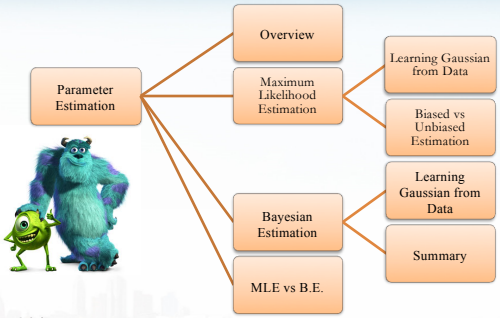
Lecture Notes on

Pattern Recognition

Session 03(A): Parameter Estimation (I)

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

Road map



Parameter Estimation

- Maximum Likelihood
- Bayesian Estimation

Parameter Estimation

- **Bayesian Decision Theory** allows us to design an optimal classifier if we know the prior probabilities $P(\omega_i)$ and the class-conditional densities $p(x/\omega_i)$.

$$P(\omega_j / x) = \frac{p(x / \omega_j)P(\omega_j)}{p(x)}$$

- Estimate $P(\omega_i)$ and $p(x/\omega_i)$ from training examples.
  - Estimating $P(\omega_i)$ is usually not difficult.
  - Estimating $p(x/\omega_i)$ is more difficult!
    - Number of samples is often too small
    - Dimensionality of feature space is large.

Parameter Estimation

- Assumptions
  - We are given a set of training samples $D = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, where the samples were drawn according to $p(\mathbf{x} \mid w_j)$
  - $p(\mathbf{x} \mid w_j)$ has known parametric form , e.g.,

$$p(\mathbf{x}/\omega_j) \sim N(\boldsymbol{\mu}_i, \Sigma_j), \quad \boldsymbol{\theta}=(\mu_i, \Sigma_j)$$

- Parameter estimation problem:

  Given *D*, find the best possible θ

Main Methods in Parameter Estimation

- **Maximum Likelihood (ML)**
  - Assumes that the values of the parameters are ***fixed*** but unknown.
  - Best estimate $\hat{\boldsymbol{\theta}}$ is obtained by ***maximizing*** the probability of obtaining the samples actually observed (i.e., training data)

$$p(D/\boldsymbol{\theta})$$

Main Methods in Parameter Estimation

- **Bayesian Estimation (BE)**
  - Assumes that the parameters $\boldsymbol{\theta}$ are ***random*** variables that have some known a-priori distribution $p(\boldsymbol{\theta})$.
  - Estimates a ***distribution*** rather than making point estimates like ML.
  - BE solution ***might not*** be of the parametric form assumed.

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

ML Estimation

- A General Strategy
- Learning Gaussian from Data
- Biased and Unbiased Estimation

Why we should care

- **Maximum Likelihood Estimation** is a very very very very fundamental part of data analysis.
- "*MLE for Gaussians*" is training wheels for our future techniques
  - Learning Gaussians is more useful than you might guess…

ML Estimation (Assumptions)

- Training data is divided in ***c*** sets, based on ***c*** classes.

$$D_1, D_2, ..., D_c$$

- Samples in D_j have been drawn independently according to $p(x/\omega_j)$.
- $p(x/\omega_j)$ has known parametric form with parameters $\boldsymbol{\theta}_j$:
  - e.g., $\boldsymbol{\theta}_j = (\mu_j, \Sigma_j)$ for Gaussian distributions

ML Estimation (Problem Definition)

- Problem: given $D_1, D_2, ..., D_c$ and a model for each class, estimate $\boldsymbol{\theta_1, \theta_2, \ldots, \theta_c}$
  - If the samples in D_j give no information about $\boldsymbol{\theta}_i$, where $i \neq j$. We need to solve c independent problems
    - i.e., one for each class
  - ML estimate for $D=\{x_1, x_2, .., x_n\}$ is the $\hat{\boldsymbol{\theta}}$ value that maximizes $p(D/\boldsymbol{\theta})$
    - i.e., best supports the training data

$$p(D/\boldsymbol{\theta}) = p(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n/\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k/\boldsymbol{\theta})$$

MLE Gaussians from Data (Case of Unknown $\theta = \mu$)

- Suppose you have $x_1, x_2, \ldots x_n \sim$ (i.i.d) $N(\mu, \sigma^2)$
- You don't know μ, but you do know σ^2
- MLE:
  - For which μ is $x_1, x_2, \ldots x_n$ most likely?

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, ...x_n \mid \mu, \sigma^2)$$

MLE Gaussians from Data (Case of Unknown θ = μ)

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_n \mid \mu, \sigma^2)$$

$=$ (by i.i.d)

$=$ (monotonicity of log)

$=$ (plug in formula for Gaussian)

$=$ (after simplification)

MLE Gaussians from Data (Case of Unknown θ = μ)

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_n \mid \mu, \sigma^2)$$

$$= \arg\max_{\mu} \prod_{i=1}^{R} p(x_i \mid \mu, \sigma^2)$$ (by i.i.d)

$$= \arg\max_{\mu} \sum_{i=1}^{R} \log p(x_i \mid \mu, \sigma^2)$$ (monotonicity of log)

$$= \arg\max_{\mu} \frac{1}{\sqrt{2\pi}\,\sigma} \sum_{i=1}^{R} -\frac{(x_i - \mu)^2}{2\sigma^2}$$ (plug in formula for Gaussian)

$$= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$ (after simplification)

MLE Gaussians from Data (Case of Unknown θ = μ)

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_n \mid \mu, \sigma^2)$$

$$= \arg\min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$subject\ to\ 0 = \frac{\partial LL}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{n} (x_i - \mu)^2 = -\sum_{i=1}^{n} 2(x_i - \mu)$$

$$\text{Thus}\quad \mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

MLE Gaussians from Data (Case of Unknown θ = μ)

- The best estimate of the mean of a distribution is the mean of the sample!

$$\mu^{mle} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

At first sight:
This kind of pedantic, algebra-filled and ultimately unsurprising fact is exactly the reason people throw down their "Statistics" book and pick up their "Agent Based Evolutionary Data Mining Using The Neuro-Fuzz Transform" book.

A General Scalar MLE strategy

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus
3. Set ∂LL/∂θ=0 for a maximum, creating an equation in terms of θ
4. Solve it*
5. Check that you have found a maximum rather than a minimum or saddle-point, and be careful if θ is constrained

**This is a perfect example of something that works perfectly in all textbook examples and usually involves surprising pain if you need it for something new.*

A General MLE strategy

- Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T$ is a vector of parameters. Find MLE $\boldsymbol{\theta}$ assuming known form for p(Data|$\boldsymbol{\theta}$,stuff)
  1. Write LL = log P(Data|$\boldsymbol{\theta}$,stuff)
  2. Work out $\partial LL/\partial \boldsymbol{\theta}$ using high-school calculus

$$\frac{\partial LL}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial LL}{\partial \theta_1} \\ \frac{\partial LL}{\partial \theta_2} \\ \vdots \\ \frac{\partial LL}{\partial \theta_n} \end{pmatrix}$$

A General MLE strategy

- Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T$ is a vector of parameters. Find MLE $\boldsymbol{\theta}$ assuming known form for p(Data | $\boldsymbol{\theta}$,stuff)
  1. Write LL = log P(Data | $\boldsymbol{\theta}$,stuff)
  2. Work out $\partial LL/\partial \boldsymbol{\theta}$ using high-school calculus
  3. Solve the set of simultaneous equations

$$\frac{\partial LL}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial LL}{\partial \theta_1} \\ \frac{\partial LL}{\partial \theta_2} \\ \vdots \\ \frac{\partial LL}{\partial \theta_n} \end{pmatrix} \quad \begin{matrix} \frac{\partial LL}{\partial \theta_1} = 0 \\ \frac{\partial LL}{\partial \theta_2} = 0 \\ \vdots \\ \frac{\partial LL}{\partial \theta_n} = 0 \end{matrix}$$

A General MLE strategy

- Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T$ is a vector of parameters. Find MLE $\boldsymbol{\theta}$ assuming known form for p(Data | $\boldsymbol{\theta}$,stuff)
  1. Write LL = log P(Data | $\boldsymbol{\theta}$,stuff)
  2. Work out $\partial LL/\partial \boldsymbol{\theta}$ using high-school calculus
  3. Solve the set of simultaneous equations
  4. Check that you are at a maximum

MLE Gaussians from Data (Case of Unknown $\theta=(\theta_1,\theta_2)=(\mu,\sigma^2)$)

- Suppose you have $x_1, x_2, \ldots x_n \sim$ (i.i.d) $N(\mu,\sigma^2)$
- You don't know μ or σ^2
- MLE:
  - For which $\boldsymbol{\theta} = (\mu,\sigma^2)$ is $x_1, x_2, \ldots x_n$ most likely?

$$\log p(x_1, x_2, \ldots x_n \mid \mu, \sigma^2) = -n(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial LL}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \qquad \frac{\partial LL}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

MLE Gaussians from Data (Case of Unknown $\theta=(\theta_1,\theta_2)=(\mu,\sigma^2)$)

- Suppose you have $x_1, x_2, \ldots x_n \sim$ (i.i.d) $N(\mu,\sigma^2)$
- You don't know μ or σ^2
- MLE:
  - For which $\boldsymbol{\theta} = (\mu,\sigma^2)$ is $x_1, x_2, \ldots x_n$ most likely?

$$\log p(x_1, x_2, \ldots x_n \mid \mu, \sigma^2) = -n(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \qquad 0 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

MLE Gaussians from Data (Case of Unknown $\theta=(\theta_1,\theta_2)=(\mu,\sigma^2)$)

- Suppose you have $x_1, x_2, \ldots x_n \sim$ (i.i.d) $N(\mu,\sigma^2)$
- You don't know μ or σ^2
- MLE:
  - For which $\boldsymbol{\theta} = (\mu,\sigma^2)$ is $x_1, x_2, \ldots x_n$ most likely?

$$\log p(x_1, x_2, \ldots x_n \mid \mu, \sigma^2) = -n(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\mu^{mle} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2_{mle} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu^{mle})^2$$

Unbiased Estimators

- An estimator of a parameter is ***unbiased*** if the expected value of the estimate is the ***same*** as the true value of the parameters.
- *If* $x_1, x_2, \ldots x_R \sim$(i.i.d) $N(\mu,\sigma^2)$ then

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

- μ^{mle} is unbiased

Biased Estimators

- An estimator of a parameter is ***biased*** if the expected value of the estimate is the ***different from*** the true value of the parameters.
- *If* $x_1, x_2, \ldots x_R \sim$(i.i.d) $N(\mu,\sigma^2)$ then

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

- σ^2_{mle} is biased

MLE Variance Bias

- *If* $x_1, x_2, \ldots x_R \sim$(i.i.d) $N(\mu,\sigma^2)$ then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{n}\left(\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2\right] = \left(1-\frac{1}{n}\right)\sigma^2 \neq \sigma^2$$

Intuition check: consider the case of n=1

Why should our guts expect that σ^2_{mle} would be an underestimate of true σ^2?

How could you prove that?

- σ^2_{mle} is biased

MLE Variance Bias

- *If* $x_1, x_2, \ldots x_R \sim$(i.i.d) $N(\mu,\sigma^2)$ then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{n}\left(\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2\right] = \left(1-\frac{1}{n}\right)\sigma^2 \neq \sigma^2$$

$$\sigma^2_{\text{unbiased}} = \frac{\sigma^2_{mle}}{\left(1-\frac{1}{n}\right)} \qquad \text{So } E\left[\sigma^2_{\text{unbiased}}\right] = \sigma^2$$

$$\sigma^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu^{mle})^2$$

Which one is better?

- It depends on the task
- no much difference when R→ large

MLE Gaussians from Data (m-Dimensional $\theta=(\theta_1,\theta_2)=(\mu, \Sigma)$)

- Suppose you have $x_1, x_2, \ldots x_n \sim$ (i.i.d) $N(\mu, \boldsymbol{\Sigma})$
- You don't know μ or $\boldsymbol{\Sigma}$
- MLE:
  - For which $\boldsymbol{\theta} = (\mu, \boldsymbol{\Sigma})$ is $x_1, x_2, \ldots x_n$ most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

MLE Gaussians from Data (m-Dimensional $\theta=(\theta_1,\theta_2)=(\mu, \Sigma)$)

- Suppose you have $x_1, x_2, \ldots x_n \sim$ (i.i.d) $N(\mu, \boldsymbol{\Sigma})$
- You don't know μ or $\boldsymbol{\Sigma}$
- MLE:
  - For which $\boldsymbol{\theta} = (\mu, \boldsymbol{\Sigma})$ is $x_1, x_2, \ldots x_n$ most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \qquad \mu_i^{mle} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_{ki}$$

Where $1 \le i \le m$

And x_{ki} is value of the i^{th} component of $\mathbf{x}_k$ (the i^{th} attribute of the k^{th} record)

And μ_i^{mle} is the i^{th} component of $\boldsymbol{\mu}^{mle}$

MLE Gaussians from Data (m-Dimensional $\theta=(\theta_1,\theta_2)=(\mu, \Sigma)$)

- Suppose you have $x_1, x_2, \ldots x_n \sim$ (i.i.d) $N(\mu, \boldsymbol{\Sigma})$
- You don't know μ or $\boldsymbol{\Sigma}$
- MLE:
  - For which $\boldsymbol{\theta} = (\mu, \boldsymbol{\Sigma})$ is $x_1, x_2, \ldots x_n$ most likely?

$$\sigma_{ij}^{mle} = \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_{ki} - \mu_i^{mle}\right)\left(\mathbf{x}_{kj} - \mu_j^{mle}\right)$$

Where $1 \le i \le m$, $1 \le j \le m$

And x_{ki} is value of the i^{th} component of $\mathbf{x}_k$ (the i^{th} attribute of the k^{th} record)

And σ_{ij}^{mle} is the $(i,j)^{th}$ component of $\boldsymbol{\Sigma}^{mle}$

MLE Gaussians from Data (m-Dimensional $\theta=(\theta_1,\theta_2)=(\mu, \Sigma)$)

- Suppose you have $x_1, x_2, \ldots x_n \sim$ (i.i.d) $N(\mu, \boldsymbol{\Sigma})$
- You don't know μ or $\boldsymbol{\Sigma}$
- MLE:
  - For which $\boldsymbol{\theta} = (\mu, \boldsymbol{\Sigma})$ is $x_1, x_2, \ldots x_n$ most likely?

$$\sigma_{ij}^{mle} = \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_{ki} - \mu_i^{mle}\right)\left(\mathbf{x}_{kj} - \mu_j^{mle}\right)$$

$$\boldsymbol{\Sigma}^{mle} = \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)^T$$

Where $1 \le i \le m$, $1 \le j \le m$
And x_{ki} is value of the i^{th} component of $\mathbf{x}_k$ (the i^{th} attribute of the k^{th} record)
And σ_{ij}^{mle} is the $(i,j)^{th}$ component of $\boldsymbol{\Sigma}^{mle}$

MLE Gaussians from Data (m-Dimensional $\theta=(\theta_1,\theta_2)=(\mu, \Sigma)$)

- Suppose you have $x_1, x_2, \ldots x_n \sim$ (i.i.d) $N(\mu, \boldsymbol{\Sigma})$
- You don't know μ or $\boldsymbol{\Sigma}$
- MLE:
  - For which $\boldsymbol{\theta} = (\mu, \boldsymbol{\Sigma})$ is $x_1, x_2, \ldots x_n$ most likely?

$$\boldsymbol{\Sigma}^{\text{unbiased}} = \frac{\boldsymbol{\Sigma}^{mle}}{1-\frac{1}{n}} = \frac{1}{n-1}\sum_{k=1}^{n}\left(\mathbf{x}_k - \mu^{mle}\right)\left(\mathbf{x}_k - \mu^{mle}\right)^T$$

$$\boldsymbol{\Sigma}^{mle} = \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)^T$$

Note how $\boldsymbol{\Sigma}^{mle}$ is forced to be symmetric non-negative definite

How many data points would you need before the Gaussian has a chance of being non-degenerate?

Summary on MLE

- ML estimation is usually simpler than alternative methods.
  - More accurate estimates as the number of training samples increases.
  - If the model chosen for $p(x/\theta)$ is correct, and independence assumptions among variables are true, ML will give very good results.
  - Otherwise, ML will give poor results.

Bayesian Estimation

- A General Strategy
- Learning Gaussian from Data
- Summary

Bayesian Estimation

- Assumes that the parameters $\boldsymbol{\theta}$ are ***random variables*** that have some known a-priori distribution $p(\boldsymbol{\theta})$.
- Using the training examples D, BE converts $p(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|D)$.
  - BE estimates a ***distribution*** rather than making point estimates like ML.

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

- **Note:** BE solution *might not* be of the parametric form assumed.

The Role of Training Examples

- If $p(x|\omega_i)$ and $P(\omega_i)$ are known, Bayes' rule allows us to compute the posterior probabilities $P(\omega_i|x)$:

$$P(\omega_i/\mathbf{x}) = \frac{p(\mathbf{x}/\omega_i)P(\omega_i)}{\sum_j p(\mathbf{x}/\omega_j)P(\omega_j)}$$

  - The role of the ***training examples*** D_i can be emphasized by using them in the computation of the posterior probabilities:

$$P(\omega_i/\mathbf{x}, D_i)$$

The Role of Training Examples

$$P(\omega_i / \mathbf{x}, D_i) = \frac{p(\mathbf{x}, D_i / \omega_i) P(\omega_i)}{p(\mathbf{x}, D_i)} = \frac{p(\mathbf{x} / D_i, \omega_i) p(D_i / \omega_i) P(\omega_i)}{p(\mathbf{x} / D_i) p(D_i)}$$

$$= \frac{p(\mathbf{x} / \omega_i, D_i) P(\omega_i / D_i)}{p(\mathbf{x} / D_i)} = \frac{p(\mathbf{x} / \omega_i, D_i) P(\omega_i / D_i)}{\sum_j p(\mathbf{x}, \omega_j / D_j)}$$

$$= \frac{p(\mathbf{x} / \omega_i, D_i) P(\omega_i / D_i)}{\sum_j p(\mathbf{x} / \omega_j, D_j) P(\omega_j / D_j)}$$

The Role of Training Examples

- This implies that the training examples D_i can help us to determine both the class-conditional densities and the prior probabilities:

$$P(\omega_i / \mathbf{x}, D_i) = \frac{p(\mathbf{x} / \omega_i, D_i) P(\omega_i / D_i)}{\sum_j p(\mathbf{x} / \omega_j, D_j) P(\omega_j / D_j)}$$

- To simplify things, let's assume that $P(\omega_i/D) = P(\omega_i)$:

$$P(\omega_i / \mathbf{x}, D_i) = \frac{p(\mathbf{x} / \omega_i, D_i) P(\omega_i)}{\sum_j p(\mathbf{x} / \omega_j, D_j) P(\omega_j)}$$

Bayesian Estimation (Problem Definition)

- Problem: need to estimate $p(x|\omega_i, D_i)$ for every class ω_i
  - If samples in D_j give no information about $\boldsymbol{\theta}_i$ ($i \neq j$), we need to solve c independent problems (i.e., one for each class)

"given D, estimate p(x|D)"

A General BE strategy

- Estimate $p(x|D)$ as follows:

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

- Since the distribution is known completely given θ, we have:

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

  - Important equation to link $p(x|D)$ with $p(\theta|D)$

A General BE strategy

- Steps
  1. Compute $p(\theta/D)$

$$p(\boldsymbol{\theta} / D) = \frac{p(D / \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(D)} = a \prod_{k=1}^{n} p(\mathbf{x}_k / \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

  2. Compute $p(x/D)$

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

- $p(x/D)$ will replace $p(x/\omega_i, D_i)$ in Bayes' rule:

$$P(\omega_i / \mathbf{x}, D_i) = \frac{p(\mathbf{x} / \omega_i, D_i) P(\omega_i)}{\sum_j p(\mathbf{x} / \omega_j, D_j) P(\omega_j)}$$

BE Gaussians from Data (Case of Unknown θ = μ)

- Estimate $\boldsymbol{\theta}$ using the a-posteriori density $P(\boldsymbol{\theta}|D)$
  - The univariate Gaussian case: $P(\mu|D)$
  - μ is the only unknown parameter
  - Prior distribution $P(\mu)$ is provided

$$P(x \mid \mu) \sim N(\mu, \sigma^2)$$

$$P(\mu) \sim N(\mu_0, \sigma_0^2)$$

BE Gaussians from Data (Case of Unknown θ = μ)

$$P(\mu \mid D) = \frac{P(D \mid \mu).P(\mu)}{\int P(D \mid \mu).P(\mu)d\mu} \qquad (1)$$

$$= \alpha \prod_{k=1}^{k=n} P(x_k \mid \mu).P(\mu)$$

$$P(\mu \mid D) \sim N(\mu_n, \sigma_n^2) \qquad (2)$$

- The updated parameters of the prior:

$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}.\mu_0, \quad and \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

- μ_n is our best guess for μ; σ_n^2 measures the uncertainty.

BE Gaussians from Data (Case of Unknown θ = μ)

- The updated parameters of the prior:

$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}.\mu_0, \quad and \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

- μ_n is a linear combination of μ_n and μ_0

- If $\sigma \neq 0$, then μ_n approaches μ_n (maximum likelihood estimate !!)

- If σ_0=0, then $\mu = \mu_0$; If $\sigma_0 \gg \sigma$ then $\mu_n = \bar{\mu}_n$

BE Gaussians from Data (Case of Unknown θ = μ)

- The updated parameters of the prior:

$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}.\mu_0, \quad and \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

- σ_n^2 approaches σ^2/n as n increases (more observations will decrease our uncertainty about μ)

- $p(\mu/D)$ becomes sharply peaked (*Bayesian Learning*) as n increases.

BE Gaussians from Data (Case of Unknown θ = μ)

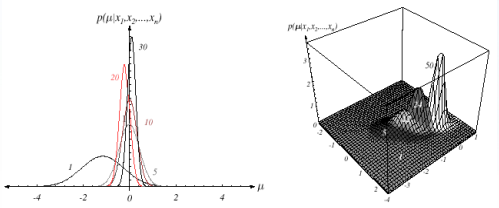


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

BE Gaussians from Data (Case of Unknown θ = μ)

- The univariate case P(x | D)
  - P(μ | D) has been computed
  - P(x | D) remains to be computed!

$$P(x \mid D) = \int P(x \mid \mu).P(\mu \mid D)d\mu \quad \text{is Gaussian}$$

$$P(x \mid D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

BE Gaussians from Data (Case of Unknown θ = μ)

- Bayesian Classification Rule
  - Desired class-conditional density P(x | D_j, ω_j)
  - P(x | D_j, ω_j) together with P(ω_j) and using Bayes formula, we obtain the Bayesian classification rule:

$$\underset{\omega_j}{Max}\left[P(\omega_j \mid x, D\right] \equiv \underset{\omega_j}{Max}\left[P(x \mid D_j, \omega_j).P(\omega_j)\right]$$

Summary on BE

- P(x|D) computation can be applied to any situation in which the unknown density can be parameterized. The basic assumptions are:
  - The form of P(x|**θ**) is assumed known, but the value of **θ** is not known exactly
  - Our knowledge about **θ** is assumed to be contained in a known prior density P(**θ**)
  - The rest of our knowledge about **θ** is contained in a set D of n random variables $x_1, x_2, \ldots, x_n$ that follows P(x)

Summary on BE

- The basic problem is:
  "**Compute the posterior density P(θ|D)**"
  then "**Derive P(x|D)**"
- Using Bayes formula, we have:

$$P(\theta \mid D) = \frac{P(D \mid \theta).P(\theta)}{\int P(D \mid \theta).P(\theta)d\theta},$$

  - by independence assumption:

$$P(D \mid \theta) = \prod_{k=1}^{k=n} P(x_k \mid \theta)$$

Summary on BE

- If we are less certain about the exact value of θ, we should consider a weighted average of p(x|θ) over the possible values of θ

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

  - Bayesian estimation approach estimates a distribution for p(x|D) rather than making point estimates like ML

Summary on BE

- Suppose p(θ|D) peaks very sharply at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $p(\hat{\boldsymbol{\theta}}) \neq 0$ then p(x/D) can be approximated as follows:

$$p(\mathbf{x}/D) \cong p(\mathbf{x}/\hat{\boldsymbol{\theta}})$$

  - i.e., the best estimate is obtained by setting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$
  - This is the ML solution
    - i.e., p(D|θ) peaks at $\hat{\boldsymbol{\theta}}$ too

since $p(\boldsymbol{\theta}/D) = \frac{p(D/\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}$

Summary on BE

- Given a large number of samples, p(θ|D) will have a very strong peak at $\hat{\boldsymbol{\theta}}$; in this case:

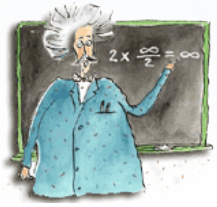
$$p(\mathbf{x}/D) \cong p(\mathbf{x}/\hat{\boldsymbol{\theta}})$$

  - When p(θ|D) contains more than one peaks (i.e., more than one θ explains the data), the solution p(x|θ) should be obtained by integration.

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

ML vs BE

- Comparisons



MLE vs Bayesian Estimation

Number of training data

- The two methods are equivalent assuming infinite number of training data
  - prior distributions do not exclude the true solution
- For small training data sets, they give different results in most cases.

Computational complexity

- ML uses differential calculus or gradient search for maximizing the likelihood.
- Bayesian estimation requires complex multidimensional integration techniques.

MLE vs Bayesian Estimation

Solution complexity

- Easier to interpret ML solutions (i.e., ***must be*** of the assumed parametric form).
- A Bayesian estimation solution ***might not*** be of the parametric form assumed.

Prior distribution

- If the prior distribution $p(\theta)$ is uniform, Bayesian estimation solutions are equivalent to ML solutions.
- Otherwise, the two methods will give different solutions.

MLE vs Bayesian Estimation

- General comments
  - There are strong theoretical and methodological arguments supporting Bayesian estimation.
  - In practice, ML estimation is simpler and can lead to comparable performance.

Seminar S03

- **Topic**
  - Give a presentation on how to choose the prior density in Bayesian Estimation
- **Requirements**
  - Prepare a **15 minutes** talk on your chosen topic
  - Make **ppt** to assist your talk
  - Prepare **at least 3 questions** to ask the audience after your talk
  - Get ready to **take questions** from the audience
- **Hints**
  - Check concepts such as
    - **Conjugate Prior**
    - **Beta distribution, Gammar distrbution**
    - **Dirichlet distribution**
    - **Inverse-Wishart distribution**

Exercise

- Suppose categorical arity-n inputs $x_1, x_2, \ldots x_n \sim$ (i.i.d.) from a multinomial

$$M(p_1, p_2, \ldots p_n)$$

where

$$P(x_k=j|\mathbf{p})=p_j$$

- What is the MLE $\mathbf{p}=(p_1, p_2, \ldots p_n)$?

Questions?