Pattern Classification - Duda et al.

Chapters 3.2 and 3.3

Miguel A. Veganzones

Grupo de Inteligencia Computacional
Universidad del País Vasco

2012-02-17

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

Learning from samples

- We could design an optimal classifier if we knew the prior probabilities $P(\omega_i)$ and the class-conditional densities $p(\mathbf{x}|\omega_i)$.
- We rarely have this kind of complete knowledge about the probabilistic structure of the problem.
  - Unknown distributions.
  - Samples.
- The problem, then, is to find some way to use this information to design or train the classifier.

Generative approach

- One approach to this problem is to use the samples to estimate the unknown probabilities and probability densities, and to use the resulting estimates as if they were the true values.
- Making assumptions about the distributions makes the problem easy:
  - Assuming $p(\mathbf{x}|\omega_i)$ follows a Gaussian distribution with mean μ and covariance Σ.
  - We simplify the problem from estimating an unknown function $p(\mathbf{x}|\omega_i)$ to estimate the unkown parameters μ and Σ.

Parameter estimation

- *Maximum Likelihood* (ML) views the parameters as quantities whose values are fixed but unknown.
  - The best estimate of their value is defined to be the one that maximizes the probability of obtaining the samples actually observed.
- Bayesian methods view the parameters as random variables having some known a priori distribution.
  - Observation of the samples converts this to a posterior density, thereby revising our opinion about the true values of the parameters.
  - *Maximum A Posteriori* (MAP).

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

Problem definition

Suppose that we separate a collection of samples according to class, so that we have c sets, $\mathcal{D}_1, ..., \mathcal{D}_c$, with the samples in $\mathcal{D}_j$ having been drawn independently according to the probability law $p(\mathbf{x}|\omega_j)$. We say such samples are *i.i.d.* — independent identically distributed random variables. We assume that $p(\mathbf{x}|\omega_j)$ has a known parametric form, and is therefore determined uniquely by the value of a parameter vector $\boldsymbol{\theta}_j$. For example, we might have $p(\mathbf{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\theta}_j$ consists of the components of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$. To show the dependence of $p(\mathbf{x}|\omega_j)$ on $\boldsymbol{\theta}_j$ explicitly, we write $p(\mathbf{x}|\omega_j)$ as $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$. Our problem is to use the information provided by the training samples to obtain good estimates for the unknown parameter vectors $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_c$ associated with each category.

Problem simplification

To simplify treatment of this problem, we shall assume that samples in $\mathcal{D}_i$ give no information about $\boldsymbol{\theta}_j$ if $i \neq j$ — that is, we shall assume that the parameters for the different classes are functionally independent. This permits us to work with each class separately, and to simplify our notation by deleting indications of class distinctions. With this assumption we thus have c separate problems of the following form: Use a set $\mathcal{D}$ of training samples drawn independently from the probability density $p(\mathbf{x}|\boldsymbol{\theta})$ to estimate the unknown parameter vector $\boldsymbol{\theta}$.

Maximum Likelihood

- Suppose that $\mathscr{D}$ contains n samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ that were drawn independently.
- The likelihood $p(\mathscr{D}|\theta)$ of θ with respect to the set of samples $\mathscr{D}$ is given by:

$$p(\mathscr{D}|\theta) = \prod_{k=1}^{n} p(\mathbf{x}_k|\theta)$$

- The maximum likelihood estimate of θ is the value $\hat{\theta}$ that maximizes $p(\mathscr{D}|\theta)$.
- $\hat{\theta}$ corresponds to the value of θ that best agrees with or supports the actually observed training samples.

Log-likelihood

- For analytical purposes, it is usually easier to work with the logarithm of the likelihood than with the likelihood itself:

$$l(\theta) = \ln p(\mathscr{D}|\theta) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\theta)$$

- Since the logarithm is monotonically increasing, the $\hat{\theta}$ that maximizes the log-likelihood also maximizes the likelihood:

$$\hat{\theta} = \arg\max_{\theta} l(\theta) = \arg\max_{\theta} p(\mathscr{D}|\theta)$$

- If $p(\mathscr{D}|\theta)$ is a well behaved, differentiable function of θ, $\hat{\theta}$ can be found by the standard methods of differential calculus:

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln p(\mathbf{x}_k|\theta) = \mathbf{0}$$

Visual example

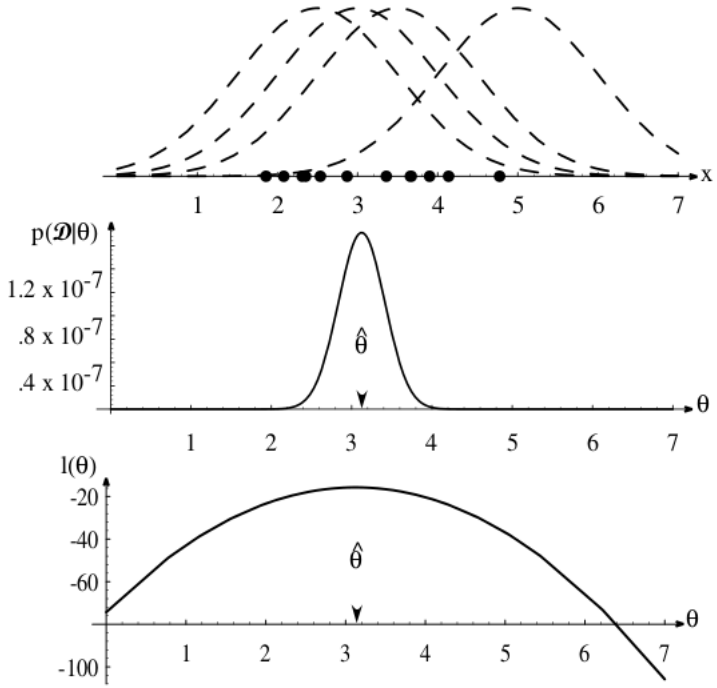

Figure 3.1: The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figures shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood — i.e., the log-likelihood $l(\theta)$, shown at the bottom. Note especially that the likelihood lies in a different space from $p(x|\hat{\theta})$, and the two can have different functional forms.

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

Gaussian log-likelihood

To see how maximum likelihood methods results apply to a specific case, suppose that the samples are drawn from a multivariate normal population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For simplicity, consider first the case where only the mean is unknown. Under this condition, we consider a sample point $\mathbf{x}_k$ and find

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \tag{8}$$

and

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}). \tag{9}$$

μ estimate

Identifying $\boldsymbol{\theta}$ with $\boldsymbol{\mu}$, we see from Eq. 9 that the maximum likelihood estimate for $\boldsymbol{\mu}$ must satisfy

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0}, \tag{10}$$

that is, each of the d components of $\hat{\boldsymbol{\mu}}$ must vanish. Multiplying by $\boldsymbol{\Sigma}$ and rearranging, we obtain

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k. \tag{11}$$

This is a very satisfying result. It says that the maximum likelihood estimate for the unknown population mean is just the arithmetic average of the training samples — the *sample mean*, sometimes written $\hat{\boldsymbol{\mu}}_n$ to clarify its dependence on the number of samples. Geometrically, if we think of the n samples as a cloud of points, the sample mean is the centroid of the cloud. The sample mean has a number of desirable statistical properties as well, and one would be inclined to use this rather obvious estimate even without knowing that it is the maximum likelihood solution.

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

Univariate case

In the more general (and more typical) multivariate normal case, neither the mean $\boldsymbol{\mu}$ nor the covariance matrix $\boldsymbol{\Sigma}$ is known. Thus, these unknown parameters constitute the components of the parameter vector $\boldsymbol{\theta}$. Consider first the univariate case with $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. Here the log-likelihood of a single point is

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2 \tag{12}$$

and its derivative is

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}. \tag{13}$$

Applying Eq. 7 to the full log-likelihood leads to the conditions

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \tag{14}$$

μ and Σ estimates for the univariate case

$$-\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2}+\sum_{k=1}^{n}\frac{(x_k-\hat{\theta}_1)^2}{\hat{\theta}_2^2}=0, \tag{15}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the maximum likelihood estimates for θ_1 and θ_2, respectively. By substituting $\hat{\mu}=\hat{\theta}_1$, $\hat{\sigma}^2=\hat{\theta}_2$ and doing a little rearranging, we obtain the following maximum likelihood estimates for μ and σ^2:

$$\hat{\mu}=\frac{1}{n}\sum_{k=1}^{n}x_k \tag{16}$$

and

$$\hat{\sigma}^2=\frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})^2. \tag{17}$$

μ and Σ estimates for the multivariate case

While the analysis of the multivariate case is basically very similar, considerably more manipulations are involved (Problem 6). Just as we would predict, though, the result is that the maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \tag{18}$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t. \tag{19}$$

Thus, once again we find that the maximum likelihood estimate for the mean vector is the sample mean. The maximum likelihood estimate for the covariance matrix is the arithmetic average of the n matrices $(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$. Since the true covariance matrix is the expected value of the matrix $(\mathbf{x} - \hat{\boldsymbol{\mu}})\ (\mathbf{x} - \hat{\boldsymbol{\mu}})^t$, this is also a very satisfying result.

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

σ estimate

The maximum likelihood estimate for the variance σ^2 is *biased*; that is, the expected value over all data sets of size n of the sample variance is not equal to the true variance:*

$$\mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2. \tag{20}$$

We shall return to a more general consideration of bias in Chap. **??**, but for the moment we can verify Eq. 20 for an underlying distribution with non-zero variance, σ^2, in the extreme case of $n = 1$, in which the expectation value $\mathcal{E}[\cdot] = 0 \neq \sigma^2$. The maximum likelihood estimate of the covariance matrix is similarly biased.

Unbiased estimators

Elementary *unbiased* estimators for σ^2 and $\mathbf{\Sigma}$ are given by

$$\mathcal{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\sigma^2 \qquad \text{and} \tag{21}$$

$$\mathbf{C}=\frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^t, \tag{22}$$

where $\mathbf{C}$ is the so-called *sample covariance matrix*, as explored in Problem 33. If an estimator is unbiased for *all* distributions, as for example the variance estimator in Eq. 21, then it is called *absolutely unbiased*. If the estimator tends to become unbiased as the number of samples becomes very large, as for instance Eq. 20, then the estimator is *asymptotically unbiased*. In many pattern recognition problems with large training data sets, asymptotically unbiased estimators are acceptable.

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

Bayes' formula

Given the sample $\mathcal{D}$, Bayes' formula then becomes

$$P(\omega_i|\mathbf{x},\mathcal{D}) = \frac{p(\mathbf{x}|\omega_i,\mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j,\mathcal{D})P(\omega_j|\mathcal{D})}. \quad (23)$$

As this equation suggests, we can use the information provided by the training samples to help determine both the class-conditional densities and the a priori probabilities.

Simplification

Although we could maintain this generality, we shall henceforth assume that the true values of the a priori probabilities are known or obtainable from a trivial calculation; thus we substitute $P(\omega_i) = P(\omega_i|\mathcal{D})$. Furthermore, since we are treating the supervised case, we can separate the training samples by class into c subsets $\mathcal{D}_1, ..., \mathcal{D}_c$, with the samples in $\mathcal{D}_i$ belonging to ω_i. As we mentioned when addressing maximum likelihood methods, in most cases of interest (and in all of the cases we shall consider), the samples in $\mathcal{D}_i$ have no influence on $p(\mathbf{x}|\omega_j, \mathcal{D})$ if $i \neq j$. This has two simplifying consequences. First, it allows us to work with each class separately, using only the samples in $\mathcal{D}_i$ to determine $p(\mathbf{x}|\omega_i, \mathcal{D})$. Used in conjunction with our assumption that the prior probabilities are known, this allows us to write Eq. 23 as

$$P(\omega_i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|\omega_i, \mathcal{D}_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \mathcal{D}_j)P(\omega_j)}. \tag{24}$$

Second, because each class can be treated independently, we can dispense with needless class distinctions and simplify our notation. In essence, we have c separate problems of the following form: use a set $\mathcal{D}$ of samples drawn independently according to the fixed but unknown probability distribution $p(\mathbf{x})$ to determine $p(\mathbf{x}|\mathcal{D})$. This is the central problem of Bayesian learning.

Outline

1. Introduction
2. Maximum Likelihood estimation
   - The general principle
   - The Gaussian case: unknown μ
   - The Gaussian case: unknown μ and Σ
   - Bias
3. Bayesian estimation
   - The class-conditional densities
   - The parameter distribution

Parametric form

Although the desired probability density $p(\mathbf{x})$ is unknown, we assume that it has a known parametric form. The only thing assumed unknown is the value of a parameter vector $\boldsymbol{\theta}$. We shall express the fact that $p(\mathbf{x})$ is unknown but has known parametric form by saying that the function $p(\mathbf{x}|\boldsymbol{\theta})$ is completely known. Any information we might have about $\boldsymbol{\theta}$ prior to observing the samples is assumed to be contained in a *known* prior density $p(\boldsymbol{\theta})$. Observation of the samples converts this to a posterior density $p(\boldsymbol{\theta}|\mathcal{D})$, which, we hope, is sharply peaked about the true value of $\boldsymbol{\theta}$.

Unupervised density estimation

Note that we are changing our supervised learning problem into an unsupervised density estimation problem. To this end, our basic goal is to compute $p(\mathbf{x}|\mathcal{D})$, which is as close as we can come to obtaining the unknown $p(\mathbf{x})$. We do this by integrating the joint density $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ over $\boldsymbol{\theta}$. That is,

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}, \tag{25}$$

where the integration extends over the entire parameter space. Now as discussed in Problem 12 we can write $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ as the product $p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta}|\mathcal{D})$. Since the selection of $\mathbf{x}$ and that of the training samples in $\mathcal{D}$ is done independently, the first factor is merely $p(\mathbf{x}|\boldsymbol{\theta})$. That is, the distribution of $\mathbf{x}$ is known completely once we know the value of the parameter vector. Thus, Eq. 25 can be rewritten as

$$\boxed{p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}.} \tag{26}$$

$p(\mathbf{x}|\mathscr{D})$ to $p(\boldsymbol{\theta}|\mathscr{D})$ link

This key equation links the desired class-conditional density $p(\mathbf{x}|\mathcal{D})$ to the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$ for the unknown parameter vector. If $p(\boldsymbol{\theta}|\mathcal{D})$ peaks very sharply about some value $\hat{\boldsymbol{\theta}}$, we obtain $p(\mathbf{x}|\mathcal{D}) \simeq p(\mathbf{x}|\hat{\boldsymbol{\theta}})$, i.e., the result we would obtain by substituting the estimate $\hat{\boldsymbol{\theta}}$ for the true parameter vector. This result rests on the assumption that $p(\mathbf{x}|\boldsymbol{\theta})$ is smooth, and that the tails of the integral are not important. These conditions are typically but not invariably the case, as we shall see in Sect. **??**. In general, if we are less certain about the exact value of $\boldsymbol{\theta}$, this equation directs us to average $p(\mathbf{x}|\boldsymbol{\theta})$ over the possible values of $\boldsymbol{\theta}$. Thus, when the unknown densities have a known parametric form, the samples exert their influence on $p(\mathbf{x}|\mathcal{D})$ through the posterior density $p(\boldsymbol{\theta}|\mathcal{D})$. We should also point out that in practice, the integration in Eq. 26 is often performed numerically, for instance by Monte-Carlo simulation.