Parametric Techniques
Lecture 3

Jason Corso

SUNY at Buffalo

22 January 2009

Introduction

- In Lecture 2, we learned how to form optimal decision boundaries when the full probabilistic structure of the problem is known.
- However, this is rarely the case in practice.
- Instead, we have some knowledge of the problem and some example data and we must estimate the probabilities.
- **Focus of this lecture** is to study a pair of techniques for estimating the parameters of the likelihood models (given a particular form of likelihood function, such as a Gaussian).

Introduction

- In Lecture 2, we learned how to form optimal decision boundaries when the full probabilistic structure of the problem is known.
- However, this is rarely the case in practice.
- Instead, we have some knowledge of the problem and some example data and we must estimate the probabilities.
- **Focus of this lecture** is to study a pair of techniques for estimating the parameters of the likelihood models (given a particular form of likelihood function, such as a Gaussian).
- **Parametric Models** – For a particular class ω_i, we consider a set of parameters $\boldsymbol{\theta}_i$ to fully define the likelihood model.
    - For the Guassian, $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

Introduction

- In Lecture 2, we learned how to form optimal decision boundaries when the full probabilistic structure of the problem is known.
- However, this is rarely the case in practice.
- Instead, we have some knowledge of the problem and some example data and we must estimate the probabilities.
- **Focus of this lecture** is to study a pair of techniques for estimating the parameters of the likelihood models (given a particular form of likelihood function, such as a Gaussian).
- **Parametric Models** – For a particular class ω_i, we consider a set of parameters $\boldsymbol{\theta}_i$ to fully define the likelihood model.
  - For the Guassian, $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
- **Supervised Learning** – we are working in a supervised situation where we have an set of training data:

$$\mathcal{D} = \{(\mathbf{x}, \omega)_1, (\mathbf{x}, \omega)_2, \dots (\mathbf{x}, \omega)_N\} \tag{1}$$

Overview of the Methods

- **Intuitive Problem:** Given a set of training data, $\mathcal{D}$, containing labels for c classes, train the likelihood models $p(\mathbf{x}|\omega_i, \theta_i)$ by estimating the parameters θ_i for $i = 1, \ldots, c$.

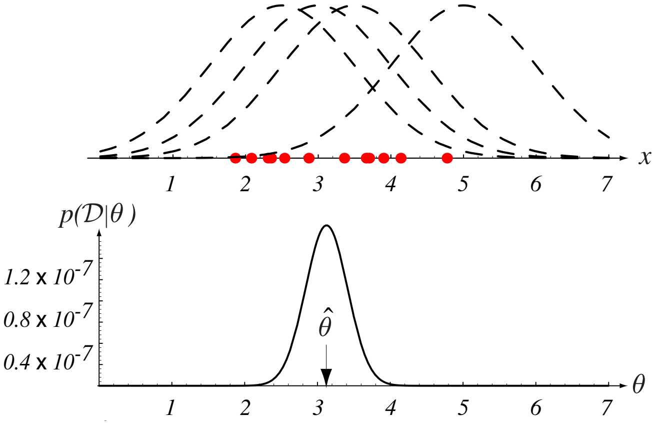
Overview of the Methods

- **Intuitive Problem:** Given a set of training data, $\mathcal{D}$, containing labels for c classes, train the likelihood models $p(\mathbf{x}|\omega_i, \theta_i)$ by estimating the parameters θ_i for $i = 1, \ldots, c$.
- **Maximum Likelihood Parameter Estimation**
  - Views the parameters as quantities that are fixed by unknown.
  - The best estimate of their value is the one that maximizes the probability of obtaining the samples in $\mathcal{D}$.

Overview of the Methods

- **Intuitive Problem:** Given a set of training data, $\mathcal{D}$, containing labels for c classes, train the likelihood models $p(\mathbf{x}|\omega_i, \theta_i)$ by estimating the parameters θ_i for $i = 1, \dots, c$.
- **Maximum Likelihood Parameter Estimation**
  - Views the parameters as quantities that are fixed by unknown.
  - The best estimate of their value is the one that maximizes the probability of obtaining the samples in $\mathcal{D}$.
- **Bayesian Parameter Estimation**
  - Views the parameters as random variables having some known prior distribution.
  - The samples converts this prior into a posterior and revises our estimate of the distribution over the parameters.
  - We shall typically see that the posterior is increasingly peaked for larger $\mathcal{D}$.

Maximum Likelihood Intuition



- Underlying model is assumed to be a Gaussian of particular variance but unknown mean.

Preliminaries

- Separate our training data according to class; i.e., we have c data sets $\mathcal{D}_1, \ldots, \mathcal{D}_c$.
- Assume that samples in $\mathcal{D}_i$ give no information for $\boldsymbol{\theta}_j$ for all $i \neq j$.

Preliminaries

- Separate our training data according to class; i.e., we have c data sets $\mathcal{D}_1, \dots, \mathcal{D}_c$.
- Assume that samples in $\mathcal{D}_i$ give no information for $\boldsymbol{\theta}_j$ for all $i \neq j$.
- Assume the samples in $\mathcal{D}_j$ have been drawn independently according to the (unknown but) fixed density $p(\mathbf{x}|\omega_j)$.
    - We say these samples are **i.i.d.** — independent and identically distributed.

Preliminaries

- Separate our training data according to class; i.e., we have c data sets $\mathcal{D}_1, \dots, \mathcal{D}_c$.
- Assume that samples in $\mathcal{D}_i$ give no information for $\boldsymbol{\theta}_j$ for all $i \neq j$.
- Assume the samples in $\mathcal{D}_j$ have been drawn independently according to the (unknown but) fixed density $p(\mathbf{x}|\omega_j)$.
  - We say these samples are **i.i.d.** — independent and identically distributed.
- Assume $p(\mathbf{x}|\omega_j)$ has some fixed parametric form and is fully described by $\boldsymbol{\theta}_j$; hence we write $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$.

Preliminaries

- Separate our training data according to class; i.e., we have c data sets $\mathcal{D}_1, \ldots, \mathcal{D}_c$.
- Assume that samples in $\mathcal{D}_i$ give no information for $\boldsymbol{\theta}_j$ for all $i \neq j$.
- Assume the samples in $\mathcal{D}_j$ have been drawn independently according to the (unknown but) fixed density $p(\mathbf{x}|\omega_j)$.
  - We say these samples are **i.i.d.** — independent and identically distributed.
- Assume $p(\mathbf{x}|\omega_j)$ has some fixed parametric form and is fully described by $\boldsymbol{\theta}_j$; hence we write $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$.
- We thus have c separate problems of the form:

Definition

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

(Log-)Likelihood

- Because we assume i.i.d. we have

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}) \quad . \tag{2}$$

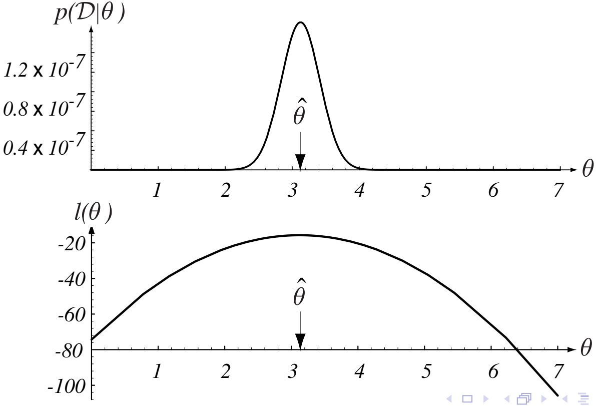
- The log-likelihood is typically easier to work with both analytically and numerically.

$$\begin{aligned} l_{\mathcal{D}}(\boldsymbol{\theta}) \equiv l(\boldsymbol{\theta}) &\doteq \ln p(\mathcal{D}|\boldsymbol{\theta}) && (3) \\ &= \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) && (4) \end{aligned}$$

Maximum (Log-)Likelihood

- The **maximum likelihood estimate** of $\boldsymbol{\theta}$ if the value $\hat{\boldsymbol{\theta}}$ that maximizes $p(\mathcal{D}|\boldsymbol{\theta})$ or equivalently maximizes $l_{\mathcal{D}}(\boldsymbol{\theta})$.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l_{\mathcal{D}}(\boldsymbol{\theta}) \tag{5}$$



Necessary Conditions for MLE

- For p parameters, $\boldsymbol{\theta} \doteq \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_p \end{bmatrix}^\mathsf{T}$.
- Let $\boldsymbol{\nabla}_{\boldsymbol{\theta}}$ be the gradient operator, then $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \doteq \begin{bmatrix} \frac{\partial}{\partial\theta_1} & \dots & \frac{\partial}{\partial\theta_p} \end{bmatrix}^\mathsf{T}$.
- The set of **necessary conditions** for the maximum likelihood estimate of $\boldsymbol{\theta}$ are obtained from the following system of p equations:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) = 0 \tag{6}$$

Necessary Conditions for MLE

- For p parameters, $\boldsymbol{\theta} \doteq \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_p \end{bmatrix}^\mathsf{T}$.
- Let $\boldsymbol{\nabla}_{\boldsymbol{\theta}}$ be the gradient operator, then $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \doteq \begin{bmatrix} \frac{\partial}{\partial \theta_1} & \dots & \frac{\partial}{\partial \theta_p} \end{bmatrix}^\mathsf{T}$.
- The set of **necessary conditions** for the maximum likelihood estimate of $\boldsymbol{\theta}$ are obtained from the following system of p equations:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k | \boldsymbol{\theta}) = 0 \tag{6}$$

- A solution $\hat{\boldsymbol{\theta}}$ to (6) can be a true global maximum, a local maximum or minimum or an inflection point of $l(\boldsymbol{\theta})$.

Necessary Conditions for MLE

- For p parameters, $\boldsymbol{\theta} \doteq \begin{bmatrix} \theta_1 & \theta_2 & \ldots & \theta_p \end{bmatrix}^{\mathsf{T}}$.
- Let $\boldsymbol{\nabla}_{\boldsymbol{\theta}}$ be the gradient operator, then $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \doteq \begin{bmatrix} \frac{\partial}{\partial \theta_1} & \ldots & \frac{\partial}{\partial \theta_p} \end{bmatrix}^{\mathsf{T}}$.
- The set of **necessary conditions** for the maximum likelihood estimate of $\boldsymbol{\theta}$ are obtained from the following system of p equations:

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) = 0 \tag{6}$$

- A solution $\hat{\boldsymbol{\theta}}$ to (6) can be a true global maximum, a local maximum or minimum or an inflection point of $l(\boldsymbol{\theta})$.
- Keep in mind that $\hat{\boldsymbol{\theta}}$ is only an estimate. Only in the limit of an infinitely large number of training samples can we expect it to be the true parameters of the underlying density.

Gaussian Case with Known Σ and Unknown μ

- For a single sample point $\mathbf{x}_k$:

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{7}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{8}$$

Gaussian Case with Known Σ and Unknown μ

- For a single sample point $\mathbf{x}_k$:

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{7}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{8}$$

- We see that the ML-estimate must satisfy

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = 0 \tag{9}$$

Gaussian Case with Known Σ and Unknown μ

- For a single sample point $\mathbf{x}_k$:

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{7}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{8}$$

- We see that the ML-estimate must satisfy

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = 0 \tag{9}$$

- And we get the sample mean!

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \tag{10}$$

Univariate Gaussian Case with Unknown μ and σ^2

The Log-Likelihood

- Let $\boldsymbol{\theta} = (\mu, \sigma^2)$. The log-likelihood of x_k is

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln\left[2\pi\sigma^2\right] - \frac{1}{2\sigma^2}(x_k-\mu)^2 \tag{11}$$

$$\nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma^2}(x_k-\mu) \\ -\frac{1}{2\sigma^2} + \frac{(x_k-\mu)^2}{2\sigma^4} \end{bmatrix} \tag{12}$$

Univariate Gaussian Case with Unknown μ and σ^2

Necessary Conditions

- The following conditions are defined:

$$\sum_{k=1}^{n} \frac{1}{\hat{\sigma}^2}(x_k - \hat{\mu}) = 0 \tag{13}$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\sigma}^2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\mu})^2}{\hat{\sigma}^4} = 0 \tag{14}$$

Univariate Gaussian Case with Unknown μ and σ^2

ML-Estimates

- After some manipulation we have the following:

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} x_k \tag{15}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu})^2 \tag{16}$$

- These are encouraging results – even in the case of unknown μ and σ^2 the ML-estimate of μ corresponds to the sample mean.

Bias

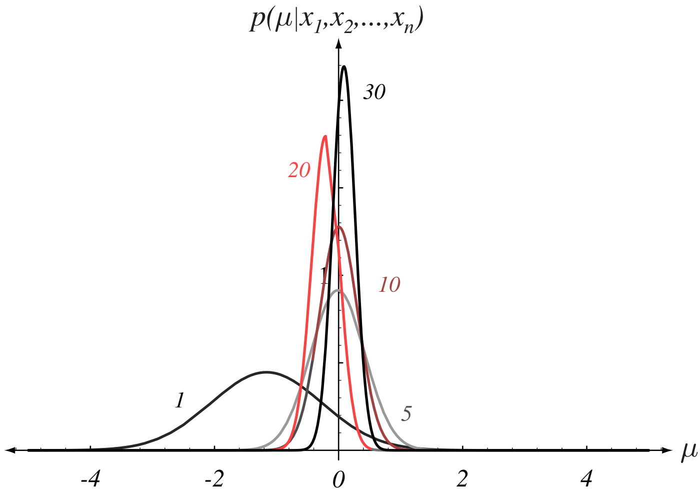
- The maximum likelihood estimate for the variance σ^2 is **biased**.
- The expected value over datasets of size n of the sample variance is not equal to the true variance

$$\mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \tag{17}$$

- In other words, the ML-estimate of the variance systematically underestimates the variance of the distribution.
- As $n \to \infty$ the problem of bias is reduced or removed, but bias remains a problem of the ML-estimator.
- An unbiased ML-estimator of the variance is

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \hat{\mu})^2 \tag{18}$$

Bayesian Parameter Estimation Intuition



General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).

General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).
- The values of the parameter vector $\boldsymbol{\theta}$ are not exactly known.

General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).
- The values of the parameter vector $\boldsymbol{\theta}$ are not exactly known.
- Our initial knowledge about the parameters is summarized in a prior distribution $p(\boldsymbol{\theta})$.

General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).
- The values of the parameter vector $\boldsymbol{\theta}$ are not exactly known.
- Our initial knowledge about the parameters is summarized in a prior distribution $p(\boldsymbol{\theta})$.
- The rest of our knowledge about $\boldsymbol{\theta}$ is contained in a set $\mathcal{D}$ of n i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ drawn according to fixed $p(\mathbf{x})$.

General Assumptions

Bayesian Parameter Estimation

- The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known (e.g., it is a Gaussian).
- The values of the parameter vector $\boldsymbol{\theta}$ are not exactly known.
- Our initial knowledge about the parameters is summarized in a prior distribution $p(\boldsymbol{\theta})$.
- The rest of our knowledge about $\boldsymbol{\theta}$ is contained in a set $\mathcal{D}$ of n i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ drawn according to fixed $p(\mathbf{x})$.

Goal

Our ultimate goal is to estimate $p(\mathbf{x}|\mathcal{D})$, which is as close as we can come to estimating the unknown $p(\mathbf{x})$.

Linking Likelihood and the Parameter Distribution

- How do we relate the prior distribution on the parameters to the samples?

Linking Likelihood and the Parameter Distribution

- How do we relate the prior distribution on the parameters to the samples?
- **Missing Data!** The samples will convert our prior $p(\boldsymbol{\theta})$ to a posterior $p(\boldsymbol{\theta}|\mathcal{D})$, by integrating the joint density over $\boldsymbol{\theta}$:

$$
\begin{aligned}
p(\mathbf{x}|\mathcal{D}) &= \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} && (19) \\
&= \int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} && (20)
\end{aligned}
$$

Linking Likelihood and the Parameter Distribution

- How do we relate the prior distribution on the parameters to the samples?
- **Missing Data!** The samples will convert our prior $p(\boldsymbol{\theta})$ to a posterior $p(\boldsymbol{\theta}|\mathcal{D})$, by integrating the joint density over $\boldsymbol{\theta}$:

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{19}$$

$$= \int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{20}$$

- And, because the distribution of $\mathbf{x}$ is known given the parameters $\boldsymbol{\theta}$, we simplify to

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{21}$$

Linking Likelihood and the Parameter Distribution

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

- We can see the link between the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ and the posterior for the unknown parameters $p(\boldsymbol{\theta}|\mathcal{D})$.

Linking Likelihood and the Parameter Distribution

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

- We can see the link between the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ and the posterior for the unknown parameters $p(\boldsymbol{\theta}|\mathcal{D})$.
- If the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply for sample point $\hat{\boldsymbol{\theta}}$, then we obtain

$$p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \quad . \tag{22}$$

Linking Likelihood and the Parameter Distribution

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

- We can see the link between the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ and the posterior for the unknown parameters $p(\boldsymbol{\theta}|\mathcal{D})$.
- If the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply for sample point $\hat{\boldsymbol{\theta}}$, then we obtain

$$p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \quad . \tag{22}$$

- And, we will see that during Bayesian parameter estimation, the distribution over the parameters will get increasingly "peaky" as the number of samples increases.

Linking Likelihood and the Parameter Distribution

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

- We can see the link between the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ and the posterior for the unknown parameters $p(\boldsymbol{\theta}|\mathcal{D})$.
- If the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply for sample point $\hat{\boldsymbol{\theta}}$, then we obtain

$$p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \quad . \tag{22}$$

- And, we will see that during Bayesian parameter estimation, the distribution over the parameters will get increasingly "peaky" as the number of samples increases.
- What if the integral is not readily analytically computed?

The Posterior Density on the Parameters

- The primary task in Bayesian Parameter Estimation is the computation of the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$.
- By Bayes formula

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{23}$$

- Z is a normalizing constant:

$$Z = \int p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{24}$$

The Posterior Density on the Parameters

- The primary task in Bayesian Parameter Estimation is the computation of the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$.
- By Bayes formula

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{23}$$

- Z is a normalizing constant:

$$Z = \int p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{24}$$

- And, by the independence assumption on $\mathcal{D}$:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}) \tag{25}$$

- Let's see some examples before we return to this general formulation.

Univariate Gaussian Case with Known σ^2

- Assume $p(x|\mu) \sim N(\mu, \sigma^2)$ with known σ^2.
- Whatever prior knowledge we know about μ is expressed in $p(\mu)$, which is known.

Univariate Gaussian Case with Known σ^2

- Assume $p(x|\mu) \sim N(\mu, \sigma^2)$ with known σ^2.
- Whatever prior knowledge we know about μ is expressed in $p(\mu)$, which is known.
- Indeed, we assume it took is a Gaussian

$$p(\mu) \sim N(\mu_0, \sigma_0^2) \ . \tag{26}$$

μ_0 represents our best guess of the value of the mean and σ_0^2 represents our uncertainty about this guess.

Univariate Gaussian Case with Known σ^2

- Assume $p(x|\mu) \sim N(\mu, \sigma^2)$ with known σ^2.
- Whatever prior knowledge we know about μ is expressed in $p(\mu)$, which is known.
- Indeed, we assume it took is a Gaussian

$$p(\mu) \sim N(\mu_0, \sigma_0^2) \ . \tag{26}$$

  μ_0 represents our best guess of the value of the mean and σ_0^2 represents our uncertainty about this guess.
- Note: the choice of the prior as a Gaussian is not so crucial. Rather, the assumption that **we know the prior is crucial**.

Univariate Gaussian Case with Known σ^2

Training samples

- We assume that we are given samples $\mathcal{D} = \{x_1, \dots, x_n\}$ from $p(x, \mu)$.
- Take some time to think through this point—unlike in MLE, we cannot assume that we have a single value of the parameter in the underlying distribution.

Univariate Gaussian Case with Known σ^2

Bayes Rule

- $$p(\mu|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\mu)p(\mu) \qquad (27)$$
  $$= \frac{1}{Z} \prod_k p(x_k|\mu)p(\mu) \qquad (28)$$
- See how the training samples modulate our prior knowledge of the parameters in the posterior?

Univariate Gaussian Case with Known σ^2

Expanding...

- $$p(\mu|\mathcal{D}) = \frac{1}{Z} \prod_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right] \quad (29)$$

- After some manipulation, we can see that $p(\mu|\mathcal{D})$ is an exponential function of a quadratic of μ, which is another way of saying a normal density.

$$p(\mu|\mathcal{D}) = \frac{1}{Z'} \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_k x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right] \quad (30)$$

Univariate Gaussian Case with Known σ^2

Names of these convenient distributions...

- And, this will be true regardless of the number of training samples.
- In other words, $p(\mu|\mathcal{D})$ remains a normal as the number of samples increases.
- Hence, $p(\mu|\mathcal{D})$ is said to be a **reproducing density**.
- $p(\mu)$ is said to be a **conjugate prior**.

Univariate Gaussian Case with Known σ^2

Rewriting...

- We can write $p(\mu|\mathcal{D}) \sim N(\mu_n, \sigma_n^2)$. Then, we have

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right] \tag{31}$$

- The new coefficients are

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{32}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\overline{\mu}_n + \frac{\mu_0}{\sigma_0^2} \tag{33}$$

- $\overline{\mu}_n$ is the sample mean over the n samples.

Univariate Gaussian Case with Known σ^2

Rewriting...

- Solving explicitly for μ_n and σ_n^2

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\overline{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \tag{34}$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{35}$$

  shows explicitly how the prior information is combined with the training samples **to estimate the parameters of the posterior distribution**.

- After n samples, μ_n is our best guess for the mean of the posterior and σ_n^2 is our uncertainty about it.

Univariate Gaussian Case with Known σ^2

Uncertainty...

- What can we say about this uncertainty as n increases?

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

Univariate Gaussian Case with Known σ^2

Uncertainty...

- What can we say about this uncertainty as n increases?

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

- That each observation **monotonically decreases our uncertainty** about the distribution.

$$\lim_{n \to \infty} \sigma_n^2 = 0 \tag{36}$$

- In other terms, as n increases, $p(\mu|\mathcal{D})$ becomes more and more sharply peaked approaching a Dirac delta function.

Univariate Gaussian Case with Known σ^2

- What can we say about the parameter μ_n as n increases?

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

Univariate Gaussian Case with Known σ^2

- What can we say about the parameter μ_n as n increases?

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

- It is a convex combination between the sample mean $\overline{\mu}_n$ (from the observed data) and the prior μ_0.
- Thus, it always lives somewhere betweens $\overline{\mu}_n$ and μ_0.
- And, it approaches the sample mean as n approaches ∞:

$$\lim_{n \to \infty} \mu_n = \overline{\mu}_n \equiv \frac{1}{n} \sum_{k=1}^{n} x_k \tag{37}$$

Univariate Gaussian Case with Known σ^2

Putting it all together to obtain $p(x|\mathcal{D})$.

- Our goal has been to obtain an estimate of how likely a novel sample x is given the entire training set $\mathcal{D}$: $p(x|\mathcal{D})$.

$$p(x|\mathcal{D}) = \int p(x|\mu)p(\mu|\mathcal{D})d\mu \tag{38}$$

$$\begin{aligned} &= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ &\quad \int \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] \end{aligned} \tag{39}$$

$$= \frac{1}{2\pi\sigma\sigma_n} \exp\left[\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma,\sigma_n) \tag{40}$$

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

Some Comparisons

Maximum Likelihood

- Point Estimator

$$p(x|\mathcal{D}) = p(x|\hat{\boldsymbol{\theta}})$$

- Parameter Estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathcal{D}|\boldsymbol{\theta})$$

Some Comparisons

Maximum Likelihood

- Point Estimator

$$p(x|\mathcal{D}) = p(x|\hat{\boldsymbol{\theta}})$$

- Parameter Estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathcal{D}|\boldsymbol{\theta})$$

Bayesian

- Distribution Estimator

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

- Distribution Estimate

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z}p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Some Comparisons

- So, is the Bayesian approach like Maximum Likelihood with a prior?

Some Comparisons

- So, is the Bayesian approach like Maximum Likelihood with a prior?
- **NO!**

Maximum Posterior

- Point Estimator

$$p(x|\mathcal{D}) = p(x|\hat{\boldsymbol{\theta}})$$

- Parameter Estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Bayesian

- Distribution Estimator

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

- Distribution Estimate

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z}p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Some Comparisons

Comments on the two methods

- For reasonable priors, MLE and BPE are equivalent in the asymptotic limit of infinite training data.
- **Computationally** – MLE methods are preferred for computational reasons because they are comparatively simpler (differential calculus versus multidimensional integration).

Some Comparisons

Comments on the two methods

- For reasonable priors, MLE and BPE are equivalent in the asymptotic limit of infinite training data.
- **Computationally** – MLE methods are preferred for computational reasons because they are comparatively simpler (differential calculus versus multidimensional integration).
- **Interpretability** – MLE methods are often more readily interpretted because they give a single point answer whereas BPE methods give a distribution over answers which can be more complicated.

Some Comparisons

Comments on the two methods

- For reasonable priors, MLE and BPE are equivalent in the asymptotic limit of infinite training data.
- **Computationally** – MLE methods are preferred for computational reasons because they are comparatively simpler (differential calculus versus multidimensional integration).
- **Interpretability** – MLE methods are often more readily interpretted because they give a single point answer whereas BPE methods give a distribution over answers which can be more complicated.
- **Confidence In Priors** – But, the Bayesian methods bring more information to the table. If the underlying distribution is of a different parametric form than originally assumed, Bayesian methods will do better.

Some Comparisons

Comments on the two methods

- For reasonable priors, MLE and BPE are equivalent in the asymptotic limit of infinite training data.
- **Computationally** – MLE methods are preferred for computational reasons because they are comparatively simpler (differential calculus versus multidimensional integration).
- **Interpretability** – MLE methods are often more readily interpretted because they give a single point answer whereas BPE methods give a distribution over answers which can be more complicated.
- **Confidence In Priors** – But, the Bayesian methods bring more information to the table. If the underlying distribution is of a different parametric form than originally assumed, Bayesian methods will do better.
- **Bias-Variance** – Bayesian methods make the bias-variance tradeoff more explicit by directly incorporating the uncertainty in the estimates.

Some Comparisons

Comments on the two methods

Take Home Message

There are strong theoretical and methodological arguments supporting Bayesian estimation, though in practice maximum-likelihood estimation is simpler, and when used for designing classifiers, can lead to classifiers that are nearly as accurate.

Recursive Bayesian Estimation

- Another reason to prefer Bayesian estimation is that it provides a natural way to incorporate additional training data as it becomes available.
- Let a training set with n samples be denoted $\mathcal{D}^n$.

Recursive Bayesian Estimation

- Another reason to prefer Bayesian estimation is that it provides a natural way to incorporate additional training data as it becomes available.
- Let a training set with n samples be denoted $\mathcal{D}^n$.
- Then, due to our independence assumption:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}) \tag{41}$$

we have

$$p(\mathcal{D}^n|\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})p(\mathcal{D}^{n-1}|\boldsymbol{\theta}) \tag{42}$$

Recursive Bayesian Estimation

- And, with Bayes Formula, we see that the posterior satisfies the recursion

$$p(\boldsymbol{\theta}|\mathcal{D}^n) = \frac{1}{Z} p(\mathbf{x}_n|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}^{n-1}) \ . \tag{43}$$

- This is an instance of **on-line learning**.

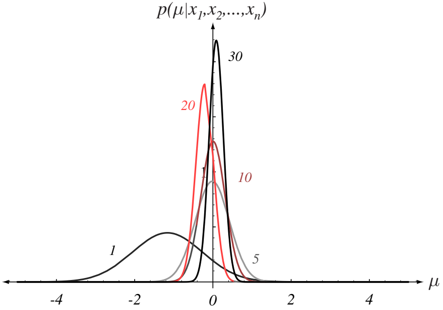
Recursive Bayesian Estimation

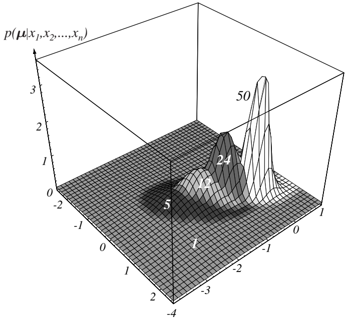
- And, with Bayes Formula, we see that the posterior satisfies the recursion

$$p(\boldsymbol{\theta}|\mathcal{D}^n) = \frac{1}{Z} p(\mathrm{x}_n|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}^{n-1}) \ . \tag{43}$$

- This is an instance of **on-line learning**.
- In principle, this derivation requires that we retain the entire training set in $\mathcal{D}^{n-1}$ to calculate $p(\boldsymbol{\theta}|\mathcal{D}^n)$. But, for some distributions, we can simply retain the sufficient statistics, which contain all the information needed.

Recursive Bayesian Estimation





Example of Recursive Bayes

- Suppose we believe our samples come from a uniform distribution:

$$p(x|\theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

- Initially, we know only that our parameter θ is bounded by 10, i.e., $0 \le \theta \le 10$.

Example of Recursive Bayes

- Suppose we believe our samples come from a uniform distribution:

$$p(x|\theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

- Initially, we know only that our parameter θ is bounded by 10, i.e., $0 \le \theta \le 10$.
- Before any data arrive, we have

$$p(\theta|\mathcal{D}^0) = p(\theta) = U(0,10) \ . \tag{45}$$

Example of Recursive Bayes

- Suppose we believe our samples come from a uniform distribution:

$$p(x|\theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

- Initially, we know only that our parameter θ is bounded by 10, i.e., $0 \le \theta \le 10$.
- Before any data arrive, we have

$$p(\theta|\mathcal{D}^0) = p(\theta) = U(0,10) \ . \tag{45}$$

- We get a training data set $\mathcal{D} = \{4, 7, 2, 8\}$.

Example of Recursive Bayes

- When the first data point arrives, $x_1 = 4$, we get an improved estimate of θ:

$$p(\theta|\mathcal{D}^1) \propto p(x|\theta)p(\theta|\mathcal{D}^0) = \begin{cases} 1/\theta & 4 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases} \tag{46}$$

Example of Recursive Bayes

- When the first data point arrives, $x_1 = 4$, we get an improved estimate of θ:

$$p(\theta|\mathcal{D}^1) \propto p(x|\theta)p(\theta|\mathcal{D}^0) = \begin{cases} 1/\theta & 4 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases} \tag{46}$$

- When the next data point arrives, $x_2 = 7$, we have

$$p(\theta|\mathcal{D}^2) \propto p(x|\theta)p(\theta|\mathcal{D}^1) = \begin{cases} 1/\theta^2 & 7 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

Example of Recursive Bayes

- When the first data point arrives, $x_1 = 4$, we get an improved estimate of θ:

$$p(\theta|\mathcal{D}^1) \propto p(x|\theta)p(\theta|\mathcal{D}^0) = \begin{cases} 1/\theta & 4 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases} \tag{46}$$

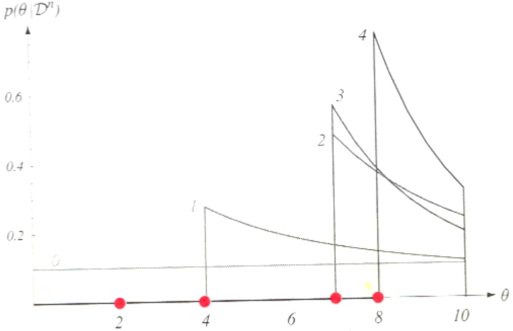
- When the next data point arrives, $x_2 = 7$, we have

$$p(\theta|\mathcal{D}^2) \propto p(x|\theta)p(\theta|\mathcal{D}^1) = \begin{cases} 1/\theta^2 & 7 \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

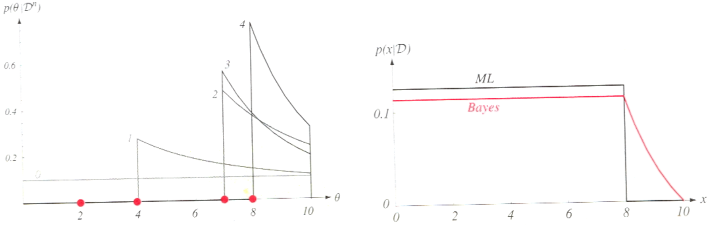
- And so on....

Example of Recursive Bayes

- Notice that each successive data sample introduces a factor of $1/\theta$ into $p(x|\theta)$.
- The distribution of samples is nonzero only for x values above the max, $p(\theta|\mathcal{D}^n) \propto 1/\theta^n$ for $\max_x[\mathcal{D}^n] \leq \theta \leq 10$.
- Our distribution is



Example of Recursive Bayes



- The maximum likelihood solution is $\hat{\theta} = 8$, implying $p(x|\mathcal{D}) \sim U(0, 8)$.
- But, the Bayesian solution shows a different character:
  - Starts out flat.
  - As more points are added, it becomes increasingly peaked at the value of the highest data point.
  - And, the Bayesian estimate has a tail for points above 8 reflecting our prior distribution.