

***First Course in Machine Learning 2nd edition Rogers
Solutions Manual***

SKU: 9781498738484-SOLUTIONS

Chapter 2

EX 2.1. The errors are real valued and hence a continuous random variable would be more appropriate.

EX 2.2. If all outcomes are equally likely, they have the same probability of occurring. Defining Y to be the random variable taking the value shown on a die, we can state the following:

$$P(Y = y) = r,$$

where r is a constant. From the definition of probabilities, we know that:

$$\sum_{y=1}^6 P(Y = y) = 1.$$

Substituting r into this gives us the following:

$$\sum_{y=1}^6 r = 1, \quad 6r = 1, \quad r = 1/6.$$

EX 2.3. (a) Y is a discrete random variable that can take any value from 0 to inf. The probability that $Y \leq 4$ is equal to the sum of all of the probabilities that satisfy $Y \leq 4$, $Y = 0, Y = 1, Y = 2, Y = 3, Y = 4$:

$$P(Y \leq 4) = \sum_{y=0}^4 P(Y = y).$$

When $\lambda = 5$, we can compute these probabilities as:

$$P(Y \leq 4) = 0.0067379 + 0.0336897 + 0.0842243 + 0.1403739 + 0.1754674 = 0.44049.$$

(b) Because Y has to satisfy either $P(Y \leq 4)$ or $P(Y > 4)$, we know that $P(Y > 4) = 1 - P(Y \leq 4)$:

$$P(Y > 4) = 0.5591.$$

EX 2.4. We require $\mathbf{E}_{p(y)} \{ \sin(y) \}$ where $p(y) = \mathcal{U}(a, b)$. The uniform density is given by:

$$p(y) = \begin{cases} \frac{1}{b-a} & a \leq y \leq b \\ 0 & \text{otherwise} \end{cases}$$

The required expectation is given by:

$$\begin{aligned}
 \mathbf{E}_{p(y)} \{ \sin(y) \} &= \int \sin(y) p(y) dy \\
 &= \int_{y=a}^b \sin(y) \frac{1}{b-a} dy \\
 &= \frac{1}{b-a} [-\cos(y)]_a^b \\
 &= \frac{\cos(a) - \cos(b)}{b-a}.
 \end{aligned}$$

When $a = 0$, $b = 1$, this is equal to

$$\mathbf{E}_{p(y)} \{ \sin(y) \} = \frac{\cos(0) - \cos(1)}{1} = 0.45970.$$

Code to compute a sample-based approximation below (`sampleexpect.m`):

```

1 clear all;
2 close all;
3 % Compute a sample based approximation to the required expectation
4 u = rand(10000,1); % Take 10000 samples
5 su = sin(u);
6 % Plot how the approximation changes as more samples are used
7 ns = 10:100:10000;
8 stages = zeros(size(ns));
9 for i = 1:length(ns)
10     stages(i) = mean(su(1:ns(i)));
11 end
12 plot(ns,stages)
13 % Plot the true value
14 hold on
15 plot([0 ns(end)], [0.4597 0.4597], 'k—')
```

EX 2.5. The multivariate Gaussian pdf is given by:

$$p(\mathbf{w}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{w} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{w} - \boldsymbol{\mu}) \right\}.$$

Setting $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ gives:

$$p(\mathbf{w}) = \frac{1}{(2\pi)^{D/2} |\sigma^2 \mathbf{I}|^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{w} - \boldsymbol{\mu})^T \mathbf{I}^{-1} (\mathbf{w} - \boldsymbol{\mu}) \right\}.$$

Because it only has elements on the diagonal, the determinant of $\sigma^2 \mathbf{I}$ is given by the product of these diagonal elements. As they are all the same, $|\sigma^2 \mathbf{I}|^{1/2} = \left(\prod_{d=1}^D \sigma^2 \right)^{1/2} = (\sigma^2)^{D/2}$. $\mathbf{I}^{-1} = \mathbf{I}$ and multiplying a vector/matrix by $\mathbf{I}$ leaves the matrix/vector unchanged. Therefore, the argument within the expectation can be written as $-\frac{1}{2\sigma^2} (\mathbf{w} - \boldsymbol{\mu})^T (\mathbf{w} - \boldsymbol{\mu})$ and recalling that $\mathbf{b}^T \mathbf{b} = \sum_i b_i^2$, we can rewrite the pdf as:

$$p(\mathbf{w}) = \frac{1}{(2\pi)^{D/2} (\sigma^2)^{D/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{d=1}^D (w_d - \mu_d)^2 \right\}.$$

Where w_d and μ_d are the d th elements of $\mathbf{w}$ and $\boldsymbol{\mu}$ respectively. The exponential of a sum is the same as a product of exponentials. Hence,

$$\begin{aligned} p(\mathbf{w}) &= \frac{1}{(2\pi)^{D/2}(\sigma^2)^{D/2}} \prod_{d=1}^D \exp \left\{ -\frac{1}{2\sigma^2} (w_d - \mu_d)^2 \right\} \\ &= \prod_{d=1}^D \frac{1}{(2\pi)^{1/2}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} (w_d - \mu_d)^2 \right\} \\ &= \prod_{d=1}^D p(w_d | \mu_d, \sigma^2), \end{aligned}$$

where $p(w_d | \mu_d, \sigma^2) = \mathcal{N}(w_d | \mu_d, \sigma^2)$. Hence, the diagonal covariance is equivalent to assuming that the elements of $\mathbf{w}$ are distributed as independent, univariate Gaussians with mean μ_d and variance σ^2 .

EX 2.6. Using the same methods as in the previous exercise, we see that the determinant of the covariance matrix is given by $\prod_{d=1}^D \sigma_d^2$ and we have the following:

$$p(\mathbf{w}) = \frac{1}{(2\pi)^{D/2} \left(\prod_{d=1}^D \sigma_d^2 \right)^{1/2}} \exp \left\{ -\frac{1}{2} \sum_{d=1}^D \frac{(w_d - \mu_d)^2}{\sigma_d^2} \right\}$$

Changing the sum to a product leaves us with

$$\begin{aligned} p(\mathbf{w}) &= \frac{1}{(2\pi)^{D/2} \left(\prod_{d=1}^D \sigma_d^2 \right)^{1/2}} \prod_{d=1}^D \exp \left\{ -\frac{1}{2\sigma_d^2} (w_d - \mu_d)^2 \right\} \\ &= \prod_{d=1}^D \frac{1}{(2\pi)^{1/2}\sigma_d} \exp \left\{ -\frac{1}{2\sigma_d^2} (w_d - \mu_d)^2 \right\}. \end{aligned}$$

This is the product of D independent univariate Gaussian densities.

EX 2.7. The Hessian for a general model of our form is given by:

$$-\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X}$$

For the linear model, $\mathbf{X}$ is defined as:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$$

Therefore $-\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X}$ is:

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} N & \sum_{n=1}^N x_n \\ \sum_{n=1}^N x_n & \sum_{n=1}^N x_n^2 \end{bmatrix}$$

The diagonal elements are $-N/\sigma^2$ and $-(1/\sigma^2)\sum_{n=1}^N x_n^2$ which are equivalent (they differ only by multiplication with a negative constant) the expressions obtained in Chapter 1.

EX 2.8. We have N values, $x_1, \dots, x_N$. Assuming that these values came from a Gaussian, we want to find the maximum likelihood estimate of the μ and want to find the maximum likelihood estimates of the mean and variance of the Gaussian. The Gaussian pdf is:

$$\frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2}(x_n - \mu)^2 \right\}$$

Assuming the IID assumption, the likelihood of all N points is given by a product over the N objects:

$$\prod_{n=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2}(x_n - \mu)^2 \right\}.$$

We'll work with the log of the likelihood:

$$\log L = \sum_{n=1}^N \left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2 \right)$$

To find the maximum likelihood estimate for μ , we differentiate with respect to μ , equate to zero and solve:

$$\begin{aligned} \frac{\partial \log L}{\partial \mu} &= \sum_{n=1}^N \frac{1}{\sigma^2}(x_n - \mu) \\ 0 &= \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu) \\ 0 &= \sum_{n=1}^N x_n - \sum_{n=1}^N \mu \\ &= \sum_{n=1}^N x_n - N\mu \\ \mu &= \frac{1}{N} \sum_{n=1}^N x_n \end{aligned}$$

Similarly, for σ^2 ,

$$\begin{aligned} \frac{\partial \log L}{\partial \sigma^2} &= \sum_{n=1}^N \left(-\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(x_n - \mu)^2 \right) = 0 \\ N\sigma^2 &= \sum_{n=1}^N (x_n - \mu)^2 \end{aligned} \tag{2.1}$$

$$\sigma^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 \tag{2.2}$$

EX 2.9. The Bernoulli distribution is defined as:

$$P(X_n = x|r) = r^x(1-r)^{1-x}$$

where x is either 0 or 1. Using the IID assumption, we have:

$$L = \prod_{n=1}^N r^{x_n}(1-r)^{1-x_n}$$

and the log likelihood is:

$$\log L = \sum_{n=1}^N x_n \log r + (1-x_n) \log(1-r)$$

Differentiating with respect to r gives us:

$$\begin{aligned} \frac{\partial \log L}{\partial r} &= \sum_{n=1}^N \left(\frac{x_n}{r} - \frac{1-x_n}{1-r} \right) = 0 \\ \sum_{n=1}^N \frac{x_n}{r} &= \sum_{n=1}^N \frac{1-x_n}{1-r} \\ \sum_{n=1}^N x_n - r \sum_{n=1}^N x_n &= rN - r \sum_{n=1}^N x_n \\ r &= \frac{1}{N} \sum_{n=1}^N x_n. \end{aligned}$$

EX 2.10. The Fisher information is defined as the expectation of the negative second derivative. From the above expression, we can see that the second derivative of the Gaussian likelihood (assuming N observations, $x_1, \dots, x_N$ is:

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{N}{\sigma^2}.$$

Hence the Fisher information is equal to N/σ^2 .

EX 2.11. Starting from the second expression, we have

$$\widehat{\sigma^2} = \frac{1}{N} \left[\sum_{n=1}^N t_n^2 - 2 \sum_{n=1}^N t_n \mathbf{x}_n^T \widehat{\mathbf{w}} + \sum_{n=1}^N (\mathbf{x}_n \widehat{\mathbf{w}})^2 \right].$$

Concentrating on the final term,

$$\begin{aligned}
 \sum_{n=1}^N (\mathbf{x}_n^\top \hat{\mathbf{w}})^2 &= \sum_{n=1}^N \mathbf{x}_n^\top \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{x}_n \\
 &= \text{Tr}(\mathbf{X} \hat{\mathbf{w}} \hat{\mathbf{w}}^\top \mathbf{X}^\top) \\
 &= \text{Tr}(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t} \mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \\
 &= \text{Tr}(\mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t} \mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}) \\
 &= \text{Tr}(\mathbf{X}^\top \mathbf{t} \mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}) \\
 &= \text{Tr}(\mathbf{t} \mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \\
 &= \text{Tr}(\mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t}) \\
 &= \mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t} \\
 &= \mathbf{t}^\top \mathbf{X} \hat{\mathbf{w}} \\
 &= \sum_{n=1}^N t_n \mathbf{x}_n^\top \hat{\mathbf{w}}.
 \end{aligned}$$

Therefore,

$$\hat{\sigma}^2 = \frac{1}{N} \left[\sum_{n=1}^N t_n^2 - \sum_{n=1}^N t_n \mathbf{x}_n^\top \hat{\mathbf{w}} \right].$$

Now, $\sum_{n=1}^N t_n^2 = \mathbf{t}^\top \mathbf{t}$ and we already know that $\sum_{n=1}^N t_n \mathbf{x}_n^\top \hat{\mathbf{w}} = \mathbf{t}^\top \mathbf{X} \hat{\mathbf{w}}$. So,

$$\hat{\sigma}^2 = \frac{1}{N} [\mathbf{t}^\top \mathbf{t} - \mathbf{t}^\top \mathbf{X} \hat{\mathbf{w}}],$$

as required.

EX 2.12. Code below (`predvar.m`):

```

1 clear all;close all;
2 % Relevant code extracted from predictive_variance_example.m
3 x = rand(50,1)*10-5;
4 x = sort(x);
5 % Compute true function values
6 f = 5*x.^3 - x.^2 + x;
7 % Generate some test locations
8 testx = [min(x):0.2:max(x)]';
9 % Add some noise
10 t = f+randn(50,1)*sqrt(1000);
11 % Remove all training data between -1.5 and 1.5
12 pos = find(x>-1.5 & x<1.5);
13 x(pos) = [];
14 f(pos) = [];
15 t(pos) = [];
16
17 % Choose model order
18 K = 5;
19
20 X = repmat(1,size(x));
21 testX = repmat(1,size(testx));

```

```

22 for k = 1:K
23     X = [X x.^k];
24     testX = [testX testx.^k];
25 end
26
27
28 w_hat = inv(X'*X)*X'*t;
29 ss_hat = mean((t - X*w_hat).^2);
30 pred_va = ss_hat*diag(testX*inv(X'*X)*testX');
31 % Make a plot
32 figure(1);hold off
33 plot(x,t,'b. ');
34 hold on
35 errorbar(testx,testX*w_hat,pred_va,'r');

```

EX 2.13. The Bernoulli distribution for a binary random variable x is:

$$p(x|\theta) = \theta^x (1 - \theta)^{1-x}$$

The Fisher information is defined as the negative expected value of the second derivative of the log density evaluated at some parameter value:

$$\mathcal{F} = -\mathbf{E}_{p(x|\theta)} \left\{ \frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} \bigg|_{\theta} \right\}$$

Differentiating $\log p(x|\theta)$ twice gives:

$$\begin{aligned} \frac{\partial \log p(x|\theta)}{\partial \theta} &= \frac{x}{\theta} - \frac{1-x}{1-\theta} \\ \frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} &= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}. \end{aligned}$$

The Fisher information is therefore:

$$\mathcal{F} = \frac{1}{\theta^2} \mathbf{E}_{p(x|\theta)} \{x\} + \frac{1}{(1-\theta)^2} \mathbf{E}_{p(x|\theta)} \{1-x\}.$$

Substituting in the expectations (θ and $1-\theta$ respectively) gives:

$$\mathcal{F} = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

EX 2.14. The multivariate Gaussian pdf is given by:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}.$$

Logging and removing terms not including $\boldsymbol{\mu}$:

$$\log p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}.$$

First and second derivatives are:

$$\begin{aligned}\frac{\partial \log p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} &= \boldsymbol{\Sigma}^{-1} \mathbf{x} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ \frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^\top} &= -\boldsymbol{\Sigma}^{-1}.\end{aligned}$$

Therefore, the Fisher information is:

$$\mathcal{F} = \boldsymbol{\Sigma}^{-1}.$$