Pattern Recognition and Machine Learning

Solutions to the Exercises: Web-Edition

Markus Svensén and Christopher M. Bishop

March 23, 2007

Pattern Recognition and Machine Learning
Solutions to the Exercises: Web-Edition

Markus Svensén and Christopher M. Bishop

This is an early partial release of the solutions manual (web-edition) for the book *Pattern Recognition and Machine Learning* (PRML; published by Springer in 2006). It contains solutions to the www exercises in PRML for chapters 1–4. This release was created March 23, 2007; further releases with solutions to later chapters will be published in due course on the PRML web-site (see below).

The authors would like to express their gratitude to the various people who have provided feedback on pre-releases of this document. In particular, the "Bishop Reading Group", held in the Visual Geometry Group at the University of Oxford provided valuable comments and suggestions.

Please send any comments, questions or suggestions about the solutions in this document to Markus Svensén, markussv@microsoft.com.

Further information about PRML is available from:

http://research.microsoft.com/∼cmbishop/PRML

Contents

Chapter 1 Pattern Recognition

1.1 Substituting (1.1) into (1.2) and then differentiating with respect to w_i we obtain

$$\sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x_n^j - t_n\right) x_n^i = 0. \tag{1}$$

Re-arranging terms then gives the required result.

1.4 We are often interested in finding the most probable value for some quantity. In the case of probability distributions over discrete variables this poses little problem. However, for continuous variables there is a subtlety arising from the nature of probability densities and the way they transform under non-linear changes of variable.

Consider first the way a function $f(x)$ behaves when we change to a new variable y where the two variables are related by $x = g(y)$. This defines a new function of y given by

$$\widetilde{f}(y) = f(g(y)). \tag{2}$$

Suppose $f(x)$ has a mode (i.e. a maximum) at $\widehat{x}$ so that $f'(\widehat{x}) = 0$. The corresponding mode of $\widetilde{f}(y)$ will occur for a value $\widehat{y}$ obtained by differentiating both sides of (2) with respect to y

$$\widetilde{f}'(\widehat{y}) = f'(g(\widehat{y}))g'(\widehat{y}) = 0. \tag{3}$$

Assuming $g'(\widehat{y}) \neq 0$ at the mode, then $f'(g(\widehat{y})) = 0$. However, we know that $f'(\widehat{x}) = 0$, and so we see that the locations of the mode expressed in terms of each of the variables x and y are related by $\widehat{x} = g(\widehat{y})$, as one would expect. Thus, finding a mode with respect to the variable x is completely equivalent to first transforming to the variable y, then finding a mode with respect to y, and then transforming back to x.

Now consider the behaviour of a probability density $p_x(x)$ under the change of variables $x = g(y)$, where the density with respect to the new variable is $p_y(y)$ and is given by ((1.27)). Let us write $g'(y) = s|g'(y)|$ where $s \in \{-1, +1\}$. Then ((1.27)) can be written

$$p_y(y) = p_x(g(y))sg'(y).$$

Differentiating both sides with respect to y then gives

$$p'_y(y) = sp'_x(g(y))\{g'(y)\}^2 + sp_x(g(y))g''(y). \tag{4}$$

Due to the presence of the second term on the right hand side of (4) the relationship $\widehat{x} = g(\widehat{y})$ no longer holds. Thus the value of x obtained by maximizing $p_x(x)$ will not be the value obtained by transforming to $p_y(y)$ then maximizing with respect to y and then transforming back to x. This causes modes of densities to be dependent on the choice of variables. In the case of linear transformation, the second term

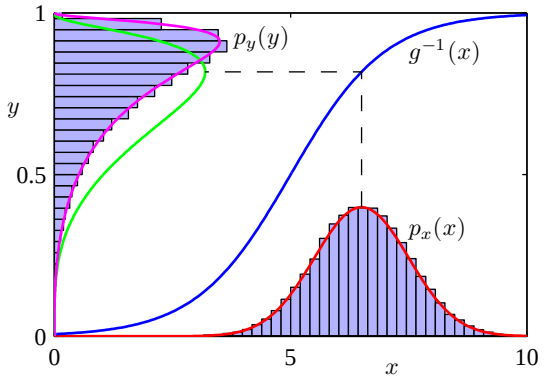

Figure 1 Example of the transformation of the mode of a density under a non-linear change of variables, illustrating the different behaviour compared to a simple function. See the text for details.

vanishes on the right hand side of (4) vanishes, and so the location of the maximum transforms according to $\widehat{x} = g(\widehat{y})$.

This effect can be illustrated with a simple example, as shown in Figure 1. We begin by considering a Gaussian distribution $p_x(x)$ over x with mean $\mu = 6$ and standard deviation $\sigma = 1$, shown by the red curve in Figure 1. Next we draw a sample of $N = 50,000$ points from this distribution and plot a histogram of their values, which as expected agrees with the distribution $p_x(x)$.

Now consider a non-linear change of variables from x to y given by

$$x = g(y) = \ln(y) - \ln(1-y) + 5. \tag{5}$$

The inverse of this function is given by

$$y = g^{-1}(x) = \frac{1}{1+\exp(-x+5)} \tag{6}$$

which is a *logistic sigmoid* function, and is shown in Figure 1 by the blue curve.

If we simply transform $p_x(x)$ as a function of x we obtain the green curve $p_x(g(y))$ shown in Figure 1, and we see that the mode of the density $p_x(x)$ is transformed via the sigmoid function to the mode of this curve. However, the density over y transforms instead according to (1.27) and is shown by the magenta curve on the left side of the diagram. Note that this has its mode shifted relative to the mode of the green curve.

To confirm this result we take our sample of $50,000$ values of x, evaluate the corresponding values of y using (6), and then plot a histogram of their values. We see that this histogram matches the magenta curve in Figure 1 and not the green curve!

1.7 The transformation from Cartesian to polar coordinates is defined by

$$x = r\cos\theta \tag{7}$$

$$y = r\sin\theta \tag{8}$$

and hence we have $x^2 + y^2 = r^2$ where we have used the well-known trigonometric result (2.177). Also the Jacobian of the change of variables is easily seen to be

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

where again we have used (2.177). Thus the double integral in (1.125) becomes

$$I^2 = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) r \,\mathrm{d}r \,\mathrm{d}\theta \tag{9}$$

$$= 2\pi \int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2}\right)\frac{1}{2}\,\mathrm{d}u \tag{10}$$

$$= \pi\left[\exp\left(-\frac{u}{2\sigma^2}\right)\left(-2\sigma^2\right)\right]_0^{\infty} \tag{11}$$

$$= 2\pi\sigma^2 \tag{12}$$

where we have used the change of variables $r^2 = u$. Thus

$$I = \left(2\pi\sigma^2\right)^{1/2}.$$

Finally, using the transformation $y = x - \mu$, the integral of the Gaussian distribution becomes

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\int_{-\infty}^{\infty}\exp\left(-\frac{y^2}{2\sigma^2}\right)\,\mathrm{d}y$$

$$= \frac{I}{\left(2\pi\sigma^2\right)^{1/2}} = 1$$

as required.

1.8 From the definition (1.46) of the univariate Gaussian distribution, we have

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} x\,\mathrm{d}x. \tag{13}$$

Now change variables using $y = x - \mu$ to give

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}y^2\right\}(y+\mu)\,\mathrm{d}y. \tag{14}$$

We now note that in the factor $(y + \mu)$ the first term in y corresponds to an odd integrand and so this integral must vanish (to show this explicitly, write the integral

as the sum of two integrals, one from $-\infty$ to 0 and the other from 0 to ∞ and then show that these two integrals cancel). In the second term, μ is a constant and pulls outside the integral, leaving a normalized Gaussian distribution which integrates to 1, and so we obtain (1.49).

To derive (1.50) we first substitute the expression (1.46) for the normal distribution into the normalization result (1.48) and re-arrange to obtain

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \mathrm{d}x = \left(2\pi\sigma^2\right)^{1/2}. \tag{15}$$

We now differentiate both sides of (15) with respect to σ^2 and then re-arrange to obtain

$$\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} (x-\mu)^2 \,\mathrm{d}x = \sigma^2 \tag{16}$$

which directly shows that

$$\mathbb{E}[(x-\mu)^2] = \mathrm{var}[x] = \sigma^2. \tag{17}$$

Now we expand the square on the left-hand side giving

$$\mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 = \sigma^2.$$

Making use of (1.49) then gives (1.50) as required.

Finally, (1.51) follows directly from (1.49) and (1.50)

$$\mathbb{E}[x^2] - \mathbb{E}[x]^2 = \left(\mu^2 + \sigma^2\right) - \mu^2 = \sigma^2.$$

1.9 For the univariate case, we simply differentiate (1.46) with respect to x to obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}\left(x|\mu,\sigma^2\right) = -\mathcal{N}\left(x|\mu,\sigma^2\right)\frac{x-\mu}{\sigma^2}.$$

Setting this to zero we obtain $x = \mu$.

Similarly, for the multivariate case we differentiate (1.52) with respect to $\mathbf{x}$ to obtain

$$\begin{aligned}\frac{\partial}{\partial\mathbf{x}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= -\frac{1}{2}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\nabla_{\mathbf{x}}\left\{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \\ &= -\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}),\end{aligned}$$

where we have used (C.19), (C.20) and the fact that $\boldsymbol{\Sigma}^{-1}$ is symmetric. Setting this derivative equal to $\mathbf{0}$, and left-multiplying by $\boldsymbol{\Sigma}$, leads to the solution $\mathbf{x} = \boldsymbol{\mu}$.

1.10 Since x and z are independent, their joint distribution factorizes $p(x,z) = p(x)p(z)$, and so

$$\mathbb{E}[x+z] = \iint (x+z)p(x)p(z)\,\mathrm{d}x\,\mathrm{d}z \tag{18}$$

$$= \int xp(x)\,\mathrm{d}x + \int zp(z)\,\mathrm{d}z \tag{19}$$

$$= \mathbb{E}[x] + \mathbb{E}[z]. \tag{20}$$