SYDE 372
Introduction to Pattern Recognition

Estimation and Learning: Part I

Alexander Wong

Department of Systems Design Engineering
University of Waterloo

Outline

Motivation

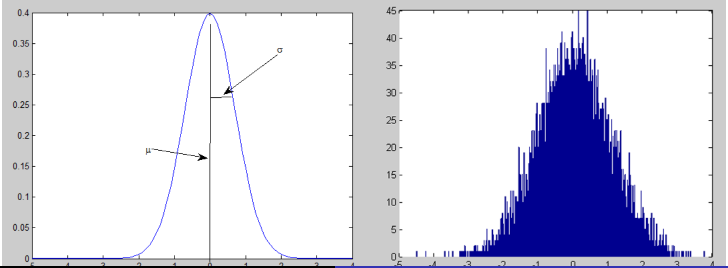
- From the previous chapter, we know that the Bayesian classifier achieves **minimum probability of error.**
- Therefore, it performs better than MICD classifier for situations where the class conditional PDFs ($p(\underline{x}|c_i)$) are **known**.
- **Problem:** In general, the PDFs are not known a priori, so how do we perform Bayesian classification?
- **Idea:**
    - What if we have samples with known class labels?
    - With these samples, we can try to learn the PDFs of the individual classes.
    - These empirical PDFs allow us to apply Bayesian classification!

Motivation

- Bayesian classification is optimal in terms of probability of error **ONLY** if the true class conditional PDFs ($p(\underline{x}|c_i)$) are known.
- The use of empirical PDFs result in sub-optimal classifiers.
- How close the performance is to the theoretical minimum $P(\epsilon)$ depends on the accuracy of the estimated PDF compared to the true PDF.

Categories of Statistical Learning

- There are two main categories of statistical learning approaches:
  - **Parametric Estimation:** functional form of PDF is assumed to be known and the necessary parameters of the PDF are estimated.
  - **Non-parametric Estimation:** functional form of PDF is not assumed to be known and the PDF is estimated directly.



Parametric Learning

- Here, we assume that we know the class conditional probability function, but we don't know the parameters that define this function.
- For example,
  - Suppose that $p(\underline{x}|A)$ is multivariate Normal, $\mathcal{N}(\underline{mu}_A, \Sigma_A)$
  - We may not know what the actual value of parameters μ_A and/or Σ_A are!
- In this scenario, what we want to do is estimate what these parameters are based on a set of labeled samples for the class!

Types of Parametric Estimation

- There are two main categories of parametric estimation approaches:
  - **Maximum Likelihood Estimation:** Treat parameters as being fixed but unknown quantities, with the goal of finding estimate values which maximize the probability that the given samples came from the resulting PDF $p(\underline{x}|\theta)$.
  - **Bayesian (Maximum a Posteriori) Estimation:** Treat parameters as random variables with an assumed a priori distribution, with the goal of obtaining an a posteriori distribution $p(\theta|\underline{x})$ which indicates the estimate value based on the given samples.

Maximum Likelihood Estimation

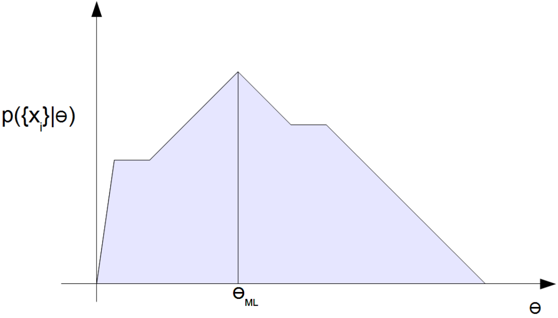
- Supposed that we are given a set of samples $\{\underline{x}_i\}$, independently drawn from a distribution $p(\underline{x})$ where the form of the PDF is known (e.g., Gaussian).
- The goal is to obtain estimates for the parameters θ that defines this PDF.
- For example, if the distribution is of the Gaussian form: $p(\underline{x}|A) = \mathcal{N}(\underline{mu}_A, \Sigma_A)$, then the set of parameters defining this distribution is $\theta = (\underline{mu}_A, \Sigma_A)$.

Maximum Likelihood Estimation

- Writing the PDF as $p(\underline{x}|\theta)$ to emphasize the dependence on parameters, the Maximum Likelihood estimate of the parameters θ is the set of parameters that maximizes the probability that the given samples $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$ are obtained given θ:

$$\hat{\theta}_{ML} = \text{argmax}_{\theta}\left[p(\{\underline{x}_i\}|\theta)\right] \tag{1}$$

Maximum Likelihood Estimation



Given an observation $\underline{x}_i$, the maximum likelihood estimate of parameter θ is chosen to be that value which maximizes the PDF $p(\underline{x}_i|\theta)$

Maximum Likelihood Estimation

- Assuming that the sample are independent of each other, $p(\{\underline{x}_i\}|\theta)$ becomes:

$$p(\{\underline{x}_i\}|\theta) = p(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N|\theta) = \prod_{i=1}^{N} p(\underline{x}_i|\theta) \qquad (2)$$

- Therefore, the sample set probability is just the product of the individual sample probabilities.

Maximum Likelihood Estimation

- To maximize $p(\{\underline{x}_i\}|\theta)$, we take the derivative and set it to zero:

$$\frac{\partial}{\partial\theta}p(\{\underline{x}_i\}|\theta)|_{\theta=\hat{\theta}_{ML}} = 0 \tag{3}$$

- It is often more convenient to deal with $p(\{\underline{x}_i\}|\theta)$ in log form

$$l(\theta) = \log[p(\{\underline{x}_i\}|\theta)] = \sum_{i=1}^{N}\log p(\underline{x}_i|\theta) \tag{4}$$

- This gives us the final maximum likelihood condition:

$$\frac{\partial}{\partial\theta}l(\theta)|_{\theta=\hat{\theta}_{ML}} = 0 \tag{5}$$

Maximum Likelihood Estimation: Example

- Example 1: Suppose that we would like the learn the underlying PDF and we are given the following information:
  - We know that the PDF is a univariate Normal distribution $p(x) = \mathcal{N}(\mu, \sigma^2)$.
  - We do not know what the mean μ is.
  - We know what the variance σ^2 is.

  What is the maximum likelihood estimate of μ?

Maximum Likelihood Estimation: Example

- Since the parameter that we do not know is the mean μ, what we have is $\theta = \mu$ and $p(x|\theta) = \mathcal{N}(\theta|\sigma^2)$.
- Therefore:

$$p(\{x_i\}|\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}(\frac{x_i-\theta}{\sigma})^2\right] \tag{6}$$

- Taking the log gives us:

$$l(\theta) = \log\left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}(\frac{x_i-\theta}{\sigma})^2\right]\right] \tag{7}$$

Maximum Likelihood Estimation: Example

- Taking the log gives us:

$$l(\theta) = \sum_{i=1}^{N} \left[\left[-\frac{1}{2} \left(\frac{x_i - \theta}{\sigma} \right)^2 \right] - \log \sqrt{2\pi}\sigma \right] \quad (8)$$

- Taking the derivative:

$$\frac{\partial}{\partial \theta} l(\theta) = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \theta) \quad (9)$$

Maximum Likelihood Estimation: Example

- Setting it to zero:

$$\frac{\partial}{\partial\theta} l(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \theta) = 0 \tag{10}$$

$$\sum_{i=1}^{N}(x_i) - N\theta = 0 \tag{11}$$

$$\theta = \frac{1}{N}\sum_{i=1}^{N}(x_i) \tag{12}$$

- Therefore, the maximum likelihood estimate for θ in this case is just the sample mean!

Maximum Likelihood Estimation: Example

- Example 2: Suppose that we would like the learn the underlying PDF and we are given the following information:
  - We know that the PDF is a univariate Normal distribution $p(x) = \mathcal{N}(\mu, \sigma^2)$.
  - We do not know what the mean μ is.
  - We do not know what the variance σ^2 is.

  What is the maximum likelihood estimates of μ and $sigma^2$?

Maximum Likelihood Estimation: Example

- Since the parameters that we do not know are the mean μ and variance σ^2, what we have is:

$$\underline{\theta} = [\theta_1 \;\; \theta_2]^T = [\mu \;\; \sigma^2]^T \tag{13}$$

and $p(x|\underline{\theta}) = \mathcal{N}(\theta_1, \theta_2)$,

$$p(\{x_i\}|\underline{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{1}{2}\frac{(x_i-\theta_1)^2}{\theta_2}\right] \tag{14}$$

- Taking the log gives us:

$$l(\underline{\theta}) = \log\left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{1}{2}\frac{(x_i-\theta_1)^2}{\theta_2}\right]\right] \tag{15}$$

Maximum Likelihood Estimation: Example

- Taking the log gives us:

$$l(\underline{\theta}) = \sum_{i=1}^{N} \left[\left[-\frac{1}{2} \frac{(x_i - \theta_1)^2}{\theta_2} \right] - \log \sqrt{2\pi\theta_2} \right] \qquad (16)$$

$$l(\underline{\theta}) = -\frac{1}{2} \sum_{i=1}^{N} \left[\frac{(x_i - \theta_1)^2}{\theta_2} \right] - \frac{N}{2} \log 2\pi\theta_2 \qquad (17)$$

Maximum Likelihood Estimation: Example

- Given that we have multiple parameters to estimate, we must maximize $l(\theta)$ with respect to each of the components of θ via a vector derivative:

$$\frac{\partial l(\underline{\theta})}{\partial \theta} = \left[\frac{\partial l(\underline{\theta})}{\partial \theta_1} \quad \frac{\partial l(\underline{\theta})}{\partial \theta_2} \right]^T \tag{18}$$

- Taking the derivative of each component gives us:

$$\frac{\partial l(\underline{\theta})}{\partial \theta_1} = \sum_{i=1}^{N} \frac{x_i - \theta_1}{\theta_2} \tag{19}$$

$$\frac{\partial l(\underline{\theta})}{\partial \theta_2} = \frac{1}{2} \sum_{i=1}^{N} \frac{(x_i - \theta_1)^2}{\theta_2^2} - \frac{N}{2\theta_2} \tag{20}$$

Maximum Likelihood Estimation: Example

- Setting $\frac{\partial l(\underline{\theta})}{\partial \theta_1} = 0$ and solving gives us:

$$\frac{\partial l(\underline{\theta})}{\partial \theta_1} = \sum_{i=1}^{N} \frac{x_i - \theta_1}{\theta_2} = 0 \tag{21}$$

$$\sum_{i=1}^{N} x_i = N\theta_1 \tag{22}$$

$$\theta_1 = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{23}$$

- Same as before! The maximum likelihood estimate for θ_1 is $\hat{\theta}_{1,ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

Maximum Likelihood Estimation: Example

- Setting $\frac{\partial l(\underline{\theta})}{\partial \theta_2} = 0$, plugging in $\hat{\theta}_{1,ML}$ gives us:

$$\frac{\partial l(\underline{\theta})}{\partial \theta_2} = \frac{1}{2}\sum_{i=1}^{N} \frac{(x_i - \hat{\theta}_{1,ML})^2}{\theta_2^2} - \frac{N}{2\theta_2} = 0 \tag{24}$$

$$\sum_{i=1}^{N} \frac{(x_i - \hat{\theta}_{1,ML})^2}{\theta_2^2} = \frac{N}{\theta_2} \tag{25}$$

$$\sum_{i=1}^{N} (x_i - \hat{\theta}_{1,ML})^2 = N\theta_2 \tag{26}$$

$$\hat{\theta}_{2,ML} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\theta}_{1,ML})^2 \tag{27}$$

Maximum Likelihood Estimation: Example

- Example 3: Suppose that we would like the learn the underlying PDF and we are given the following information:
  - We know that the PDF is a multivariate Normal distribution $p(\underline{x}) = \mathcal{N}(\underline{\mu}, \Sigma)$.
  - We do not know what the mean vector $\underline{\mu}$ is.
  - We do not know what the covariance matrix Σ is.

  What is the maximum likelihood estimates of $\underline{\mu}$ and Σ?

Maximum Likelihood Estimation: Example

- To simplify derivation, let us defined our parameters as follows:
  - $\underline{\theta}_1 = \underline{\mu}$
  - $\theta_2 = \Sigma^{-1}$
- Therefore, $p(\underline{x}|\underline{\theta})$ can be written as:

$$p(\underline{x}|\underline{\theta}) = \frac{|\theta_2|^{1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\underline{x} - \underline{\theta}_1)^T \theta_2 (\underline{x} - \underline{\theta}_1)\right] \quad (28)$$

- Taking the log gives us:

$$l(\underline{\theta}) = \sum_{i=1}^{N} \frac{1}{2} \log |\theta_2| - \frac{n}{2} \log 2\pi - \frac{1}{2}(\underline{x}_i - \underline{\theta}_1)^T \theta_2 (\underline{x}_i - \underline{\theta}_1) \quad (29)$$

Maximum Likelihood Estimation: Example

- Taking the derivative $\frac{\partial l(\underline{\theta})}{\partial \theta_1}$ and setting it to zero

$$\frac{\partial l(\underline{\theta})}{\partial \theta_1} = \frac{\partial}{\partial \theta_1}\left[-\frac{1}{2}\sum_{i=1}^{N}(\underline{x}_i - \underline{\theta}_1)^T \theta_2 (\underline{x}_i - \underline{\theta}_1)\right] = 0 \quad (30)$$

$$\sum_{i=1}^{N}(\underline{x}_i - \underline{\theta}_1) = 0 \quad (31)$$

$$\sum_{i=1}^{N}(\underline{x}_i) = N\underline{\theta}_1 \quad (32)$$

$$\underline{\theta}_1 = \frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i) = \underline{\hat{\mu}}_{ML} \quad (33)$$

Maximum Likelihood Estimation: Example

- Taking the derivative $\frac{\partial l(\underline{\theta})}{\partial \theta_2}$

$$\frac{\partial l(\underline{\theta})}{\partial \theta_2} = \frac{1}{2}\sum_{i=1}^{N} \frac{\partial \log|\theta_2|}{\partial \theta_2} - \frac{1}{2}\sum_{i=1}^{N} \frac{\partial}{\partial \theta_2}\left[(\underline{x}_i - \underline{\theta}_1)^T \theta_2 (\underline{x}_i - \underline{\theta}_1)\right] \tag{34}$$

$$\frac{\partial l(\underline{\theta})}{\partial \theta_2} = \frac{1}{2}\sum_{i=1}^{N} \frac{cof\theta_2}{|\theta_2|} - \frac{1}{2}\sum_{i=1}^{N}(\underline{x}_i - \underline{\theta}_1)(\underline{x}_i - \underline{\theta}_1)^T \tag{35}$$

$$\frac{\partial l(\underline{\theta})}{\partial \theta_2} = \frac{1}{2}\sum_{i=1}^{N}[\theta_2^{-1}]^T - \frac{1}{2}\sum_{i=1}^{N}(\underline{x}_i - \underline{\theta}_1)(\underline{x}_i - \underline{\theta}_1)^T \tag{36}$$

Maximum Likelihood Estimation: Example

- Setting $\frac{\partial l(\underline{\theta})}{\partial \theta_2}$ to zero:

$$\frac{1}{2}\sum_{i=1}^{N}[\theta_2^{-1}]^T - \frac{1}{2}\sum_{i=1}^{N}(\underline{x}_i - \underline{\theta}_1)(\underline{x}_i - \underline{\theta}_1)^T = 0 \qquad (37)$$

$$\sum_{i=1}^{N}[\theta_2^{-1}]^T = \sum_{i=1}^{N}(\underline{x}_i - \underline{\theta}_1)(\underline{x}_i - \underline{\theta}_1)^T \qquad (38)$$

$$[\theta_2^{-1}]^T = \frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i - \underline{\theta}_1)(\underline{x}_i - \underline{\theta}_1)^T \qquad (39)$$

Maximum Likelihood Estimation: Example

- Since we are dealing with a symmetric matrix:

$$[\theta_2^{-1}] = \frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i - \underline{\theta}_1)(\underline{x}_i - \underline{\theta}_1)^T \quad (40)$$

- With $\theta_2^{-1} = \Sigma$:

$$\hat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i - \underline{\hat{\mu}}_{ML})(\underline{x}_i - \underline{\hat{\mu}}_{ML})^T \quad (41)$$

Maximum Likelihood Estimation: Example

- Example 3: Suppose we wish to find the maximum likelihood estimate of the performance of a classifier
- We are told that the classifier has a true error rate of θ.
- On any given set of N test samples, the probability that k samples are misclassified is given by the binomial distribution:

$$p(k|\theta) = \left[\begin{array}{c} N \\ k \end{array} \right] \theta^k (1-\theta)^{N-k} \quad (42)$$

- where $\left[\begin{array}{c} N \\ k \end{array} \right] = \frac{N!}{k!(N-k)!}$ is the number of ways that any k out of the N samples can be misclassified.

Maximum Likelihood Estimation: Example

- Taking the log gives us:

$$l(\theta) = \log p(k|\theta) = \log \left[\begin{array}{c} N \\ k \end{array} \right] + \log[\theta^k] + \log[(1-\theta)^{N-k}] \tag{43}$$

$$l(\theta) = \log p(k|\theta) = \log \left[\begin{array}{c} N \\ k \end{array} \right] + k\log[\theta] + (N-k)\log[1-\theta] \tag{44}$$

Maximum Likelihood Estimation: Example

- Taking the derivative and setting to zero:

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{k}{\theta} - \frac{N-k}{1-\theta} = 0 \tag{45}$$

$$(1-\theta)k = \theta(N-k) \tag{46}$$

$$\hat{\theta}_{ML} = \frac{k}{N} \tag{47}$$

- Therefore, the ML estimate of error rate is just the fraction of samples misclassified.

Estimation Bias

- ML estimates are optimal in the sense of maximizing probability of observing the given samples
- However, we may also require that the estimates be unbiased. What does that mean?
- **Formal definition**: an estimate $\underline{\hat{\theta}}$ is unbiased if its expected value is equal to the true value:

$$E[\underline{\hat{\theta}}] = \underline{\theta} \tag{48}$$

Estimation Bias

- Example: Is the ML estimate of the mean unbiased?

$$E[\underline{\hat{\mu}}_{ML}] = E[\frac{1}{N}\sum_{i=1}^{N}\underline{x}_i] \tag{49}$$

$$E[\underline{\hat{\mu}}_{ML}] = \frac{1}{N}\sum_{i=1}^{N}E[\underline{x}_i] \tag{50}$$

- Since $\mu = E[\underline{x}]$,

$$E[\underline{\hat{\mu}}_{ML}] = \frac{1}{N}\sum_{i=1}^{N}\underline{\mu} = \underline{\mu} \tag{51}$$

- Therefore, the ML estimate of the mean is unbiased!

Estimation Bias

- Example: Is the ML estimate of the covariance matrix unbiased?

$$E[\hat{\Sigma}_{ML}] = E[\frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i - \underline{\hat{\mu}}_{ML})(\underline{x}_i - \underline{\hat{\mu}}_{ML})^T] \tag{52}$$

$$E[\hat{\Sigma}_{ML}] = \frac{1}{N}\sum_{i=1}^{N}E[(\underline{x}_i - \underline{\hat{\mu}}_{ML})(\underline{x}_i - \underline{\hat{\mu}}_{ML})^T] \tag{53}$$

Estimation Bias

- Since variance is expressed in terms of mean $\underline{\mu}$,

$$E[\hat{\Sigma}_{ML}] = \frac{1}{N}\sum_{i=1}^{N} E[((\underline{x}_i - \underline{\mu}) - (\hat{\underline{\mu}}_{ML} - \underline{\mu}))((\underline{x}_i - \underline{\mu}) - (\hat{\underline{\mu}}_{ML} - \underline{\mu}))^T] \tag{54}$$

- Expanding and plugging in $\hat{\underline{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N} \underline{x}_i$:

$$E[\hat{\Sigma}_{ML}] = E[\frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})^T] - E[(\hat{\underline{\mu}}_{ML} - \underline{\mu})(\hat{\underline{\mu}}_{ML} - \underline{\mu})^T] \tag{55}$$

- Does the first term look familiar?

Estimation Bias

- The first term is just the variance Σ!

$$E[\hat{\Sigma}_{ML}] = \Sigma - E[(\underline{\hat{\mu}}_{ML} - \underline{\mu})(\underline{\hat{\mu}}_{ML} - \underline{\mu})^T] \tag{56}$$

- Substituting $\underline{\hat{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N} \underline{x}_i$ back in:

$$E[\hat{\Sigma}_{ML}] = \Sigma - E[(\frac{1}{N}\sum_{i=1}^{N} \underline{x}_i - \underline{\mu})(\frac{1}{N}\sum_{i=1}^{N} \underline{x}_i - \underline{\mu})^T] \tag{57}$$

$$E[\hat{\Sigma}_{ML}] = \Sigma - E[(\frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i - \underline{\mu}))(\frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i - \underline{\mu}))^T] \tag{58}$$

$$E[\hat{\Sigma}_{ML}] = \Sigma - \frac{1}{N^2}E[\sum_{i=1}^{N}\sum_{j=1}^{N}(\underline{x}_i - \underline{\mu})(\underline{x}_j - \underline{\mu})^T] \tag{59}$$

Estimation Bias

- Since the samples are independent,

$$E[(\underline{x}_i - \underline{\mu})(\underline{x}_j - \underline{\mu})^T] = 0 \text{ for } i \neq j \tag{60}$$

- Therefore,

$$E[\hat{\Sigma}_{ML}] = \Sigma - \frac{1}{N^2}\sum_{j=1}^{N} E[(\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})^T] \tag{61}$$

- Since $\Sigma = E[(\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})^T]$,

$$E[\hat{\Sigma}_{ML}] = \Sigma - \frac{1}{N^2} N\Sigma \tag{62}$$

Estimation Bias

- Continuing to simplify:

$$E[\hat{\Sigma}_{ML}] = \Sigma - \frac{1}{N^2} N\Sigma \tag{63}$$

$$E[\hat{\Sigma}_{ML}] = \Sigma - \frac{1}{N}\Sigma \tag{64}$$

$$E[\hat{\Sigma}_{ML}] = \frac{N-1}{N}\Sigma \tag{65}$$

- Therefore, the ML estimate for the covariance matrix is biased!
- As $N \leftarrow \infty$, the bias becomes negligible.

Estimation Bias

- Then how do we get an unbiased estimate?
- Answer: Just multiply your ML estimate by $\frac{N}{N-1}$!

$$E[\frac{N}{N-1}\hat{\Sigma}_{ML}] = \frac{N}{N-1}\frac{N-1}{N}\Sigma = \Sigma \tag{66}$$

- Bias stems from the use of ML estimate for mean, $\underline{\hat{\mu}}_{ML}$, rather than the true mean in the expression for $\hat{\Sigma}_{ML}$.

Bayesian Estimation

- Idea: Instead of treating the parameters as fixed and finding the parameters that maximize the probability that the given samples come from the resulting PDF, we do the following:
  - Treat the parameters as **random variables** with an assumed a priori distribution
  - Use the observed samples to obtain an a posterior distribution which indicates the parameters!

Bayesian Estimation

- Let $p(\theta)$ be the a priori distribution and $\{\underline{x}_i\}$ be the set of samples.
- The a posteriori distribution can be written as:

$$p(\underline{\theta}|\{\underline{x}_i\}) = \frac{p(\{\underline{x}_i\}|\underline{\theta})p(\underline{\theta})}{p(\{\underline{x}_i\})} \tag{67}$$

- The term $p(\{\underline{x}_i\})$ is treated as a scale factor which may be obtained from the requirement for PDFs:

$$\int p(\underline{\theta}|\{\underline{x}_i\})d\underline{\theta} = 1 \tag{68}$$

Bayesian Estimation

- Example: Suppose that we would like the learn the underlying PDF and we are given the following information:
  - We know that the PDF is a univariate Normal distribution $p(x) = \mathcal{N}(\mu, \sigma^2)$.
  - We do not know what the mean μ is.
  - We know what the variance σ^2 is.

  What is the Bayesian estimate of μ?

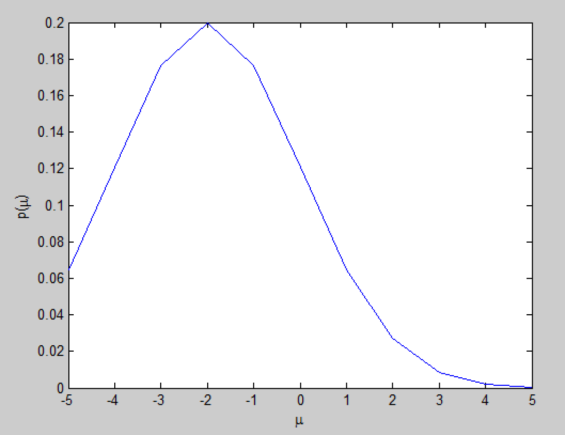
Bayesian Estimation

- Step 1: Assume an a priori PDF for parameter $\theta = \mu$

$$p(\mu) = \mathcal{N}(\mu|\mu_o, \sigma_o^2). \tag{69}$$

- What this means is that:
  - Initial guess for μ is μ_o, and
  - Uncertainty of our guess is normally distributed with variance σ_o^2.

Bayesian Estimation



A Priori PDF for $\theta = \mu$

Bayesian Estimation

- Step 2: Given samples, compute $p(\mu|\{x_i\})$

$$p(\mu|\{x_i\}) = \alpha p(\{x_i\}|\mu)p(\mu) \tag{70}$$

- Assuming that the samples are independent:

$$p(\mu|\{x_i\}) = \alpha \prod_{i=1}^{N} p(x_i|\mu)p(\mu) \tag{71}$$

where $\alpha = \frac{1}{p(\{x_i\})}$ is a scale factor independent of μ.

Bayesian Estimation

- Substituting $p(x_i|\mu)$ and $p(\mu) = \mathcal{N}(\mu|\mu_o, \sigma_o^2)$:

$$p(\mu|\{x_i\}) = \alpha \prod_{i=1}^{N} p(x_i|\mu)p(\mu) \quad (72)$$

$$p(\mu|\{x_i\}) = \alpha \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi\sigma_o}} exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_o}{\sigma_o}\right)^2\right] \quad (73)$$

Bayesian Estimation

- This can be rewritten as:

$$p(\mu|\{x_i\}) = \alpha' \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{N}\frac{x_i^2-2x_i\mu+\mu^2}{\sigma^2}+\frac{\mu^2-2\mu\mu_o+\mu_o^2}{\sigma_o^2}\right\}\right]. \quad (74)$$

$$p(\mu|\{x_i\}) = \alpha'' \exp\left[-\frac{1}{2}\left[\frac{N}{\sigma^2}+\frac{1}{\sigma_o^2}\right]\mu^2 - 2\left[\frac{Nm_N}{\sigma^2}+\frac{\mu_o}{\sigma_o^2}\right]\mu\right]. \quad (75)$$

  where m_N is the sample mean.
- It can be seen that the exponent is quadratic in μ, making it of the Gaussian form!

Bayesian Estimation

- If we complete the square, the a posteriori density is of the form:

$$p(\mu|\{x_i\}) = \mathcal{N}(\mu|\mu_N, \mu_N^2) \tag{76}$$

where μ_N can be defined as:

$$\mu_N = \frac{N\sigma_o^2}{N\sigma_o^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_o^2 + \sigma^2}\mu_o \tag{77}$$

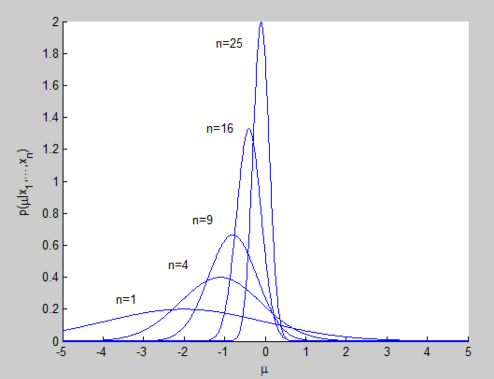
- The peak of the density is at μ_N, with a variance of σ_N^2.
- Based on this, the Bayesian estimate of μ is

$$\hat{\mu}_B = \mu_N \tag{78}$$

Bayesian Estimation

- Observations:
  - Bayesian estimate can be interpreted as weighted average of initial guess μ_o and sample mean m_N.
  - If $\sigma_o = 0$, we are so sure of initial guess that we ignore the samples.
  - If $\sigma_o > 0$, there is some uncertainty and the sample mean has greater dominance.
  - If $\sigma_o >> \sigma$, initial uncertainty is relatively large and samples weighted more heavily.
  - As $N \rightarrow \infty$, $\sigma_N^2 \rightarrow 0$ and $\mu_N \rightarrow m_N$.
  - This means that as the number of samples increases, the density narrows and peaks at true mean during the Bayesian learning process!

Bayesian Estimation



As measures arrive, the PDF of the estimate narrows, implying that the estimation error is decreasing.