Chapter 3

Maximum likelihood and Bayesian parameter estimation

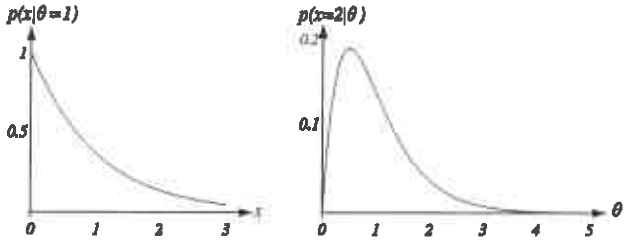
Problem Solutions

Section 3.2

1. Our exponential function is:

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) SEE FIGURE.



(b) The log-likelihood function is

$$l(\theta) = \sum_{k=1}^{n} \ln p(x_k|\theta) = \sum_{k=1}^{n} [\ln \theta - \theta x_k] = n \ln \theta - \theta \sum_{k=1}^{n} x_k.$$

We solve $\nabla_\theta l(\theta) = 0$ to find $\hat{\theta}$ as

$$\nabla_\theta l(\theta) = \frac{\partial}{\partial \theta} \left[n \ln \theta - \theta \sum_{k=1}^{n} x_k \right]$$

$$= \frac{n}{\theta} - \sum_{k=1}^{n} x_k = 0.$$

Thus the maximum-likelihood solution is

$$\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{k=1}^{n} x_k}.$$

2. Our (normalized) distribution function is

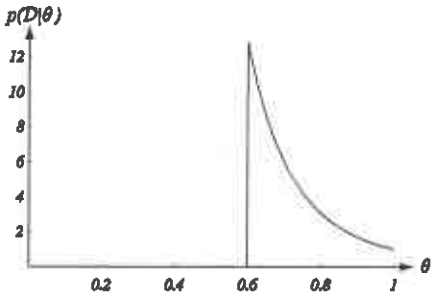
$$p(x|\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

(a) We will need to use the convention of an *indicator function* $I(\cdot)$, whose value is equal to 1.0 if the logical value of its argument is TRUE, and 0.0 otherwise. We can write the likelihood function using $I(\cdot)$ as

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{k=1}^{n} p(x_k|\theta) \\ &= \prod_{k=1}^{n} \frac{1}{\theta} I\left(0 \le x_k \le \theta\right) \\ &= \frac{1}{\theta^n} I\left(\theta \ge \max_k x_k\right)\ I\left(\min_k x_k \ge 0\right). \end{aligned}$$

We note that $1/\theta^n$ decreases monotonically as θ increases but also that $I(\theta \ge \max_k x_k)$ is 0.0 if θ is less than the maximum value of x_k. Therefore, our likelihood function is maximized at $\hat{\theta} = \max_k x_k$.

(b) SEE FIGURE.



3. We are given that

$$z_{ik} = \begin{cases} 1 & \text{if the state of nature for the } k^{th} \text{ sample is } \omega_i \\ 0 & \text{otherwise.} \end{cases}$$

(a) The samples are drawn by successive independent selection of a state of nature ω_i with probability $P(\omega_i)$. We have then

$$\begin{aligned} P(z_{ik} = 1|P(\omega_i)) &= P(\omega_i) \text{ and} \\ P(z_{ik} = 0|P(\omega_i)) &= 1 - P(\omega_i). \end{aligned}$$

These two equations can be unified as

$$P(z_{ik}|P(\omega_i)) = [P(\omega_i)]^{z_{ik}}[1-P(\omega_i)]^{1-z_{ik}}.$$

By the independence of the successive selections, we have

$$\begin{aligned} P(z_{i1},\cdots,z_{in}|P(\omega_i)) &= \prod_{k=1}^{n} P(z_{ik}|P(\omega_i)) \\ &= \prod_{k=1}^{n}[P(\omega_i)]^{z_{ik}}[1-P(\omega_i)]^{1-z_{ik}}. \end{aligned}$$

(b) The log-likelihood function for $P(\omega_i)$ is

$$\begin{aligned} l(P(\omega_i)) &= \ln P(z_{i1},\cdots,z_{in}|P(\omega_i)) \\ &= \ln\left[\prod_{k=1}^{n}[P(\omega_i)]^{z_{ik}}[1-P(\omega_i)]^{1-z_{ik}}\right] \\ &= \sum_{k=1}^{n}\left[z_{ik}\ln P(\omega_i) + (1-z_{ik})\ln(1-P(\omega_i))\right]. \end{aligned}$$

Therefore, the maximum-likelihood values for the $P(\omega_i)$'s must satisfy

$$\nabla_{P(\omega_i)} l(P(\omega_i)) = \frac{1}{P(\omega_i)}\sum_{k=1}^{n} z_{ik} - \frac{1}{1-P(\omega_i)}\sum_{k=1}^{n}(1-z_{ik}) = 0.$$

We solve this equation and find

$$(1-\hat{P}(\omega_i))\sum_{k=1}^{n} z_{ik} = \hat{P}(\omega_i)\sum_{k=1}^{n}(1-z_{ik}),$$

which can be rewritten as

$$\sum_{k=1}^{n} z_{ik} = \hat{P}(\omega_i)\sum_{k=1}^{n} z_{ik} + n\hat{P}(\omega_i) - \hat{P}(\omega_i)\sum_{k=1}^{n} z_{ik}.$$

The solution is then

$$\hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} z_{ik}.$$

4. We have an n samples $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ from the discrete distribution

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d}\theta_i^{x_i}(1-\theta_i)^{1-x_i}.$$

The likelihood for a particular sequence of n samples is

$$P(\mathbf{x}_1,\ldots,\mathbf{x}_n|\boldsymbol{\theta}) = \prod_{k=1}^{n}\prod_{i=1}^{d}\theta_i^{x_{ki}}(1-\theta_i)^{1-x_{ki}},$$

and the log-likelihood function is then

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n}\sum_{i=1}^{d} x_{ki} \ln \theta_i + (1 - x_{ki}) \ln (1 - \theta_i).$$

We set $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = 0$ and evaluate component by component ($i = 1, \ldots, d$) to get:

$$\begin{aligned}[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})]_i &= \nabla_{\boldsymbol{\theta}_i} l(\boldsymbol{\theta}) \\ &= \frac{1}{\theta_i}\sum_{k=1}^{n} - \frac{1}{1-\theta_i}\sum_{k=1}^{n}(1 - x_{ki}) \\ &= 0.\end{aligned}$$

This implies that for any i

$$\frac{1}{\hat{\theta}_i}\sum_{k=1}^{n} x_{ki} = \frac{1}{1-\hat{\theta}_i}\sum_{k=1}^{n}(1 - x_{ki}),$$

which can be rewritten as

$$(1 - \hat{\theta}_i)\sum_{k=1}^{n} x_{ki} = \hat{\theta}_i \left(n - \sum_{k=1}^{n} x_{ki} \right).$$

The solution is then

$$\hat{\theta}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}.$$

Since this result is valid for all $i = 1, \ldots, d$, we can write this last equation in vector form as

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

5. The probability of finding feature x_i to be 1.0 in category ω_1 is denoted p:

$$p(x_i = 1|\omega_1) = 1 - p(x_i = 0|\omega_1) = p_{i1} = p > \frac{1}{2},\ i = 1, \ldots, d.$$

Moreover, the normalization condition gives $p_{i2} = p(x_i|\omega_2) = 1 - p_{i1}$.

(a) A single observation $\mathbf{x} = (x_1, \ldots, x_d)$ is drawn from class ω_1, and thus have

$$p(\mathbf{x}|\omega_1) = \prod_{i=1}^{d} p(x_i|\omega_1) = \prod_{i=1}^{d} p^{x_i}(1 - p)^{1-x_i},$$

and the log-likelihood function for p is

$$l(p) = \ln p(\mathbf{x}|\omega_1) = \sum_{i=1}^{d} [x_i \ln p + (1 - x_i) \ln (1 - p)].$$

Thus the derivative is

$$\nabla_p l(p) = \frac{1}{p}\sum_{i=1}^{d} x_i - \frac{1}{(1-p)}\sum_{i=1}^{d}(1-x_i).$$

We set this derivative to zero, which gives

$$\frac{1}{\hat{p}}\sum_{i=1}^{d} x_i = \frac{1}{1-\hat{p}}\sum_{i=1}^{d}(1-x_i),$$

which after simple rearrangement gives

$$(1-\hat{p})\sum_{i=1}^{d} x_i = \hat{p}\left(d - \sum_{i=1}^{d} x_i\right).$$

Thus our final solution is

$$\hat{p} = \frac{1}{d}\sum_{i=1}^{d} x_i.$$

That is, the maximum-likelihood estimate of the probability of obtaining a 1 in any position is simply the ratio of the number of 1's in a single sample divided by the total number of features, given that the number of features is extremely large.

(b) We define $T = 1/d \sum_{i=1}^{d} x_i$ to be the proportion of 1's in a single observation $\mathbf{x}$. As the number of dimensions d approaches infinity, we have

$$T = \frac{1}{d}\sum_{i=1}^{d} \mathcal{E}(x_i|\omega_1) = [1 \times p + 0 \times (1-p)] = p.$$

Likewise, the variance of T, given that we're considering just one class, ω_1, is

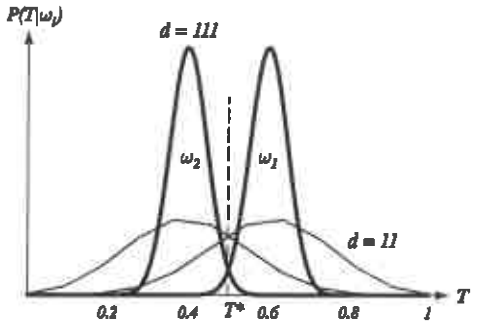
$$\begin{aligned}\mathrm{Var}(T|\omega_1) &= \frac{1}{d}\sum_{i=1}^{d}\mathrm{Var}(x_i|\omega_1) \\ &= \frac{1}{d^2}\sum_{i=1}^{d}[1^2 \times p + 0^2 \times (1-p) - p \times p] \\ &= \frac{p(1-p)}{d},\end{aligned}$$

which vanishes as $d \to \infty$. Clearly, for minimum error, we choose ω_1 if $T > T^* = 1/2$ for $p > 1/2$. Since the variance vanishes for large d, the probability of error is zero for a single sample having a sufficiently large d.

(c) SEE FIGURE.

6. The d-dimensional multivariate normal density is given by

$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\ rmexp\ \left[-\frac{1}{2}2(\mathbf{x}-\mu)^t\Sigma^{-1}(\mathbf{x}-\mu)\right].$$



We choose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ independent observations from $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The joint density is

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})\right].$$

The log-likelihood function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\left[\sum_{k=1}^{n}\mathbf{x}_k^t\mathbf{x}_k - 2\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k + n\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]. \end{aligned}$$

We set the derivative of the log-likelihood to zero, that is,

$$\frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\left[-2\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}x_k + n2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] = 0,$$

and find that

$$\hat{\boldsymbol{\Sigma}}^{-1}\sum_{k=1}^{n}\mathbf{x}_k = n\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}.$$

This gives the maximum-likelihood solution,

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

as expected. In order to simplify our calculation of $\hat{\boldsymbol{\Sigma}}$, we temporarily substitute $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$, and thus have

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln|\mathbf{A}| - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})^t\mathbf{A}(\mathbf{x}_k - \boldsymbol{\mu}).$$

We use the above results and seek the solution to

$$\frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \mathbf{A}} = \frac{n}{2}\mathbf{A}^{-1} - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t = 0.$$

We now replace $\mathbf{A}$ by $\mathbf{\Sigma}^{-1}$ and find that

$$\frac{n}{2}\hat{\mathbf{\Sigma}} = \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t.$$

We multiply both sides by $2/n$ and find our solution:

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

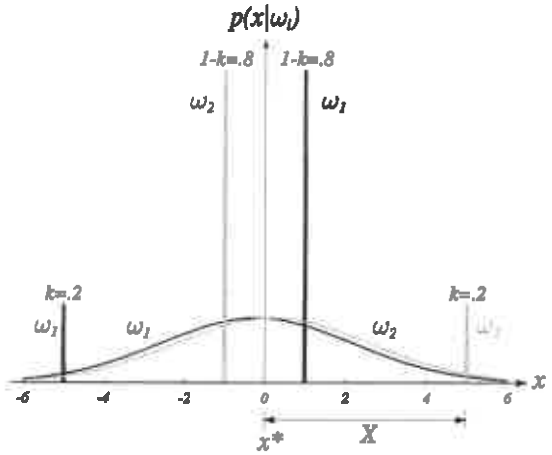
As we would expect, the maximum-likelihood estimate of the covariance matrix is merely the covariance of the samples actually found.

7. PROBLEM NOT YET SOLVED

8. Consider a case in which the maximum-likelihood solution gives the *worst* possible classifier.

(a) The symmetry operation $x \leftrightarrow -x$ takes $p(x|\omega_1) \leftrightarrow p(x|\omega_2)$, and thus we are assured that the estimated distributions have this same symmetry property. For that reason, we are guaranteed that these distribution have the same value at $x = 0$, and thus $x^* = 0$ is a decision boundary. Since the Gaussian estimates must have the same variance, we are assured that there will be only a single intersection, and hence a single decision point at $x^* = 0$.

(b) SEE FIGURE.



(c) We have the estimate of the mean as

$$\hat{\mu}_1 = \int p(x|\omega_1)dx = (1-k)1 + k(-X) = 1 - k(1+X).$$

We are asked to "switch" the mean, i.e., have $\hat{\mu}_1 < 0$. This can be assured if $X > (1-k)/k$. (We get the symmetric answer for ω_2.)

(d) Since the decision boundary is at $x^* = 0$, the error is $1-k$, i.e., the value of the distribution spikes on the "wrong" side of the decision boundary.