Machine learning : Sheet 2

Author : Djordje Zivanovic

1. On the lectures we got the following equation:

$$\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{2}\log(2\pi\sigma^2)$$

   By setting derivative over σ to 0, we get the following:

$$\begin{aligned}\frac{\partial \mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma)}{\partial \sigma} &= \frac{-2\sigma^{-3}}{2}(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{2}\frac{2\pi 2\sigma}{2\pi\sigma^2} \\ &= -\sigma^{-3}(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{\sigma} = 0\end{aligned}$$

   Multiplying the last expression with σ^3, because variance cannot be 0, and dividing by N we get

$$\sigma^2 = \frac{(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y})}{N}$$

   Since optimal $\mathbf{w}$ is calculated by setting gradient to zero, we insert $\mathbf{w_{ML}}$ to the last equation and get the necessary σ.

2. By using vector multiplication rules we can simplify the expression:

$$\begin{aligned}\mathcal{L}_{ridge}(\mathbf{w}, b) &= (\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y}) + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}b\mathbf{1} - \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + b\mathbf{1}^{\mathrm{T}}\mathbf{X}\mathbf{w} + b^2N - b\mathbf{1}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} - by^{\mathrm{T}}\mathbf{1} + \mathbf{y}^{\mathrm{T}}\mathbf{y} + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w}\end{aligned}$$

   Deriving $\mathcal{L}_{ridge}(\mathbf{w}, b)$ per b and equaling to zero we get:

$$\begin{aligned}\frac{\partial \mathcal{L}_{ridge}(\mathbf{w}, b)}{\partial b} &= 0 + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{1} - 0 + \mathbf{1}^{\mathrm{T}}\mathbf{X}\mathbf{w} + 2bN - \mathbf{1}^{\mathrm{T}}\mathbf{y} - 0 - y^{\mathrm{T}}\mathbf{1} + 0 + 0 \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{1} + \mathbf{1}^{\mathrm{T}}\mathbf{X}\mathbf{w} + 2bN - 2\mathbf{y}^{\mathrm{T}}\mathbf{1} = 0\end{aligned}$$

   Using the property that product of $\mathbf{A}^{\mathrm{T}}B = B^{\mathrm{T}}A$ when **A,B** are column vectors we got the simplification of the last equation. Further, $\mathbf{X}^{\mathrm{T}}$ when multiplied by $\mathbf{1}$ from column matrix whose cells are the sum of the same features. From the beginning condition that sum for every possible set of features is 0, the first two addends will give 0 as a result (associativity of matrix multiplication). Thus, we get $b = \mathbf{y}^{\mathrm{T}}$ which is equivalent to $\hat{b} = \frac{1}{N}\sum_{i=1}^{N} y_i$.
   Deriving $\mathcal{L}_{ridge}(\mathbf{w}, b)$ per $\mathbf{w}$ and equaling to zero we get:

$$\begin{aligned}\frac{\partial \mathcal{L}_{ridge}(\mathbf{w}, b)}{\partial \mathbf{w}} &= 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} + 0 - \mathbf{X}^{\mathrm{T}}\mathbf{y} + 0 + 0 - 0 - \mathbf{X}^{\mathrm{T}}\mathbf{y} \ - 0 + 0 + 2\lambda\mathbf{w} \\ &= 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\lambda\mathbf{w} = 0\end{aligned}$$

   In the previous derivation we simplified some expression (set them to zero) using the same observation $\mathbf{X}^{\mathrm{T}}\mathbf{1} = 0$. Using distributivity of matrix multiplication

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_D)\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

   Multiplying the both sides by $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_D)$ we get the $\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_D)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
   Consequence of centering $\mathbf{y}$ gives us that bias is unnecessary in this case (we would need neither to calculate nor to use it).

3. 1. Using properties of expectation:

$$\begin{aligned}\mathbb{E}(\hat{\mathbf{w}}_{\mathrm{LS}}(\mathcal{D})|\mathcal{D}) &= \mathbb{E}((\mathbf{X}^{\mathrm{T}}X)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}|\mathcal{D}) \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbb{E}(\mathbf{y}|\mathcal{D}) \quad (1) \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}^* \quad (2) \\ &= I\mathbf{w}^* = \mathbf{w}^*\end{aligned}$$

      The equation 1is correct by using property of expected value that it is possible to pull out constant from the parenthesis. The equation 2 is correct because it is stated that $\mathbb{E}(\mathbf{y}|\mathcal{D}) = \mathbf{X}\mathbf{w}^*$ in the text of the task.

2.

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[||\hat{\mathbf{w}}(\mathcal{D})-\mathbf{w}^*||^2\right]=\mathop{\mathbb{E}}_{\mathcal{D}}\left[(\hat{\mathbf{w}}(\mathcal{D})-\mathbf{w}^*)^{\mathrm{T}}(\hat{\mathbf{w}}(\mathcal{D})-\mathbf{w}^*)\right] \tag{3}$$

$$=\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})^{\mathrm{T}}\hat{\mathbf{w}}(\mathcal{D})-(\mathbf{w}^*)^{\mathrm{T}}\hat{\mathbf{w}}(\mathcal{D})-\hat{\mathbf{w}}(\mathcal{D})^{\mathrm{T}}\mathbf{w}^*+(\mathbf{w}^*)^{\mathrm{T}}\mathbf{w}^*\right] \tag{4}$$

$$=\mathop{\mathbb{E}}_{\mathcal{D}}\left[(\hat{\mathbf{w}}(\mathcal{D}))^{\mathrm{T}}\hat{\mathbf{w}}(\mathcal{D})\right]-(\mathbf{w}^*)^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]-\mathop{\mathbb{E}}_{\mathcal{D}}\left[(\hat{\mathbf{w}}(\mathcal{D}))^{\mathrm{T}}\right]\mathbf{w}^*+(\mathbf{w}^*)^{\mathrm{T}}\mathbf{w}^* \tag{5}$$

$$=\mathop{\mathbb{E}}_{\mathcal{D}}\left[(\hat{\mathbf{w}}(\mathcal{D}))^{\mathrm{T}}\hat{\mathbf{w}}(\mathcal{D})\right]-2(\mathbf{w}^*)^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]+(\mathbf{w}^*)^{\mathrm{T}}\mathbf{w}^* \tag{6}$$

The equation 3 is the other way to calculate Eucleadean norm using scalar product. The equation 4 is simple distributivity law. The equation 5 uses properties of mathematical expectation. The last equation 6 uses only a property that a scalar product of a column and a row vector is equal to their transposed product.

$$||\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]-\mathbf{w}^*||^2=(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]-\mathbf{w}^*)^{\mathrm{T}}(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]-\mathbf{w}^*) \tag{7}$$

$$=(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right])^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]-(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right])^{\mathrm{T}}\mathbf{w}^*-(\mathbf{w}^*)^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]+(\mathbf{w}^*)^{\mathrm{T}}\mathbf{w}^* \tag{8}$$

$$=(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right])^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]-2(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right])^{\mathrm{T}}\mathbf{w}^*+(\mathbf{w}^*)^{\mathrm{T}}\mathbf{w}^* \tag{9}$$

The equation 7 is the other way to calculate Eucleadean norm using scalar product. The equation 8 is simple distributivity law. The last equation 9 uses only a property that a scalar product of a column and a row vector is equal to their transposed product.

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[||\hat{\mathbf{w}}(\mathcal{D})-\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]||^2\right]=\mathop{\mathbb{E}}_{\mathcal{D}}\left[\left(\hat{\mathbf{w}}(\mathcal{D})-\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]\right)^{\mathrm{T}}\left(\hat{\mathbf{w}}(\mathcal{D})-\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]\right)\right] \tag{10}$$

$$=\mathop{\mathbb{E}}_{\mathcal{D}}\left[(\hat{\mathbf{w}}(\mathcal{D}))^{\mathrm{T}}\hat{\mathbf{w}}(\mathcal{D})-\left(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]\right)^{\mathrm{T}}\hat{\mathbf{w}}(\mathcal{D})-(\hat{\mathbf{w}}(\mathcal{D}))^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]+\left(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]\right)^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]\right] \tag{11}$$

$$=\mathop{\mathbb{E}}_{\mathcal{D}}\left[(\hat{\mathbf{w}}(\mathcal{D}))^{\mathrm{T}}\hat{\mathbf{w}}(\mathcal{D})\right]-\left(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]\right)^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]-$$

$$-\mathop{\mathbb{E}}_{\mathcal{D}}\left[(\hat{\mathbf{w}}(\mathcal{D}))^{\mathrm{T}}\right]\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]+\left(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]\right)^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right] \tag{12}$$

$$=\mathop{\mathbb{E}}_{\mathcal{D}}\left[(\hat{\mathbf{w}}(\mathcal{D}))^{\mathrm{T}}\hat{\mathbf{w}}(\mathcal{D})\right]-\left(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right]\right)^{\mathrm{T}}\mathop{\mathbb{E}}_{\mathcal{D}}\left[\hat{\mathbf{w}}(\mathcal{D})\right] \tag{13}$$

The equation 10 is the other way to calculate Eucleadean norm using scalar product. The equation 11 is simple distributivity law. The equation 12 uses properties of mathematical expectation. The last equation 13 uses only a property that a scalar product of a column and a row vector is equal to their transposed product.
By adding equations 9 and 13 we get the expression from the equation 6.

4. 1. We can notice that

$$p(x,y|\theta)=p(y|,x,\theta)p(x|\theta)$$

Inserting corresponding probabilities we get the table as on the figure 1.

	$y=0$	$y=1$
$x=0$	$\theta_2(1-\theta_1)$	$(1-\theta_2)(1-\theta_1)$
$x=1$	$(1-\theta_2)\theta_1$	$\theta_2\theta 1$

Figure 1: Probabilities

2. If we denote with $\alpha_1,\alpha_2,\alpha_3,\alpha_4$, respectively, the number of events where tuple (x,y) is $(0,0),(0,1),(1,0),(1,1)$, we have the following:

$$MLE(p(x,y|\theta_1,\theta_2))=\theta_2^{\alpha_1}(1-\theta_1)^{\alpha_1}(1-\theta_2)^{\alpha_2}(1-\theta_1)^{\alpha_2}(1-\theta_2)^{\alpha_3}\theta_1^{\alpha_3}\theta_2^{\alpha_4}\theta_1^{\alpha_4}$$
$$=\theta_1^{\alpha_3+\alpha_4}\theta_2^{\alpha_1+\alpha 4}(1-\theta_1)^{\alpha_1+\alpha_2}(1-\theta_2)^{\alpha_2+\alpha_3}$$

By calculating derivatives $MLE(p(x,y|\theta_1,\theta_2$ per $\theta_1,\theta_2)$ and setting them to 0 we get:

$$\frac{\partial MLE(p(x,y|\theta_1,\theta_2))}{\partial\theta_1}=(\alpha_3+\alpha_4)\theta_1^{\alpha_3+\alpha_4-1}\theta_2^{\alpha_1+\alpha 4}(1-\theta_1)^{\alpha_1+\alpha_2}(1-\theta_2)^{\alpha_2+\alpha_3}$$
$$-(\alpha_1+\alpha_2)\theta_1^{\alpha_3+\alpha_4}\theta_2^{\alpha_1+\alpha 4}(1-\theta_1)^{\alpha_1+\alpha_2-1}(1-\theta_2)^{\alpha_2+\alpha_3}=0$$

By dividing both side of equation by $\theta_1^{\alpha_3+\alpha_4-1}\theta_2^{\alpha_1+\alpha 4}(1-\theta_1)^{\alpha_1+\alpha_2-1}(1-\theta_2)^{\alpha_2+\alpha_3}$ we get:

$$(\alpha_3+\alpha_4)(1-\theta_1)-(\alpha_1+\alpha_2)\theta_1=0$$

Thus $\theta_1=\frac{\alpha_3+\alpha_4}{\alpha_1+\alpha_2+\alpha_3+\alpha_4}$. Similarly $\theta_2=\frac{\alpha_1+\alpha_4}{\alpha_1+\alpha_2+\alpha_3+\alpha_4}$ Replacing the values from our dataset we get $\theta_1=\frac{4}{7}$ and $\theta_2=\frac{4}{7}$. The $p(\mathcal{D}|\hat{\theta},M_2)$ is given on figure 2. If we calculated MLE for $p(x|\theta)$ and $p(y|\theta)$ we would get the same probabilities as we can see that θ_1 is not dependent on θ_2 and vice versa in $p(x,y|\theta)$.

	$y=0$	$y=1$
$x=0$	$\frac{12}{49}$	$\frac{9}{49}$
$x=1$	$\frac{12}{49}$	$\frac{16}{49}$

Figure 2: Probabilities

3. Let us denote number of occurrences of each tuple x,y, $(0,0),(0,1),(1,0),(1,1)$ respectively, with $\alpha_1,\alpha_2,\alpha_3,\alpha_4$. Then we would like to maximize:

$$MLE=\theta_{0,0}^{\alpha_1}\theta_{0,1}^{\alpha_2}\theta_{1,0}^{\alpha_3}\theta_{1,1}^{\alpha_4}=\theta_{0,0}^{\alpha_1}\theta_{0,1}^{\alpha_2}\theta_{1,0}^{\alpha_3}(1-\theta_{0,0}-\theta_{0,1}-\theta_{1,0})^{\alpha_4}$$

Thus, let us change $A=\theta_{0,0},B=\theta_{0,1},C=\theta_{1,0}$. Further, due to maximization we need to have derivative equal to zero:

$$\frac{\partial MLE}{\partial A}=\alpha_1 A^{\alpha_1-1}B^{\alpha_2}C^{\alpha_3}(1-A-B-C)^{\alpha_4}-\alpha_4 A^{\alpha_1}B^{\alpha_2}C^{\alpha_3}(1-A-B-C)^{\alpha_4-1}=0$$

Dividing both sides of equation by $A^{\alpha_1-1}B^{\alpha_2}C^{\alpha_3}(1-A-B-C)^{\alpha_4-1}$ we get:

$$\alpha_1(1-A-B-C)-\alpha_4 A=0$$

$$(\alpha_1+\alpha_4)A+\alpha_1 B+\alpha_1 C=\alpha_1$$

Similarly:

$$(\alpha_2+\alpha_4)B+\alpha_2 A+\alpha_2 C=\alpha_2$$

$$(\alpha_3+\alpha_4)C+\alpha_3 A+\alpha_3 B=\alpha_3$$

The determinant is: $\begin{vmatrix}\alpha_1+\alpha_4 & \alpha_1 & \alpha_1\\ \alpha_2 & \alpha_2+\alpha_4 & \alpha_2\\ \alpha_3 & \alpha_3 & \alpha_3+\alpha_4\end{vmatrix}$ which is positive always (because the result is the sum of positive monomials). This means the system has a unique solution. If we try the solution:

$$(A,B,C)=\left(\frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3+\alpha_4},\frac{\alpha_2}{\alpha_1+\alpha_2+\alpha_3+\alpha_4}\frac{\alpha_3}{\alpha_1+\alpha_2+\alpha_3+\alpha_4}\right)$$

we can see this tuple satisfies solutions.

4. For the data set give in subtask 2. the 2-parameter model will be chosen (if we do all calculations manually), because the 4-parameter model overfits the data. IN generals case the calculations would have to be done, but problem of overfitting would limit the 4-parameter model ability to learn more widely.