8001652 Introduction to Pattern Recognition. Lecture 6: Maximum-likelihood parameter estimation

Jussi Tohka

jussi.tohka@tut.fi

Institute of Signal Processing

Tampere University of Technology

Motivation

- During last lectures we have seen how to design the optimal classifier provided that we know the class conditional probability densities $p(\mathbf{x}|\omega_i)$ and prior probabilities $P(\omega_i)$.
- Unfortunately, in pattern classification, these are rarely known beforehand.
- This means that we must learn them somehow.
- Recall from the first lecture the difference between unsupervised and supervised learning and classification. In supervised learning, a teacher was used to train the classifier and in the unsupervised case we had to cope without a teacher.

Supervised learning

- Our 'teacher' is now a set of (correctly labeled) *training samples* - samples of feature vectors from each category.
- We denote training samples from the category i by $\mathcal{D}_i$. Hence, we have c sets $\mathcal{D}_1, \ldots, \mathcal{D}_c$.
- How to obtain these samples? Usual method is to ask a human expert to assign feature vectors into categories. Of course, often it is easier for the expert to directly assign objects, and not measurements from them, to categories. But, this does not cause any problems.

Supervised learning

- Now our problem is as follows: Learn $p(\mathbf{x}|\omega_i), P(\omega_i)$ from training samples in $\mathcal{D}_1, \ldots, \mathcal{D}_c$.
- If we have collected the objects for the expert to label at random and if there are enough training samples, prior probabilities are easy to estimate. Just set $P(\omega_i) = |\mathcal{D}_i|/(\sum_{k=1}^{c} |\mathcal{D}_k|)$ - the number of training samples in category i divided by the total number of training samples.
- Otherwise, it may be wise to set each category equally probable.

Class conditional probabilities

- The class conditional probabilities $p(\mathbf{x}|\omega_i)$ are more tricky.
- There, the number of training samples seems always too small.
- At this point, we will solve this problem by parameterizing the probability densities.
- We will assume that each $p(\mathbf{x}|\omega_i)$ has a known parametric form. We write $p(\mathbf{x}|\omega_i) = p(\mathbf{x}|\omega_i, \theta_i)$ where the only unknown component is the *parameter vector* θ_i.
- E.g. $p(\mathbf{x}|\omega_i) = N(\mu_i, \Sigma_i)$.
- Actually, we should write $p_i(\mathbf{x}|\omega_i, \theta_i)$ to make it clear that parametric forms of densities for different categories need not to be same.

Class conditional probabilities

- Now the task is much easier, we have to find good values for parameter vectors $\theta_1, \ldots, \theta_c$ to specify the class conditional densities.
- Note that each of these can be estimated independently due to our assumption of conditional independence.
- Hence, the problem is: Estimate (i.e. find a good value for) θ given some training data $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
- We further assume that samples are i.i.d (independent and identically) distributed random variables distributed according to density $p(\mathbf{x}|\theta)$.

Sample mean and covariance

- $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
- Sample mean of $\mathcal{D}$ is $\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$.
- Sample covariance of $\mathcal{D}$ is $\frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mu})(\mathbf{x}_i - \bar{\mu})^T$.

Maximum-likelihood estimation

- In maximum-likelihood estimation we aim to find such θ that maximizes the probability of $\mathcal{D}$.
- The probability of $\mathcal{D}$ is

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

- Viewed as a function of θ, this is called *likelihood function*.
- Our job is now find the value $\hat{\theta}$ for θ that maximizes the likelihood function. It is called the maximum-likelihood (ML) estimate.
- If we can find an explicit formula for $\hat{\theta}$ it is called the ML estimator.

Log-likelihood

- In most cases, it is easier to maximize log-likelihood instead of likelihood. These both tactics lead to same estimators.
- Log-likelihood function:

$$l(\theta) = \ln p(\mathcal{D}|\theta) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\theta).$$

Finding the ML estimate

- The standard procedure:
  1. Find the gradient (or equivalently all the partial derivatives) of the (log-)likelihood function.
  2. Find the zeros of the gradient and all the other critical points of the (log-)likelihood function.
  3. Evaluate the (log-)likelihood function at those points and select the maximum value as the maximum likelihood estimate.
- Note: If there is just one critical point, it must be confirmed that it is a local maximum and not a local minimum. However, this is a pain and, for our purposes, not very educative. Hence, our approach is 'It is easy to show that...'.

Gradient

- The gradient of a function of several variables $\theta = (\theta_1, \ldots, \theta_k)^T$:

$$\nabla f(\theta) = \begin{bmatrix} \frac{df(\theta)}{d\theta_1} \\ \vdots \\ \frac{df(\theta)}{d\theta_k} \end{bmatrix}$$

- Example: $\theta = (x, y)^T$, $f(x, y) = x^2 y + x^2 + 2$.

$$\nabla f(x, y) = (2xy + 2x, x^2)^T$$

The Gaussian case: unknown μ

- $\theta = \mu$, the covariance Σ is known.
- The log-likelihood

$$l(\theta) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\mu) = -0.5(\ln[(2\pi)^{d/2} \det(\Sigma)] + (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu)).$$

- Gradient $\nabla l(\mu) = -\sum_{i=1}^{n} \Sigma^{-1}(\mathbf{x}_i - \mu)$.
- Set $\nabla l(\mu) = \mathbf{0}$. Solving yields the MLE

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i.$$

- (There are no other critical points and it is easy to show that $\hat{\mu}$ is a local maximizer.)

The Gaussian case: unknown μ and Σ

- In this case MLEs are more difficult to derive (except in the case of single feature).
- MLE for μ: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$
- MLE for Σ: $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^T$.

The Gaussian case: numeric example

```
 x_1  x_2   x_3    x_4   x_5
 13  -115   119    33   -19
 29   119    -4    17    73

mu =  6.2000
      46.8000

sigma =  5762.6   -3212.2
        -3212.2    1937.0
```


Example: Independent binary features

- $P(\mathbf{x}|\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$, $x_i = 0$ or 1.
- Find estimates for p_i given $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
- Features independent $\rightarrow$ each p_i can be found independently from others.
- Log-likelihood:
  $l(\mathcal{D}|p_i) = \ln P((\mathcal{D}|p_i) = \sum_{k=1}^{n} x_{k_i} \ln p_i + (1 - x_{k_i}) \ln(1 - p_i)$
- MLE: $\hat{p}_i = \frac{\sum_{k=1}^{n} x_{k_i}}{n}$.

Bias

- An estimator is *unbiased* if it, on the average, produces correct estimates.
- For example, the maximum likelihood estimator for the Gaussian mean is unbiased, i.e.

  $$E_n[\hat{\mu}] = \mu,$$

  where the expectation is taken over all possible training sets consisting of n samples.
- The maximum likelihood estimator for the Gaussian covariance is *biased* but it is *asymptotically unbiased*. That is, it is unbiased for infinite n.

Variance

- Another important quantity for assessing the estimator performance is the estimator variance, which measures the average divergence between estimates obtained from different training samples.
- Smaller the bias, better the estimator
- Smaller the variance, better the estimator
- Bias-variance tradeoff $\rightarrow$ Minimum variance unbiased estimator is the best estimator.

Sources of error

- **Bayes Error**: The error due to overlapping densities for different categories. Can be never eliminated.
- **Model Error**: The error due having incorrect model. The error can be only eliminated if the designer specifies the model which includes the true model that generated the data.
- **Estimation Error**: The error due to estimating parameters from a finite sample. Can be reduced by increasing the amount of training data.