W15 Problem Sheet Explanation: Regression

Question 1

First, let's start at looking at the multivariate Gaussian distribution. The normal Gaussian pdf is given by:

$$P(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

This means, given that we know the parameters, μ and σ, we can work out the probability that a particular data point came from a Gaussian distribution with those parameters. Now, this is the case for if x is one-dimensional, but what if it's not? This is where we extend to use the multivariate Gaussian distribution.

When we work with data that are vectors, we know that our mean, $\boldsymbol{\mu}$, will be a vector of the same size and our variance becomes a covariance matrix, $\boldsymbol{\Sigma}$. The pdf for the multivariate Gaussian distribution is:

$$P(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{-1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

If you would like to understand how this relates to the univariate Gaussian distribution then this has an extensive explanation.

Regression is the process of adding extra information to prevent an overfitting of our data. So, once we have our log-likelihood we can add weights using the term $\boldsymbol{w}^T\boldsymbol{\Lambda}\boldsymbol{w}$. The aim is to find the weights, $\hat{\boldsymbol{w}}$, that will be the best to solve overfitting. When $\boldsymbol{\Lambda} = \boldsymbol{0}$, our estimate will be the same as our log-likelihood estimate. When we give $\boldsymbol{\Lambda}$ a value, then we penalise the weights with higher value more.

In the question we start with the log-likelihood function:

$$\mathcal{L}(\boldsymbol{w}) = \log P(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) - \frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Lambda}\boldsymbol{w}$$

The first term we will assume is the $\log$ of the multivariate Gaussian distribution, which is introduced in the second notebook. The second notebook also goes through the calculation of the derivative of this and its simplification, which we will just use for this solution.

The second term is the adjustment of the weights - we need to work out the derivative of this.

$$\frac{d}{d\boldsymbol{w}}\left(-\frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Lambda}\boldsymbol{w}\right)$$

We are trying to estimate $\boldsymbol{w}$, let's pretend that it has 2-dimensions and hence $\boldsymbol{w} = (w_1\, w_2)$.

$$\begin{align}
-\frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Lambda}\boldsymbol{w} &= -\frac{1}{2}\cdot\begin{pmatrix}w_1 & w_2\end{pmatrix}\cdot\begin{pmatrix}\Lambda_{11} & 0\\ 0 & \Lambda_{22}\end{pmatrix}\cdot\begin{pmatrix}w_1\\ w_2\end{pmatrix} \tag{1}\\
&= -\frac{1}{2}\cdot\begin{pmatrix}w_1 & w_2\end{pmatrix}\cdot\begin{pmatrix}\Lambda_{11}w_1\\ \Lambda_{22}w_2\end{pmatrix} \tag{2}\\
&= -\frac{1}{2}(\Lambda_{11}w_1^2+\Lambda_{22}w_2^2) \tag{3}
\end{align}$$

Now, we need to differentiate this with respect to $\boldsymbol{w}$. This is slightly different to normal differentiation and $\boldsymbol{w}$ is a vector, not a singular matrix. To differentiate by a vector we calculate the partial derivatives with respect to each element in the vector, and put them altogether in a vector.

Let's define f to be:

$$f = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Lambda}\boldsymbol{w}$$

Then to calculate the derivative w.r.t $\boldsymbol{w}$ we do:

$$\begin{align}
\frac{df}{d\boldsymbol{w}} &= \begin{pmatrix}\frac{\partial f}{\partial w_1}\\ \frac{\partial f}{\partial w_2}\end{pmatrix} \tag{4}\\
&= \begin{pmatrix}-\Lambda_{11}w_1\\ -\Lambda_{22}w_2\end{pmatrix} \tag{5}\\
&= -\begin{pmatrix}\Lambda_{11} & 0\\ 0 & \Lambda_{22}\end{pmatrix}\cdot\begin{pmatrix}w_1\\ w_2\end{pmatrix} \tag{6}\\
&= -\boldsymbol{\Lambda}\boldsymbol{w} \tag{7}
\end{align}$$

Now, here we only looked at if $\boldsymbol{w}$ had 2-dimensions, but this will still hold no matter how many dimensions $\boldsymbol{w}$ has. You can check this if you want, or look at the workings in the solution sheet (this proves it for any dimension).

So, now we have calculated the derivative of our extra term, we have our full derivative of our log-likelihood (first part comes from the derivation in in the second notebook):

$$\frac{d\mathcal{L}(\boldsymbol{w})}{d\boldsymbol{w}} = \frac{1}{\sigma^2}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{\Lambda}\boldsymbol{w}$$

Let's set the derivative equal to 0 and solve for an estimate of $\hat{\boldsymbol{w}}$:

$$\begin{aligned}
& \mathbf{0} = \frac{1}{\sigma^2}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}}) - \boldsymbol{\Lambda}\hat{\boldsymbol{w}} && (8)\\
\iff\ & \mathbf{0} = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y} - \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{\Lambda}\hat{\boldsymbol{w}} && (9)\\
\iff\ & \mathbf{0} = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{w}} - \sigma^2\boldsymbol{\Lambda}\hat{\boldsymbol{w}} && (10)\\
\iff\ & \mathbf{0} = \boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})\hat{\boldsymbol{w}} && (11)\\
\iff\ & (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})\hat{\boldsymbol{w}} = \boldsymbol{X}^T\boldsymbol{y} && (12)\\
\iff\ & (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})^{-1}(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{y} && (13)\\
\iff\ & \hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{y} && (14)
\end{aligned}$$

And now we have our estimate for $\boldsymbol{w}$.

Question 2

Now, let's use the equation we derived in question 1 to get a real solution. We have been given the data:

$$\boldsymbol{x} = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} -6.2 \\ -2.6 \\ 0.5 \\ 2.7 \\ 5.7 \end{pmatrix}$$

And we have also been given that the data are related through the equation:

$$y = w_0 + w_1 x$$

We assume that we have not been given data from the true function and there has been some added noise, hence we cannot substitute our values for x and y and use simultaneous equations to find w_0 and w_1. This is where we apply regression.

The reason we add a column of 1s to our $\boldsymbol{x}$ data is so we can write the equation above in matrix form as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

And hence:

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{w}^T$$

So, to find our values for w_0 and w_1 all we need to do in substitute our values into the equation and simplify. Let's start small and build it up:

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{X} &= \begin{pmatrix}1 & 1 & 1 & 1 & 1\\ -2 & -1 & 0 & 1 & 2\end{pmatrix}\begin{pmatrix}1 & -2\\ 1 & -1\\ 1 & 0\\ 1 & 1\\ 1 & 2\end{pmatrix} && (15)\\ &= \begin{pmatrix}5 & 0\\ 0 & 10\end{pmatrix} && (16)\end{aligned}$$

So, next we will bring in $\sigma^2\boldsymbol{\Lambda}$:

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda} &= \begin{pmatrix}5 & 0\\ 0 & 10\end{pmatrix} + (1^2)\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix} && (17)\\ &= \begin{pmatrix}7 & 0\\ 0 & 12\end{pmatrix} && (18)\end{aligned}$$

Now, we need to find the inverse of this matrix. The formula for working out the inverse of a 2x2 matrix is:

$$\begin{pmatrix}a & b\\ c & d\end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix}d & -b\\ -c & a\end{pmatrix}$$

So, our inverse is:

$$(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})^{-1} = \frac{1}{84}\begin{pmatrix}12 & 0\\ 0 & 7\end{pmatrix} \tag{19}$$

Now all we have left is to put it all together:

$$\begin{aligned}\hat{\boldsymbol{w}} &= (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{y} && (20)\\ &= \frac{1}{84}\begin{pmatrix}12 & 0\\ 0 & 7\end{pmatrix}\begin{pmatrix}1 & 1 & 1 & 1 & 1\\ -2 & -1 & 0 & 1 & 2\end{pmatrix}\begin{pmatrix}-6.2\\ -2.6\\ 0.5\\ 2.7\\ 5.7\end{pmatrix} && (21)\\ &= \frac{1}{84}\begin{pmatrix}12 & 0\\ 0 & 7\end{pmatrix}\begin{pmatrix}0.1\\ 29.1\end{pmatrix} \qquad = \frac{1}{84}\begin{pmatrix}1.2\\ 203.7\end{pmatrix} && (22)\\ &= \begin{pmatrix}0.0143\\ 2.425\end{pmatrix} && (23)\end{aligned}$$

Which gives us our estimates that the true function is:

$$y = 0.0143 + 2.425x$$