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 2008

Introduction

- 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 $p(\mathbf{x}|w_i)$.
- 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.

Introduction

- 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.

Introduction

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

$$L(\boldsymbol{\theta}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{x_1}, \ldots, \mathbf{x_n}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x_i}|\boldsymbol{\theta}).$$

Maximum Likelihood Estimation

- The *maximum likelihood estimate* (MLE) of $\boldsymbol{\theta}$ is, by definition, the value $\hat{\boldsymbol{\theta}}$ that maximizes $L(\boldsymbol{\theta}|\mathcal{D})$ and can be computed as

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

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

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\mathcal{D}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}).$$

Maximum Likelihood Estimation

- If the number of parameters is p, i.e., $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_p})^T$, define the gradient operator

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

- Then, the MLE of $\boldsymbol{\theta}$ should satisfy the necessary conditions

$$\nabla_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}) = 0.$$

Maximum Likelihood Estimation

- 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 $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then for any function $f(\boldsymbol{\theta})$, the MLE of $f(\boldsymbol{\theta})$ is $f(\hat{\boldsymbol{\theta}})$.

The Gaussian Case

- Suppose that $p(\mathbf{x}|\boldsymbol{\theta}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  - When $\boldsymbol{\Sigma}$ is known but $\boldsymbol{\mu}$ is unknown:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_{\boldsymbol{i}}$$

  - When both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_{\boldsymbol{i}} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}})(\mathbf{x}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}})^T$$

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 θ 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 θ.
- The MLE of $\boldsymbol{\mu}$ is an unbiased estimator for $\boldsymbol{\mu}$ because $E[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$.
- The MLE of Σ is not an unbiased estimator for Σ because $E[\hat{\Sigma}] = \frac{n-1}{n}\Sigma \neq \Sigma$.
- The *sample covariance*

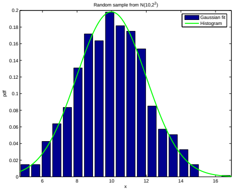
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

is an unbiased estimator for Σ.

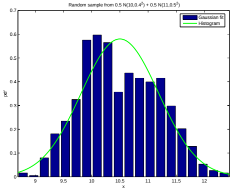
Goodness-of-fit

- 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.

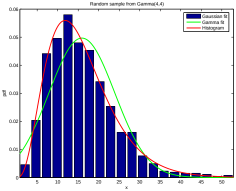
Maximum Likelihood Estimation Examples



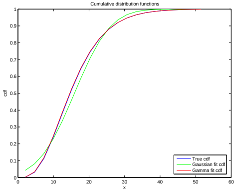
(a) True pdf is $N(10, 4)$. Estimated pdf is $N(10.1, 3.9)$.



(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.

Bayesian Estimation

- Suppose the set $\mathcal{D} = \{\mathbf{x_1}, \ldots, \mathbf{x}_{\boldsymbol{n}}\}$ contains the samples drawn independently from the density $p(\mathbf{x}|\boldsymbol{\theta})$ whose form is assumed to be known but $\boldsymbol{\theta}$ is not known exactly.
- Assume that $\boldsymbol{\theta}$ is a quantity whose variation can be described by the prior probability distribution $p(\boldsymbol{\theta})$.

Bayesian Estimation

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

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

where

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

and

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

Bayesian Estimation

- The posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ can be used to find estimates for $\boldsymbol{\theta}$ (e.g., the expected value of $p(\boldsymbol{\theta}|\mathcal{D})$ can be used as an estimate for $\boldsymbol{\theta}$).
- Then, the conditional density $p(\mathbf{x}|\mathcal{D})$ can be computed as

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

and can be used in the Bayesian classifier.

MLEs vs. Bayes Estimates

- Maximum likelihood estimation finds an estimate of $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$, and instead of taking a single estimate, we take a weighted sum of the densities $p(\mathbf{x}|\boldsymbol{\theta})$ weighted by the distribution $p(\boldsymbol{\theta}|\mathcal{D})$.

The Gaussian Case

- Consider the univariate case $p(x|\mu) = N(\mu, \sigma^2)$ where μ is the only unknown parameter with a prior distribution $p(\mu) = N(\mu_0, \sigma_0^2)$ $\quad$ (σ^2, μ_0 and σ_0^2 are all known).
- This corresponds to drawing a value for μ 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

$$
\begin{aligned}
p(\mu|\mathcal{D}) &\propto \prod_{i=1}^{n} p(x_i|\mu)p(\mu) \\
&\propto \exp\left[-\frac{1}{2}\left(\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma_0^2} \right)\mu \right) \right] \\
&= N(\mu_n, \sigma_n^2)
\end{aligned}
$$

where

$$
\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right)\hat{\mu}_n + \left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \right)\mu_0 \qquad \left(\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \right)
$$

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

The Gaussian Case

- μ_0 is our best prior guess and σ_0^2 is the uncertainty about this guess.
- μ_n is our best guess after observing $\mathcal{D}$ and σ_n^2 is the uncertainty about this guess.
- μ_n always lies between $\hat{\mu}_n$ and μ_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, μ_n approaches $\hat{\mu}_n$ as n approaches infinity.

The Gaussian Case

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

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

where the conditional mean μ_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 μ.

The Gaussian Case

- Consider the multivariate case $p(\mathbf{x}|\boldsymbol{\mu}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is the only unknown parameter with a prior distribution $p(\boldsymbol{\mu}) = N(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$ ($\boldsymbol{\Sigma}$, $\boldsymbol{\mu_0}$ and $\boldsymbol{\Sigma_0}$ are all known).
- Given $\mathcal{D} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, we obtain

$$p(\boldsymbol{\mu}|\mathcal{D}) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^T\left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma_0}^{-1}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{x}_{\boldsymbol{i}} + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0}\right)\right)\right].$$

The Gaussian Case

- It follows that

$$p(\boldsymbol{\mu}|\mathcal{D}) = N(\boldsymbol{\mu_n}, \boldsymbol{\Sigma_n})$$

where

$$\boldsymbol{\mu_n} = \boldsymbol{\Sigma_0}\left(\boldsymbol{\Sigma_0} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\mu}}_{\boldsymbol{n}} + \frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma_0} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu_0},$$

$$\boldsymbol{\Sigma_n} = \frac{1}{n}\boldsymbol{\Sigma_0}\left(\boldsymbol{\Sigma_0} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}.$$

The Gaussian Case

- Given the posterior density $p(\boldsymbol{\mu}|\mathcal{D})$, the conditional density $p(\mathbf{x}|\mathcal{D})$ can be computed as

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

  which can be viewed as the sum of a random vector $\boldsymbol{\mu}$ with $p(\boldsymbol{\mu}|\mathcal{D}) = N(\boldsymbol{\mu_n}, \boldsymbol{\Sigma_n})$ and an independent random vector $\mathbf{y}$ with $p(\mathbf{y}) = N(0, \boldsymbol{\Sigma})$.

The Bernoulli Case

- Consider $P(x|\theta) = \text{Bernoulli}(\theta)$ where θ is the unknown parameter with a prior distribution $p(\theta) = \text{Beta}(\alpha, \beta)$ (α and β 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 Bernoulli Case

- The Bayes estimate of θ can be computed as the expected value of $p(\theta|\mathcal{D})$, i.e.,

$$\begin{aligned}\hat{\theta} &= \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \beta + n} \\ &= \left(\frac{n}{\alpha + \beta + n}\right) \frac{1}{n} \sum_{i=1}^{n} x_i + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \frac{\alpha}{\alpha + \beta}.\end{aligned}$$

Conjugate Priors

- 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.

pdf generating the sample	*corresponding conjugate prior*
Gaussian	Gaussian
Exponential	Gamma
Poisson	Gamma
Binomial	Beta
Multinomial	Dirichlet

Recursive Bayes Learning

- What about the convergence of $p(\mathbf{x}|\mathcal{D})$ to $p(\mathbf{x})$?
- Given $\mathcal{D}^n = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, for $n > 1$

$$p(\mathcal{D}^n|\boldsymbol{\theta}) = p(\mathbf{x_n}|\boldsymbol{\theta})p(\mathcal{D}^{n-1}|\boldsymbol{\theta})$$

and

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

where

$$p(\boldsymbol{\theta}|\mathcal{D}^0) = p(\boldsymbol{\theta})$$

$\Rightarrow$ quite useful if the distributions can be represented using only a few parameters (*sufficient statistics*).

Recursive Bayes Learning

- Consider the Bernoulli case $P(x|\theta) = \text{Bernoulli}(\theta)$ where $p(\theta) = \text{Beta}(\alpha, \beta)$, the Bayes estimate of θ 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 θ 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 $m' = \sum_{i=1}^{n'} x_i = \#\{x_i | x_i = 1, x_i \in \mathcal{D}'\}$.

Recursive Bayes Learning

- The Bayes estimate of θ 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.

	MLE	*Bayes*
computational complexity	differential calculus, gradient search	multidimensional integration
interpretability	point estimate	weighted average of models
prior information	assume the parametric model $p(\mathbf{x}\|\boldsymbol{\theta})$	assume the models $p(\boldsymbol{\theta})$ and $p(\mathbf{x}\|\boldsymbol{\theta})$ but the resulting distribution $p(\mathbf{x}\|\mathcal{D})$ may not have the same form as $p(\mathbf{x}\|\boldsymbol{\theta})$

- If there is much data (strongly peaked $p(\boldsymbol{\theta}|\mathcal{D})$) and the prior $p(\boldsymbol{\theta})$ 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 $p(\mathbf{x}|w_i, \mathcal{D}_i)$ 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).