{D}^n) &= \frac{p(\mathbf{x}_n|\boldsymbol{\theta} p(\boldsymbol{\theta}|\mathcal{D}^{n-1})}{\int p(\mathbf{x}_n|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}^{n-1})d\boldsymbol{\theta}}
\end{aligned}
$$

- $p(\boldsymbol{\theta}|\mathcal{D}^0) = p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{x}_1)$, $p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2), \cdots$
- To estimate $p(\boldsymbol{\theta}|\mathcal{D}^n)$, all the training data in $\mathcal{D}^{n-1}$ should be kept
- sufficient statistics, where distributions can be represented using only a few parameters

Comparison

Different behaviors

	MLE	Bayes
computation	differential, gradient	multidimensional integration
Interpretability	single best model $\hat{\theta}$	weighted average of models
prior	$p(\mathbf{x}\|\hat{\theta}) = p(\mathbf{x}\|\theta)$	$p(\mathbf{x}\|\mathcal{D}) \neq p(\mathbf{x}\|\hat{\theta})$

Similar behavior

- asymptotic limit of infinite training data
- strongly peaked $p(\mathbf{x}|\hat{\theta})$ and the prior $p(\theta)$ is uniform of flat

Classification error

Three sources

- Bayes error: overlapping densities for different category $p(\mathbf{x}|\omega_i)$, intrinsic and cannot be eliminated
- Model error: domain knowledge dependent
- Estimation error: parameters estimated from a finite set of samples, error can be reduced by increasing the number of training data