# Bishop Exercises

Nicolas Pacheco

January 2021

## 1 Chapter 1

**1.1**

Consider the sum-of-squares error function given by $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2$ in which the function $y(x, \mathbf{w})$ is given by the polynomial

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + . + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Show that the coefficients w = $w_i$ that minimize this error function are given by the solutions to the following set of linear equations:

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{1.122}$$

where

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j} \qquad T_i = \sum_{n=1}^{N} (x_n)^i t_n \tag{1.123}$$

Solution:

If we replace in $E(\mathbf{w})$, with $y(x, \mathbf{w})$, we get:

$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x^j - t_n)^2$

and differentiating with respect to $w_i$

$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x^j - t_n)x_n^i$

rearranging:

$\sum_{n=1}^{N}\sum_{j=0}^{M} w_j x_n^i x_n^j = \sum_{n=1}^{N} t_n x_n^i$

In the right side we have $\sum_{n=1}^{N} t_n x_n^i = T_i$

If in the left side we replace $\sum_{n=1}^{N} x_n^i x_n^j$ by $A_{ij}$

we prove the statement:

$\sum_{j=0}^{M} A_{ij} w_j = T_i$

**1.2**

Write down the set of couple linear equations, analogous to (1.122), satisfied by the coefficients $w_i$ which minimize the regularized sum of squares error function given by

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Solution: The solution is very similar to the last one.

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + . + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

If we replace in $E(\mathbf{w})$, with $y(x, \mathbf{w})$, we get:

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x^j - t_n)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

remember:$\|\mathbf{w}\| = \sqrt{\sum_{j=0}^{M} w_j^2}$

and again differentiating with respect to $w_i$

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x^j - t_n)x_n^i + \lambda\sum_{j=0}^{M} w_j = 0$$

rearranging and taken out common factor

$$\sum_{j=0}^{M} w_j(\sum_{n=1}^{N} x_n^{i+j} + \lambda) = \sum_{n=1}^{N} t_n x_n^i$$

The left side of the equation is the same that 1.1 .

$$\sum_{n=1}^{N} t_n x_n^i = T_i$$

Using $A_{ij} = \sum_{n=1}^{N}(x_n)^{i+j}$ from 1.1 we can create $\tilde{A_{ij}}$

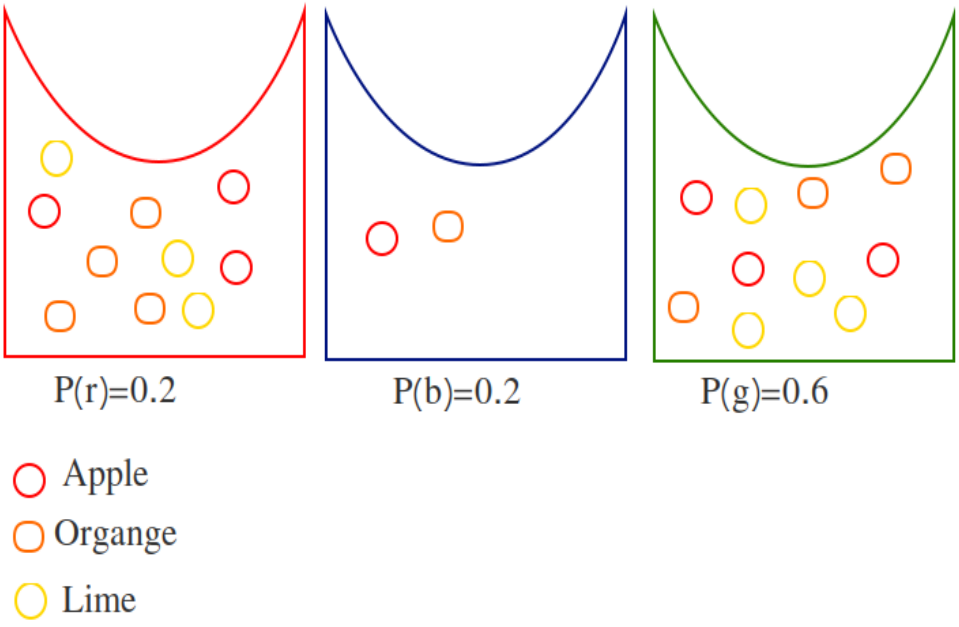
$$\tilde{A_{ij}} = A_{ij} + \lambda$$

**1.3**

Suppose that we have three coloured boxes $r$ red, $b$ blue, $g$ green. Box $r$ contains 3 apples, 4 oranges and 3 limes, box $b$ contains one apple, one orange and zero limes, and box $g$ contains 3 apples, 3 oranges and 4 limes. If a box is chosen at random with probabilities $p(r) = 0.2$, $p(b) = 0.2$ and $p(g) = 0.6$ and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then:

what is the probability of selecting an apple?

If we observe that the selected fruit is in fact orange, what is the probability that is came from the green box?




The boxes of the exercise

Solution: For the first question we have to compute $p(a)$ where a is apple. In words it is the probability of taking and apple given that we have selected the

red box or the probability of taking and apple given that we have selected the blue box or the probability of taking and apple given that we have selected the green box. If we express it as a probability

$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g)$

we know from the statement $p(r)$, $p(b)$ and $p(g)$, so we only need to compute the conditional probabilities. They are pretty easy! I mean, $p(a|r)$ is the probability of taking out an apple of the red box, and this is $\frac{\#ApplesInBoxRed}{\#FruitsInBoxRed}$, so:

$p(a|r) = \frac{\#ApplesInBoxRed}{\#FruitsInBoxRed} = \frac{3}{10}$

$p(a|b) = \frac{\#ApplesInBoxBlue}{\#FruitsInBoxBlue} = \frac{1}{2}$

$p(a|g) = \frac{\#ApplesInBoxGreen}{\#FruitsInBoxGreen} = \frac{3}{10}$

Now replacing in $p(a)$:

$p(a) = \frac{3}{10} * 0.2 + \frac{1}{2} * 0.2 + \frac{3}{10} * 0.6 = 0.34$

Now, the second question ask us to compute the probability of the box to be the green one given that the fruit is an orange. First we should write this en probability notation $p(g|o)$. But we do not have information to compute this. What can we do? Use the Bayes theorem.

$p(g|o) = \frac{p(o|g)p(g)}{p(o)}$. We know $p(g)$ from statement, and we can compute $p(o)$ and $p(o|g)$ likewise the first question.

$p(g) = 0.6$

$p(o) = p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g) =$

$\frac{4}{10} * 0.2 + \frac{1}{2} * 0.2 + \frac{3}{10} * 0.6 = 0.36$

$p(o|g) = \frac{3}{10}$

Replacing in the equation:

$p(g|o) = \frac{p(o|g)p(g)}{p(o)} = \frac{\frac{3}{10} * 0.6}{0.36} = \frac{1}{2}$

**1.4**

Consider a probability density $p_x(x)$ defined over a continuous variable x, and suppose that we make a nonlinear change of variable using $x = g(y)$ so that the density transforms according to $p_y(y) = p_x(x)|\frac{dx}{dy}| = p_x(g(y))|g'(y)|$. By differentiating that result, show that the location $\tilde{y}$ of the maximum of the density in $y$ is not in general related to the location $x$ of the maximum of the density over $x$ by the simple functional relation $\tilde{x} = g(\tilde{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent on the choice of variable. Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

Solution: First we will check what happen with a simple function. Suppose we have a function $f(x)$ and we have a transformation such that $x = g(y)$. So now we have a new function given by : $\tilde{f(y)} = f(g(y))$

Now suppose that $f(x)$ has a maximum at $\hat{x}$ such that $f'(\hat{x}) = 0$. The corresponding maximum of $\tilde{f(y)}$ will be at $\hat{y}$ and we can compute it by differentiating both sides of $\tilde{f(y)} = f(g(y))$. So, if we differentiate and use the chain rule we get:

$\tilde{f(\hat{y})} = f'(g(\hat{y}))g(\hat{y}) = 0$

Assuming $g'(y) \neq 0$ at the maximum, then $f'(g(\hat{y})) = 0$, and we see that the maximums expressed in terms of $x$ and $y$ are related by $\hat{x} = g(\hat{y})$. We can conclude that finding the maximum with respect to $x$ is equivalent to transform to $y$, find the maximum with respect to $y$ and then transform again to $x$.

Now we will check the behaviour of a probability density $p_x(x)$. Suppose we have the same transformation $x = g(y)$, where our new density will be $p_y(y)$. Using the transformation of the statement we know that the density with respect to $y$ is $p_y(y) = p_x(x)|\frac{dx}{dy}| = p_x(g(y))|g'(y)|$. To avoid the absolute value, we will re-write $g'(y) = s|g'(y)|$ where $s \in \{-1, 1\}$. Then

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

Differentiating both sides using the chain rule:

$$p'_y(y) = p'_x(g(y))sg'(y)^2 + p_x(g(y))sg''(y)$$

Now, we have a new term in the right of our result, because of this new term the relationship $\hat{x} = g(\hat{y})$ does not holds anymore. Therefore 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 $x$. We can conclude that the maximum of a density is dependent on the choice of the variable. This is in the general case, but what happen in the linear case?

First we should define the tranformation. So our new linear transformation will be $x = \alpha y$

$g(y) = \alpha y$

$g'(y) = \alpha$

Now we replace in our equation: $p_y(y) = p_x(\alpha y)|\alpha| = p_x(\alpha y)s * \alpha$

Differentiating both sides using the chain rule:

$p'_y(y) = p'_x(\alpha y)s\alpha^2$

Now we are in the same case as in the simple function. Is the same to find the maximum of $x$ or to transform to $y$, then to find the maximum with respect to $y$ and finally transform back to $x$ again.

**1.5**

Using the definition $var[f] = E[(f(x) - E[f(x))^2]$ show that $var[f(x)]$ satisfies $var[f] = E[f(x)^2] - E[x]^2$

Solution: I will provide two solutions: the first one using expectation properties and the second one using integrals.

For the first case we should know the following properties that are valid because of the linearity of expectation: $E[X + Y] = E[X] + E[Y]$ and $E[aX + b] = aE[X] + b$ where $a, b \in \mathbb{R}$

$var[f] = E[(f(x) - E[f(x))^2]$ first we should open the square.

$var[f] = E[f(x)^2 - 2f(x)E[f(x)] + E[f(x)^2]$ Using the property I mentioned above we can open the expectation

$var[f] = E[f(x)^2] - E[2f(x)E[f(x)]] + E[f(x)^2]$. In the middle term 2 and E[f(x)] are constants, so we can apply the property again

$var[f] = E[f(x)^2] - 2E[f(x)]E[f(x)] + E[f(x)^2]$

$var[f] = E[f(x)^2] - 2E[f(x)^2] + E[f(x)^2]$

$var[f] = E[f(x)^2] - E[f(x)^2]$

For the other solution we can use integrals. This one is my favourite. First we should know that $E[f(x)] = \int_{-\infty}^{\infty} p(x)f(x)dx$ and $E[f(x)] = \mu$. Now we can re-write $var[f] = E[(f(x) - E[f(x))^2] = E[(X - \mu)^2]$ as:

$var[f] = \int_{-\infty}^{\infty}(x - \mu)^2 f(x)dx =$

$\int_{-\infty}^{\infty} x^2 f(x) - 2x\mu f(x) + \mu^2 f(x)dx$. Now using the following property: The integral of the sum of two functions is equal to the sum of the integrals of these functions.

$\int_{-\infty}^{\infty} x^2 f(x)dx - \int_{-\infty}^{\infty} 2x\mu f(x)dx + \int_{-\infty}^{\infty} \mu^2 f(x)dx$. Now we can take out the constants, and because f(x) is a probability function if we integrate over all the domain we get 1: $\int_{-\infty}^{\infty} f(x)dx = 1$

$\int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu \int_{-\infty}^{\infty} xf(x)dx + \mu^2 \int_{-\infty}^{\infty} f(x)dx =$

$E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - E[X]^2$

**1.6**

Show that if two variables $x$ and $y$ are independent, then their covariance is zero.

Solution: I will provide here two solution, one for the continuous case and the other one for the discrete case.

As the variables are independent we know that $P(x, y) = P(x)P(y)$ Continuous case:

$cov[x,y] = E_{x,y}[xy] - E[x]E[y] = \int_{-\infty}^{\infty} xf(x)dx \int_{-\infty}^{\infty} yf(y)dy - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yxf(x,y)dydx$

Using that $x$ and $y$ are independent

$\int_{-\infty}^{\infty} xf(x)dx \int_{-\infty}^{\infty} yf(y)dy - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yxf(y)f(x)dydx =$

$\int_{-\infty}^{\infty} xf(x)dx \int_{-\infty}^{\infty} yf(y)dy - \int_{-\infty}^{\infty} xf(x) \int_{-\infty}^{\infty} yf(y)dydx =$

$E[x]E[y] - E[x]E[y] = 0$

Discrete case:

$cov[x,y] = E_{x,y}[xy] - E[x]E[y] = t\sum_x xf(x)dx \sum_y yf(y)dy - \sum_x \sum_y yxf(x,y)dydx =$

Using that $x$ and $y$ are independent

$\sum_x xf(x)dx \sum_y yf(y)dy - \sum_x \sum_y yxf(y)f(x)dydx =$

$\sum_x xf(x)dx \sum_y yf(y)dy - \sum_x xf(x) \sum_y yf(y)dydx =$

$E[x]E[y] - E[x]E[y] = 0$

**1.7**

In this exercise, we prove the normalization condition

$\int_{-\infty}^{\infty} N(x|\mu, \sigma^2)dx = 1$ for the univariaite Gaussian. To do this consider the integral:

$I = \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}x^2}dx$

which we can evaluate by first writing its square in the form:

$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}x^2-\frac{1}{2\sigma^2}y^2}dxdy$ Now make the transformation from Cartesian coordinates$(x, y)$ to polar coordinates $(r, \sigma)$ and the substitute $u = r^2$.Show that, by performing the integrals over $\sigma$ and $u$, and then making the square root of both sides, we obtain: $I = (2\pi\sigma^2)^{\frac{1}{2}}$

Finally,use this result to show that the Gaussian distribution $N(x|u, \sigma^2)$, is normalized

Solution: The transformation from Cartesian to Polar coordinates is defined by:

$x = cos(\theta)r$

$y = sin(\theta)r$

Using the famous trigonometric result $((cos(\theta))^2 + (sin(\theta))^2) = 1$ we can define:

$x^2 + y^2 = (cos(\theta)r)^2 + (sin(\theta)r)^2 = r^2((cos(\theta))^2 + (sin(\theta))^2) = r^2$

Also, we should compute the Jacobian of the transformation:

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

So JT $= rcos(\theta)^2 + rsin(\theta)^2 = r(cos(\theta)^2 + sin(\theta)^2) = r$ Using the trigonometric result.

Now with the jacobian we can start computing the integral:

$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}x^2-\frac{1}{2\sigma^2}y^2}dxdy = \int_0^{2\pi}\int_0^{\infty}(e^{-\frac{1}{2\sigma^2}(cos(\theta)r)^2-\frac{1}{2\sigma^2}(sin(\theta)r)^2}r)drd\theta =$

$\int_0^{2\pi}\int_0^{\infty}(e^{-\frac{1}{2\sigma^2}((cos(\theta)r)^2+(sin(\theta)r)^2)}r)drd\theta = \int_0^{2\pi}\int_0^{\infty}(e^{-\frac{1}{2\sigma^2}(r^2(cos(\theta)^2+rsin(\theta)^2)}r)drd\theta =$

$\int_0^{2\pi}\int_0^{\infty}(e^{-\frac{r^2}{2\sigma^2}}r)drd\theta$

We do not have $\theta$ anymore so we can integrate 1 over the whole domain and also we use the substitution method :

$r^2 = u$

$du = 2rdr$

$\frac{1}{2}du = rdr$

$2\pi\int_0^{\infty}(e^{-\frac{u}{2\sigma^2}}(12))du = \pi\left[e^{-\frac{u}{2\sigma^2}}(-2\sigma^2)\right]_0^{\infty} = 2\pi\theta = I^2$ So:

$I = (2\pi\theta)^{\frac{1}{2}}$

Now we should show that the Gaussian distribution $N(x|u,\sigma^2)$, is normalized:

$\int_{-\infty}^{\infty} N(x|u,\sigma^2)dx = \int_{-\infty}^{\infty}\frac{1}{(2\pi\theta)^{\frac{1}{2}}}e^{-\frac{1}{2\sigma^2}(x-u)^2}dx =$

$\frac{1}{(2\pi\theta)^{\frac{1}{2}}}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-u)^2}dx$ Using the transformation $y = x-u$ and the previous result

$\frac{1}{(2\pi\theta)^{\frac{1}{2}}}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}y^2}dy = \frac{I}{(2\pi\theta)^{\frac{1}{2}}} = \frac{(2\pi\theta)^{\frac{1}{2}}}{(2\pi\theta)^{\frac{1}{2}}} = 1$

**1.8**

By using a change of variables, verify that the univariate Gaussian distribution satisfies: $E[x] = \int_{-\infty}^{\infty}(N(x|u,\sigma^2)x)dx = u$. Next by differentiating both sides of the normalization condition:$\int_{-\infty}^{\infty} N(x|u,\sigma^2)dx = 1$ with respect to $\sigma^2$, verify that the Gaussian satisfies: $E[x^2] = \int_{-\infty}^{\infty}(N(x|u,\sigma^2)x^2)dx = u^2+\sigma^2$. Finally show that $var[x] = E[x^2]-E[x]^2 = \sigma^2$ holds

Solution: The first ask us to proves that the mean of a Gaussian distribution is given by $\mu$

$$E[x] = \int_{-\infty}^{\infty}(N(x|u,\sigma^2)x)dx = \int_{-\infty}^{\infty}(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}e^{-\frac{1}{2\sigma^2}(x-u)^2}x)dx = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(x-u)^2}x)dx$$

Now using the substitution method we will aply the following change $y = x - u$ and $dy = dx$

$\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2}(y+u))dy$

Distribuiting over $(y+u)$

$\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2}y + e^{-\frac{1}{2\sigma^2}(y)^2}u)dy$

Using the property that an integral could be splited in two integrals

$\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}(\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2}y)dy + \mu\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2})dy)$

Now, we will split the left integral in two integrals: one from $-\infty$ to 0 and the other from 0 to $\infty$

$\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}(\int_{-\infty}^{0}(e^{-\frac{1}{2\sigma^2}(y)^2}y)dy + (\int_{0}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2}y)dy + \mu\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2})dy)$

Now we will swap the boundaries of the first term using this property : $\int_a^b x = -\int_b^a x$ and we will change the boundaries sign by changing the sign of $y$

$\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}(-\int_{0}^{-\infty}(e^{-\frac{1}{2\sigma^2}(y)^2}y)dy + (\int_{0}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2}y)dy + \mu\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2})dy)$

$\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}(-\int_{0}^{\infty}(e^{-\frac{1}{2\sigma^2}(-y)^2}(-y))(-dy) + (\int_{0}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2}y)dy + \mu\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2})dy)$

$\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}(-\int_{0}^{\infty}(e^{-\frac{1}{2\sigma^2}(-y)^2}(y))(dy) + (\int_{0}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2}y)dy + \mu\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2})dy)$

The first and the second terms are the same but one is the oposite of the other, so it is 0

$\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}(\mu\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2})dy)$

The integral is the one of the last exercise and we know the result:

$\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(y)^2})dy = (2\pi\sigma^2)^{\frac{1}{2}}$ So, if we replace in our equation:

$\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}(\mu(2\pi\sigma^2)^{\frac{1}{2}}) = \mu$ and we prove the statement.

For the $\sigma^2$case we have:

$\int_{-\infty}^{\infty}(N(x|u,\sigma^2))dx = \int_{-\infty}^{\infty}(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}e^{-\frac{1}{2\sigma^2}(x-u)^2})dx = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(x-u)^2})dx = 1$

$\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(x-u)^2})dx = (2\pi\sigma^2)^{\frac{1}{2}}$ Now differentitating both sides with respect to $\sigma^2$:

$(2\sigma^2)^{-2}4\sigma\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(x-u)^2})(x-u)^2dx = (2\pi\sigma^2)^{-\frac{1}{2}}\sigma$ and then re-arranging to obatin:

$(\frac{1}{2\pi\sigma^2})^{\frac{1}{2}}\int_{-\infty}^{\infty}(e^{-\frac{1}{2\sigma^2}(x-u)^2})(x-u)^2dx = \sigma^2$ which is:

$E[(x-u)^2] = var[x] = \sigma^2$ If we expand the left side:

$E[x^2 + \mu^2 - 2x\mu] = E[x^2] + \mu^2 - 2\mu E[x] = E[x^2] - E[x] = \sigma^2$

**1.9**

Show that the mode (i.e. the maximum) of the Gaussian distribution $N(x|u, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ is given by $\mu$. Similarly, show that the mode of the multivariate Gaussian $N(\boldsymbol{x}|\boldsymbol{\mu}, \sum) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\sum|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \sum^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$ is given by $\boldsymbol{\mu}$

Solution:To find the mode, we just need to differentiate with respect to x:

$$\frac{d}{dx} N(x|u, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} (-\frac{1}{2\sigma^2}(x-\mu)^2)'$$

Using the Chain rule...

$= N(x|u, \sigma^2)(\frac{-(x-\mu)}{\sigma^2})$

Setting this to zero...

$0 = N(x|u, \sigma^2)(\frac{-(x-\mu)}{\sigma^2}) \Leftrightarrow x = \mu$

Now for the multivariate case we will use some properties of The Matrix CookBook.

$N(\boldsymbol{x}|\boldsymbol{\mu}, \sum) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\sum|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \sum^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$

$\frac{\partial}{\partial \boldsymbol{x}} N(\boldsymbol{x}|\boldsymbol{\mu}, \sum) = N(\boldsymbol{x}|\boldsymbol{\mu}, \sum)(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \sum^{-1} (\boldsymbol{x}-\boldsymbol{\mu}))'$

As in the first case the first term remains equal and then using chain rule and property 85 from

The Matrix CookBook which is: Assuming $\boldsymbol{W}$ symmetric

$\frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{x}-\boldsymbol{\mu})^T W (\boldsymbol{x}-\boldsymbol{\mu}) = 2\boldsymbol{W}(\boldsymbol{x}-\boldsymbol{\mu})$

Now applying this property...

$\frac{\partial}{\partial \boldsymbol{x}} N(\boldsymbol{x}|\boldsymbol{\mu}, \sum) = N(\boldsymbol{x}|\boldsymbol{\mu}, \sum)(\sum^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$

Setting this to zero...

$0 = N(\boldsymbol{x}|\boldsymbol{\mu}, \sum)(\sum^{-1}(\boldsymbol{x}-\boldsymbol{\mu})) \Leftrightarrow \boldsymbol{x} = \boldsymbol{\mu}$

**1.10**

Suppose that the two variables $x$ and $z$ are statistically independent. Show that the mean and variance of their sums satisfies:

$E[x + z] = E[x] + E[z]$

$var[x + z] = var[x] + var[z]$

I will prove it for the continuous case, but it is analogous for the discrete one.

Solution: Since $x$ and $z$ are independent, their joint distribution is $p(x, z) = p(x, z)$. Now using the definition of expectation ($E[x] = \int xp(x)dx$)

$E[x + z] = \int \int (x + z)p(x)p(z)dxdz$

Doing some algebra...

$= \int \int (x)p(x)p(z) + (z)p(x)p(z)dxdz$

Using integrals properties...

$= \int \int (x)p(x)p(z)dxdz + \int \int (z)p(x)p(z)dxdz$

$= \int (x)p(x) \int p(z)dzdx + \int (z)p(z) \int p(x)dxdz$

As we know, because $p(x)$ is a probability density function it integrates 1. So...

$= \int (x)p(x)dx + \int (z)p(z)dz$

Finally, we have the definition of expectation.

$= E[x] + E[z]$

For the variance case we have two solutions. The easy one and the tricky one. First I will show you the easy one.

Remember how to compute the variance: $var[x] = E[x^2] - E[x]^2$

$var[x + z] = E[(x + z)^2] - E[x + z]^2 = E[x^2 + 2xz + z^2] - (E[x] + E[z])^2$

Using expectation properties...

$= E[x^2] + 2E[xz] + E[z^2] - (E[x]^2 + 2E[x]E[z] + E[z]^2)$

Because $x$ and $z$ are independent $E[xz] = E[x]E[z]$

$= E[x^2] + 2E[x]E[z] + E[z^2] - E[x]^2 - 2E[x]E[z] - E[z]^2$

$= E[x^2] - E[x]^2 + E[z^2] - E[z]^2 = var[x] + var[z]$

The other solution is using integrals. Remember that $var[x] = \int (x - E[x])^2 p(x) dx$

In our example we have:

$var[x + z] = \int\int (x + z - E[x + z])^2 p(z)p(x)dxdz$

$(x + z - E[x + z])^2 = (x - E[x])^2 + (z - E[z])^2 + 2(x - E[x])(z - E[z])$

$= \int\int ((x - E[x])^2 + (z - E[z])^2 + 2(x - E[x])(z - E[z]))p(z)p(x)dxdz$

$= \int\int ((x - E[x])^2 p(z)p(x) + (z - E[z])^2 p(z)p(x) + 2(x - E[x])(z - E[z])p(z)p(x))dxdz$

$= \int\int (x-E[x])^2 p(z)p(x)dxdz + \int\int (z-E[z])^2 p(z)p(x)dxdz + \int\int 2(x-E[x])(z-E[z])p(z)p(x)dxdz$

Rearranging...

$= \int (x-E[x])^2 p(x) \int p(z)dzdx + \int (z-E[z])^2 p(z) \int p(x)dxdz + \int\int 2(x-E[x])(z-E[z])p(z)p(x)dxdz$

As we know, because $p(x)$ is a probability density function it integrates 1. So...

$= \int (x - E[x])^2 p(x)dx + \int (z - E[z])^2 p(z)dz + \int\int 2(x - E[x])(z - E[z])p(z)p(x)dxdz$

$= var[x] + var[z] + \int\int 2(x - E[x])(z - E[z])p(z)p(x)dxdz$

And the last term...

$= \int\int 2(x - E[x])(z - E[z])p(z)p(x)dxdz = 2\int (z - E[z])p(z) \int (x - E[x])p(x)dxdz$

$= 2\int (zp(z) - E[z]p(z)) \int (xp(x) - E[x]p(x))dxdz =$

$= 2\int (zp(z) - E[z]p(z))(\int xp(x)dx - E[x] \int p(x))dx)dz =$

$= 2\int (zp(z) - E[z]p(z))dz(E[x] - E[x]) =$

$= 2\int zp(z)dz - E[z] \int p(z)dz(E[x] - E[x]) =$

$= 2(E[z] - E[z])(E[x] - E[x]) = 0$

Finally...

$var[x + z] = var[x] + var[z] + 0 = var[x] + var[z]$

**1.11**

By setting the derivatives of the log likelihood function with respect to $\mu$ and $\sigma^2$ equal to zero, verify the results:

$\mu_{ML} = \frac{1}{N}\sum_n^N x_n, \sigma^2_{ML} = \frac{1}{N}\sum_n^N (x_n - \mu_{ML})^2$

Solution: We know that the joint probability of two independent events is given by the product of the marginal probabilities for each event separately.If we have a data set $x$ that is i.i.d we can therefore write the probability of the data set, given $\mu$ and $\sigma^2$, in the form:

$p(x|\mu, \sigma^2) = \prod_{n=1}^N N(x_n|\mu, \sigma^2)(1)$

Now we should determine the values for the unknown parameters $\mu$ and $\sigma^2$ in the Gaussian by maximizing the likelihood function (1). We would maximize the log of the likelihood function because it is more convenient. Now, we can start:

$ln(p(x|\mu, \sigma^2)) = ln(\prod_{n=1}^N N(x_n|\mu, \sigma^2)) =$

Using properties from the Natural logarithm.

$= -\frac{1}{2\sigma^2}\sum_n^N (x_n - u)^2 - \frac{N}{2}ln(\sigma^2) - \frac{N}{2}ln(2\pi)$

Now we to maximize with respect to $\mu$ we differentiate and then set to 0 and rearrange...

The last two terms do not have $\mu$ so they vanished.

$0 = \frac{1}{\sigma^2}\sum_n^N (x_n - u) = \frac{1}{\sigma^2}(\sum_n^N (x_n) - N\mu)$

$\frac{1}{\sigma^2}(N\mu) = \frac{1}{\sigma^2}(\sum_n^N (x_n))$

$N\mu = \sum_n^N (x_n)$

$\mu_{ML} = \frac{\sum_n^N (x_n)}{N}$

Let's go with the variance...

$= -\frac{1}{2\sigma^2}\sum_n^N (x_n - u)^2 - \frac{N}{2}ln(\sigma^2) - \frac{N}{2}ln(2\pi)$

Now we to maximize with respect to $\sigma^2$ we differentiate and then set to 0 and rearrange...

$0 = \frac{1}{2(\sigma^2)^2}\sum_n^N (x_n - u)^2 - \frac{N}{2}\frac{1}{(\sigma^2)}$

$\frac{N}{2}\frac{1}{(\sigma^2)} = \frac{1}{2(\sigma^2)^2}\sum_n^N (x_n - u)^2$

$N = \frac{1}{(\sigma^2)}\sum_n^N (x_n - u)^2$

$\sigma^2_{ML} = \frac{1}{N}\sum_n^N (x_n - u_{ML})^2$

**1.12**

Using the results: $E[x] = \int_{\infty}^{-\infty} N(x|u,\sigma^2)xdx = \mu$ and $E[x^2] = \int_{\infty}^{-\infty} N(x|u,\sigma^2)x^2dx = \mu^2 + \sigma^2$ show that $E[x_n x_m] = \mu^2 + I_{nm}\sigma^2$ where $x_n$ and $x_m$ denote data points sampled from a Gaussian distribution with mean $\mu$ and variance $\sigma^2$, and $I_{nm}$ = 1 if $n = m$ and $I_{nm} = 0$ otherwise. Hence prove the results $E[\mu_{ML}] = \mu$ and $E[\sigma^2_{ML}] = (\frac{N-1}{N})\sigma^2$

Solution: The case where $n = m$ is almost trivial. First $x_n x_m = x_n^2$ and $I_{nm} = 1$

So, using the result given in the statement: $E[x_n^2] = \mu + I_{nm}\sigma^2 = \mu + \sigma^2$.

For the case where $n \neq m$: We know that $x_n and x_m$ are independet so using expectation properties we can re-write $E[x_n x_m] = E[x_n]E[x_m]$

So: $E[x_n x_m] = E[x_n]E[x_m] = \mu * \mu = \mu^2$ which is what we want to prove, because $I_{nm} = 0$

Finally: $E[\mu_{ML}] = E[\frac{\sum_n^N (x_n)}{N}] == \frac{N\mu}{N} = \mu$

For the $\sigma^2 case$ :

$\mathrm{E}[\sigma^2_{ML}] = E[\frac{\sum_n^N (x_n - \mu_{ML})^2}{N}] = \frac{\sum_n^N E[(x_n - \mu_{ML})^2]}{N} = \frac{\sum_n^N E[(x_n^2 - 2\mu_{ML}x_n + \mu^2_{ML})]}{N} = \frac{\sum_n^N E[x_n^2] - 2E[\mu_{ML}x_n] + E[\mu^2_{ML}]}{N}$.

Now we use:

$E[x_n^2] = \mu^2 + \sigma^2$

$-2E[\mu_{ML}x_n] = -2\frac{\sum_m^M x_n x_m}{N}$. Using the properties above and taking into account when $n = m$ and $n \neq m$:

$= -2\frac{(N\mu^2 + \sigma^2)}{N} = -2\frac{N(\mu^2 + \frac{\sigma^2}{N})}{N} = -2(\mu^2 + \frac{\sigma^2}{N})$

$E[\mu_{ML}^2] = \frac{\sum_m^M \sum_l^L x_m x_l}{N^2}$. Using the properties above and taking into account when $l = m$ and $l \neq m$:

$= \frac{N^2\mu + N\sigma^2}{N^2} = \frac{N^2(\mu + \frac{\sigma^2)}{N}}{N^2} = (\mu + \frac{\sigma^2)}{N}$

Finally replacing above:

$= \frac{\sum_n^N (\mu^2+\sigma^2) - 2(\mu^2 + \frac{\sigma^2}{N}) + (\mu + \frac{\sigma^2)}{N}}{N} = \frac{N(\mu^2+\sigma^2) - (\mu^2 + \frac{\sigma^2}{N})}{N} = \mu^2 + \sigma^2 - \mu^2 - \frac{\sigma^2}{N} = \sigma^2 - \frac{\sigma^2}{N} = \sigma^2(1 - \frac{1}{N}) = \sigma^2(\frac{N-1}{N})$

**1.13**

Suppose that the variance of a Gaussian is estimated using the result $\sigma_{ML}^2 = \frac{1}{N}\sum_n^N (x_n - \mu_{ML})^2$ but with the maximum likelihood estimate $\mu_{ML}$ replaced with the true value $\mu$ of the mean. Show that this estimator has the property that its expectation is given by the true variance $\sigma^2$

Solution:

$E[\sigma_{ML}^2] = E[\frac{1}{N}\sum_n^N (x_n - \mu)^2] = \frac{\sum_n^N E[(x_n - \mu^2)]}{N} = \frac{\sum_n^N E[(x_n^2 - 2x_n\mu + \mu_2)]}{N} = \frac{N(\mu^2 + \sigma^2 - 2\mu^2 + \mu^2)}{N} = \sigma^2$

**1.14**

Show that an arbitrary square matrix with elements $w_{ij} can be written$ in the form $w_{ij} = w_{ij}^S + w_{ij}^A$ where $w_{ij}^S$ and $w_{ij}^A$ are synnetric and anti-symmetric matrices, respectively, satisfying $w_{ij}^S = w_{ji}^S$ and $w_{ij}^S = -w_{ji}^A$ for all $i$ and $j$. Now consider the second order termin in a higher order polynomial in D dimensions, given by

$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j$

Show that:

$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_i x_j$

so that the contribution from the anti-symmetric matrix vanishes. We therefore see that, without loss of generality the matrix of coefficientes $w_{ij}$ can be chosen to be symmetric, and so not all of the $D^2$ elements of this matrix can be chosen independently. Show that the number of independent parammeters in the matrix $w_{ij}^S$ is given by $\frac{D(D+1)}{2}$

Solution: There is an important property from matrix that will be useful for this excercise.

Any square matrix can be expressed as the sum of symmetric and anti-symmetric parts:

$A = \frac{1}{2}(A + A^t) + \frac{1}{2}(A - A^t)$ where $(A + A^t)$ is symmetric and $(A - A^t)$ anti-symmetric

Now replacing in our formula:

$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}^S + w_{ij}^A)x_i x_j$

Using our property...

$\sum_{i=1}^{D}\sum_{j=1}^{D}(\frac{1}{2}(w_{ij}+w_{ji})+\frac{1}{2}(w_{ij}-w_{ji}))x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D}(\frac{1}{2}(w_{ij}+w_{ji})x_i x_j+\sum_{i=1}^{D}\sum_{j=1}^{D}\frac{1}{2}(w_{ij}-w_{ji}))x_i x_j$

If we check tha last term...

$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D}\frac{1}{2}(w_{ij}-w_{ji}))x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D}\frac{1}{2}w_{ij}x_i x_j - \sum_{i=1}^{D}\sum_{j=1}^{D}\frac{1}{2}w_{ji}x_i x_j = 0$

Finally

$$\sum_{i=1}^{D}\sum_{j=1}^{D}(\tfrac{1}{2}(w_{ij}+w_{ji})+\tfrac{1}{2}(w_{ij}-w_{ji}))x_ix_j = \sum_{i=1}^{D}\sum_{j=1}^{D}(\tfrac{1}{2}(w_{ij}+w_{ji})x_ix_j = \sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{S}x_ix_j$$

In our matrix we have $D^2$ elements. If we subtract the diagonal which is unique, we have $D^2 - D$. If know we divide by two, we get the number of independent parameters $\frac{D^2-D}{2}$. Finally we add again the numbers in the diagonal: $\frac{D^2-D}{2} + D = \frac{D^2}{2} - \frac{D}{2} + \frac{2D}{2} = \frac{D^2+D}{2} = \frac{D(D+1)}{2}$

**1.15**

**1.16**

**1.17**

The gamma function is defined by:

$\Gamma(x) \equiv \int_0^\infty \mu^{x-1}e^{-\mu}d\mu$

Using integration by parts, prove the relation $\Gamma(x+1) = x\Gamma(x)$. Show also that $\Gamma(1) = 1$ and hence that $\Gamma(x+1) = x!$ when $x$ is an integer.

Solution: $\Gamma(x) \equiv \int_0^\infty \mu^{x-1}e^{-\mu}d\mu$ and I want to prove:

$\Gamma(x+1) = x\Gamma(x)$

$\Gamma(x+1) = \int_0^\infty \mu^{x}e^{-\mu}d\mu$

Using integration by parts...

$w = u^x$ and $dw = \mu^{x-1}x$. $dv = e^{-\mu}$ and $v = -e^{-\mu}$

$\Gamma(x+1) = -u^xe^{-\mu}\Big|_0^\infty + \int_0^\infty e^{-\mu}\mu^{x-1}xd\mu = 0 + x\Gamma(x)$

If $x = 1$

$\Gamma(1) = \int_0^\infty \mu^0 e^{-\mu} d\mu = \int_0^\infty e^{-\mu} d\mu = 0 - (-1) = 1$

To prove $\Gamma(x+1) = x!$ we should use induction.

We know that for $x = 0$ we have $\Gamma(0+1) = \Gamma(0+1) == 1 = 0!$. Our hypothesis is $\Gamma(x+1) = x!$

For $x + 1$...

$\Gamma(x+2) = x+1!$?

We will use the first property we proved and the hypothesis. $\Gamma(x+2) = (x+1)\Gamma(x+1) = (x+1)(x!) = x+1!$

**1.18**

We can use the result $I = (2\pi\sigma^2)^{\frac{1}{2}}$ to derive an expression for the surface area $S_D$, and the volume $V_D$, of a sphere of unit radius in D dimensions. To di this, consider the following result, which is obtained by transforming from cartesian to polar coordinates:

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} dx_i = S_D \int_0^\infty e^{-r^2} r^{D-1} dr$$

Using the definition of the Gamma function, together with the last equation, evaluate both sides of this equation, and hence show that:

$$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2})}$$

Next, by integrating with respect to radius from 0 to 1, show that the volume of the unit sphere in D dimensions is given by:

$$V_D = \frac{S_D}{D}$$

Finally. use the result $\Gamma(1) = 1$ and $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$ to show that $S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2})}$ and $V_D = \frac{S_D}{D}$ reduce to the usual expression for D = 2 and D = 3

Solution: First we will make the change of variables $\mu = r^2$. So...

$$\frac{1}{2}\prod_{i=1}^{D}\int_{-\infty}^{\infty} e^{-x_i^2} dx_i = \frac{1}{2} S_D \int_0^{\infty} e^{-\mu}(\mu^{\frac{D}{2}-1}) dr$$

Where the last term corresponds to a Gamma Function...

$$= \frac{1}{2} S_D \Gamma(\frac{D}{2})$$

Using excerise 1.7

$$I = \int_{-\infty}^{\infty} e^{-x_i^2} dx_i$$

$$I^2 = \int_{-\infty}^{\infty} -\infty^{\infty} e^{-x_i^2} e^{-y_i^2} dx_i dy_i$$

$$x = cos(\theta) \text{ and } y = sen(\theta)$$

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-cos(\theta)^2 r^2 + sen(\theta)^2 r^2} r dr d\theta$$

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2(cos(\theta)^2 + sen(\theta)^2)} r dr d\theta$$

Using the identity :$(cos(\theta)^2 + sen(\theta)^2 = 1$

$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta$

$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta$

Integrating using change of variable...

$\mu = r^2$ and $d\mu = 2rdr$ so $rdr = \frac{1}{2}du$

$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-u} \frac{1}{2} du d\theta$

$I^2 = \int_0^{2\pi} \frac{1}{2} d\theta$

$I^2 = \pi$

$I = \sqrt{\pi}$

Finally:

$\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} dx_i = \prod_{i=1}^{D} \sqrt{\pi} = \pi^{\frac{D}{2}}$

$S_D \Gamma(\frac{D}{2}) \frac{1}{2} = \pi^{\frac{D}{2}}$

$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2})}$

$V_D = \int_0 1 r^{D-1} dr = \frac{S_D}{D}$