Parametric Models
Part I: Maximum Likelihood and Bayesian Density Estimation

Selim Aksoy

Department of Computer Engineering
Bilkent University
saksoy@cs.bilkent.edu.tr

CS 551, Spring 2009
Bayesian Decision Theory shows us how to design an optimal classifier if we know the prior probabilities $P(w_i)$ and the class-conditional densities $p(x|w_i)$.

Unfortunately, we rarely have complete knowledge of the probabilistic structure.

However, we can often find design samples or training data that include particular representatives of the patterns we want to classify.
To simplify the problem, we can assume some parametric form for the conditional densities and estimate these parameters using training data.

Then, we can use the resulting estimates as if they were the true values and perform classification using the Bayesian decision rule.

We will consider only the supervised learning case where the true class label for each sample is known.
We will study two estimation procedures:

- **Maximum likelihood estimation**
  - Views the parameters as quantities whose values are fixed but unknown.
  - Estimates these values by maximizing the probability of obtaining the samples observed.

- **Bayesian estimation**
  - Views the parameters as random variables having some known prior distribution.
  - Observing new samples converts the prior to a posterior density.
Maximum Likelihood Estimation

- Suppose we have a set \( \mathcal{D} = \{x_1, \ldots, x_n\} \) of independent and identically distributed (i.i.d.) samples drawn from the density \( p(x|\theta) \).
- We would like to use training samples in \( \mathcal{D} \) to estimate the unknown parameter vector \( \theta \).
- Define \( L(\theta|\mathcal{D}) \) as the \textit{likelihood function} of \( \theta \) with respect to \( \mathcal{D} \) as

\[
L(\theta|\mathcal{D}) = p(\mathcal{D}|\theta) = p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta).
\]
The maximum likelihood estimate (MLE) of $\theta$ is, by definition, the value $\hat{\theta}$ that maximizes $L(\theta|D)$ and can be computed as

$$\hat{\theta} = \arg \max_\theta L(\theta|D).$$

It is often easier to work with the logarithm of the likelihood function (log-likelihood function) that gives

$$\hat{\theta} = \arg \max_\theta \log L(\theta|D) = \arg \max_\theta \sum_{i=1}^n \log p(x_i|\theta).$$
If the number of parameters is $p$, i.e.,

$$\theta = (\theta_1, \ldots, \theta_p)^T,$$

define the gradient operator

$$\nabla_\theta \equiv \begin{bmatrix}
\frac{\partial}{\partial \theta_1} \\
\vdots \\
\frac{\partial}{\partial \theta_p}
\end{bmatrix}.$$

Then, the MLE of $\theta$ should satisfy the necessary conditions

$$\nabla_\theta \log L(\theta|\mathcal{D}) = \sum_{i=1}^{n} \nabla_\theta \log p(x_i|\theta) = 0.$$
Properties of MLEs:

- The MLE is the parameter point for which the observed sample is the most likely.
- The procedure with partial derivatives may result in several local extrema. We should check each solution individually to identify the global optimum.
- Boundary conditions must also be checked separately for extrema.
- Invariance property: if \( \hat{\theta} \) is the MLE of \( \theta \), then for any function \( f(\theta) \), the MLE of \( f(\theta) \) is \( f(\hat{\theta}) \).
The Gaussian Case

▶ Suppose that $p(x|\theta) = N(\mu, \Sigma)$.
  
  ▶ When $\Sigma$ is known but $\mu$ is unknown:
    
    $$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  ▶ When both $\mu$ and $\Sigma$ are unknown:
    
    $$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$
The Bernoulli Case

- Suppose that $P(x|\theta) = \text{Bernoulli}(\theta) = \theta^x(1 - \theta)^{1-x}$ where $x = 0, 1$ and $0 \leq \theta \leq 1$.

- The MLE of $\theta$ can be computed as

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$
Bias of Estimators

- **Bias** of an estimator \( \hat{\theta} \) is the difference between the expected value of \( \hat{\theta} \) and \( \theta \).

- The MLE of \( \mu \) is an unbiased estimator for \( \mu \) because 
  \[ E[\hat{\mu}] = \mu. \]

- The MLE of \( \Sigma \) is not an unbiased estimator for \( \Sigma \) because 
  \[ E[\hat{\Sigma}] = \frac{n-1}{n} \Sigma \neq \Sigma. \]

- The sample covariance 
  \[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \]

  is an unbiased estimator for \( \Sigma \).
To measure how well a fitted distribution resembles the sample data (goodness-of-fit), we can use the Kolmogorov-Smirnov test statistic. It is defined as the maximum value of the absolute difference between the cumulative distribution function estimated from the sample and the one calculated from the fitted distribution. After estimating the parameters for different distributions, we can compute the Kolmogorov-Smirnov statistic for each distribution and choose the one with the smallest value as the best fit to our sample.
Maximum Likelihood Estimation Examples

(a) True pdf is $N(10, 4)$. Estimated pdf is $N(10.1, 3.9)$.

(b) True pdf is $0.5N(10, 0.16) + 0.5N(11, 0.25)$. Estimated pdf is $N(10.5, 0.5)$.

(c) True pdf is Gamma(4, 4). Estimated pdfs are $N(15.8, 62.1)$ and Gamma(4.0, 3.9).

(d) Cumulative distribution functions for the example in (c).

Figure 1: Histograms of samples and estimated densities for different distributions.
Suppose the set $\mathcal{D} = \{x_1, \ldots, x_n\}$ contains the samples drawn independently from the density $p(x|\theta)$ whose form is assumed to be known but $\theta$ is not known exactly.

Assume that $\theta$ is a quantity whose variation can be described by the prior probability distribution $p(\theta)$. 
Bayesian Estimation

- Given $\mathcal{D}$, the prior distribution can be updated to form the posterior distribution using the Bayes rule

\[
p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}
\]

where

\[
p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta) \, d\theta
\]

and

\[
p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(x_i|\theta).
\]
The posterior distribution \( p(\theta|D) \) can be used to find estimates for \( \theta \) (e.g., the expected value of \( p(\theta|D) \) can be used as an estimate for \( \theta \)).

Then, the conditional density \( p(x|D) \) can be computed as

\[
p(x|D) = \int p(x|\theta) p(\theta|D) \, d\theta
\]

and can be used in the Bayesian classifier.
MLEs vs. Bayes Estimates

- Maximum likelihood estimation finds an estimate of $\theta$ based on the samples in $\mathcal{D}$ but a different sample set would give rise to a different estimate.

- Bayes estimate takes into account the sampling variability.

- We assume that we do not know the true value of $\theta$, and instead of taking a single estimate, we take a weighted sum of the densities $p(x|\theta)$ weighted by the distribution $p(\theta|\mathcal{D})$. 
Consider the univariate case \( p(x|\mu) = N(\mu, \sigma^2) \) where \( \mu \) is the only unknown parameter with a prior distribution \( p(\mu) = N(\mu_0, \sigma^2_0) \) \((\sigma^2, \mu_0 \text{ and } \sigma^2_0 \text{ are all known})\).

This corresponds to drawing a value for \( \mu \) from the population with density \( p(\mu) \), treating it as the true value in the density \( p(x|\mu) \), and drawing samples for \( x \) from this density.
The Gaussian Case

Given $\mathcal{D} = \{x_1, \ldots, x_n\}$, we obtain

$$p(\mu | \mathcal{D}) \propto \prod_{i=1}^{n} p(x_i | \mu) p(\mu)$$

$$\propto \exp \left[ -\frac{1}{2} \left( \left( \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left( \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma_0^2} \right) \mu \right) \right]$$

$$= N(\mu_n, \sigma_n^2)$$

where

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

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

$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}.$$
The Gaussian Case

- \( \mu_0 \) is our best prior guess and \( \sigma^2_0 \) is the uncertainty about this guess.
- \( \mu_n \) is our best guess after observing \( D \) and \( \sigma^2_n \) is the uncertainty about this guess.
- \( \mu_n \) always lies between \( \hat{\mu}_n \) and \( \mu_0 \).
  - If \( \sigma_0 = 0 \), then \( \mu_n = \mu_0 \) (no observation can change our prior opinion).
  - If \( \sigma_0 \gg \sigma \), then \( \mu_n = \hat{\mu}_n \) (we are very uncertain about our prior guess).
  - Otherwise, \( \mu_n \) approaches \( \hat{\mu}_n \) as \( n \) approaches infinity.
Given the posterior density \( p(\mu|D) \), the conditional density \( p(x|D) \) can be computed as

\[
p(x|D) = N(\mu_n, \sigma^2 + \sigma_n^2)
\]

where the conditional mean \( \mu_n \) is treated as if it were the true mean, and the known variance is increased to account for our lack of exact knowledge of the mean \( \mu \).
The Gaussian Case

Consider the multivariate case $p(x | \mu) = N(\mu, \Sigma)$ where $\mu$ is the only unknown parameter with a prior distribution $p(\mu) = N(\mu_0, \Sigma_0)$ (\Sigma, \mu_0 and \Sigma_0 are all known).

Given $D = \{x_1, \ldots, x_n\}$, we obtain

$$p(\mu | D) \propto \exp \left[ -\frac{1}{2} \left( \mu^T \left( n\Sigma^{-1} + \Sigma_0^{-1} \right) \mu ight. ight.$$

$$\left. - 2\mu^T \left( \Sigma^{-1} \sum_{i=1}^{n} x_i + \Sigma_0^{-1} \mu_0 \right) \right].$$
The Gaussian Case

- It follows that

\[ p(\mu|D) = N(\mu_n, \Sigma_n) \]

where

\[ \mu_n = \Sigma_0 \left( \Sigma_0 + \frac{1}{n} \Sigma \right)^{-1} \hat{\mu}_n + \frac{1}{n} \Sigma \left( \Sigma_0 + \frac{1}{n} \Sigma \right)^{-1} \mu_0, \]

\[ \Sigma_n = \frac{1}{n} \Sigma_0 \left( \Sigma_0 + \frac{1}{n} \Sigma \right)^{-1} \Sigma. \]
The Gaussian Case

Given the posterior density $p(\mu | D)$, the conditional density $p(x | D)$ can be computed as

$$p(x | D) = \mathcal{N}(\mu_n, \Sigma + \Sigma_n)$$

which can be viewed as the sum of a random vector $\mu$ with $p(\mu | D) = \mathcal{N}(\mu_n, \Sigma_n)$ and an independent random vector $y$ with $p(y) = \mathcal{N}(0, \Sigma)$. 
Consider $P(x|\theta) = \text{Bernoulli}(\theta)$ where $\theta$ is the unknown parameter with a prior distribution $p(\theta) = \text{Beta}(\alpha, \beta)$ ($\alpha$ and $\beta$ are both known).

Given $\mathcal{D} = \{x_1, \ldots, x_n\}$, we obtain

$$p(\theta|\mathcal{D}) = \text{Beta}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right).$$
The Bayes estimate of $\theta$ can be computed as the expected value of $p(\theta|D)$, i.e.,

$$\hat{\theta} = \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \beta + n}$$

$$= \left( \frac{n}{\alpha + \beta + n} \right) \frac{1}{n} \sum_{i=1}^{n} x_i + \left( \frac{\alpha + \beta}{\alpha + \beta + n} \right) \frac{\alpha}{\alpha + \beta}.$$
A conjugate prior is one which, when multiplied with the probability of the observation, gives a posterior probability having the same functional form as the prior.

This relationship allows the posterior to be used as a prior in further computations.

Table 1: Conjugate prior distributions.

<table>
<thead>
<tr>
<th>pdf generating the sample</th>
<th>corresponding conjugate prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>Gaussian</td>
</tr>
<tr>
<td>Exponential</td>
<td>Gamma</td>
</tr>
<tr>
<td>Poisson</td>
<td>Gamma</td>
</tr>
<tr>
<td>Binomial</td>
<td>Beta</td>
</tr>
<tr>
<td>Multinomial</td>
<td>Dirichlet</td>
</tr>
</tbody>
</table>
What about the convergence of \( p(x|\mathcal{D}) \) to \( p(x) \)?

Given \( \mathcal{D}^n = \{x_1, \ldots, x_n\} \), for \( n > 1 \)

\[
p(\mathcal{D}^n|\theta) = p(x_n|\theta)p(\mathcal{D}^{n-1}|\theta)
\]

and

\[
p(\theta|\mathcal{D}^n) = \frac{p(x_n|\theta)p(\theta|\mathcal{D}^{n-1})}{\int p(x_n|\theta)p(\theta|\mathcal{D}^{n-1}) \, d\theta}
\]

where

\[
p(\theta|\mathcal{D}^0) = p(\theta)
\]

⇒ quite useful if the distributions can be represented using only a few parameters (sufficient statistics).
Consider the Bernoulli case $P(x|\theta) = \text{Bernoulli}(\theta)$ where $p(\theta) = \text{Beta}(\alpha, \beta)$, the Bayes estimate of $\theta$ is

$$\hat{\theta} = \frac{\alpha}{\alpha + \beta}.$$ 

Given the training set $\mathcal{D} = \{x_1, \ldots, x_n\}$, we obtain

$$p(\theta|\mathcal{D}) = \text{Beta}(\alpha + m, \beta + n - m)$$

where $m = \sum_{i=1}^{n} x_i = \#\{x_i| x_i = 1, x_i \in \mathcal{D}\}$. 

Recursive Bayes Learning
The Bayes estimate of $\theta$ becomes

$$\hat{\theta} = \frac{\alpha + m}{\alpha + \beta + n}.$$

Then, given a new training set $\mathcal{D}' = \{x_1, \ldots, x_{n'}\}$, we obtain

$$p(\theta|\mathcal{D}, \mathcal{D}') = \text{Beta}(\alpha + m + m', \beta + n - m + n' - m')$$

where $m' = \sum_{i=1}^{n'} x_i = \#\{x_i|x_i = 1, x_i \in \mathcal{D}'\}$. 
The Bayes estimate of $\theta$ becomes

$$\hat{\theta} = \frac{\alpha + m + m'}{\alpha + \beta + n + n'}.$$ 

Thus, recursive Bayes learning involves only keeping the counts $m$ (related to sufficient statistics of Beta) and the number of training samples $n$. 
MLEs vs. Bayes Estimates

Table 2: Comparison of MLEs and Bayes estimates.

<table>
<thead>
<tr>
<th></th>
<th>MLE</th>
<th>Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>computational</td>
<td>differential calculus, gradient search</td>
<td>multidimensional integration</td>
</tr>
<tr>
<td>complexity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>interpretability</td>
<td>point estimate</td>
<td>weighted average of models</td>
</tr>
<tr>
<td>prior information</td>
<td>assume the parametric model $p(x</td>
<td>\theta)$</td>
</tr>
</tbody>
</table>

- If there is much data (strongly peaked $p(\theta|\mathcal{D})$) and the prior $p(\theta)$ is uniform, then the Bayes estimate and MLE are equivalent.
Classification Error

- To apply these results to multiple classes, separate the training samples to $c$ subsets $\mathcal{D}_1, \ldots, \mathcal{D}_c$, with the samples in $\mathcal{D}_i$ belonging to class $w_i$, and then estimate each density $p(x|w_i, \mathcal{D}_i)$ separately.

- Different sources of error:
  - Bayes error: due to overlapping class-conditional densities (related to the features used).
  - Model error: due to incorrect model.
  - Estimation error: due to estimation from a finite sample (can be reduced by increasing the amount of training data).