Pattern Recognition and Machine Learning: Homework 2

Qingru Hu 2020012996

March 6, 2023

Problem 1

The distance $d(v, H)$ of any point $\boldsymbol{v} \in \mathbb{R}^d$ to H must satisfies:

$$\begin{cases} \min_{x\in H} \ \|\boldsymbol{x}-\boldsymbol{v}\|_2 \\ s.t. \quad \boldsymbol{w}^\top \boldsymbol{x} + b = 0 \end{cases}$$

where x is a point in the hyperplane H. And we can simplify the target function $\|x - v\|_2$ on the condition that $\boldsymbol{v}$ is a given point.

$$\begin{aligned} &\min \ \|\boldsymbol{x}-\boldsymbol{v}\|_2 \\ \sim &\min \ (\boldsymbol{x}-\boldsymbol{v})^2 \\ \sim &\min \ (\boldsymbol{x}^2 - 2\boldsymbol{x}\boldsymbol{v}) \\ = &\min \ (\boldsymbol{x}^\top x - 2\boldsymbol{x}^\top \boldsymbol{v}) \end{aligned}$$

The optimazation problem can be simplified as:

$$\begin{cases} \min_{x\in H} \ (\boldsymbol{x}^\top x - 2\boldsymbol{x}^\top \boldsymbol{v}) \\ s.t. \quad \boldsymbol{w}^\top \boldsymbol{x} + b = 0 \end{cases}$$

It can be solved by the Lagrange multiplier method, and its Lagrange function is as follows:

$$L(\boldsymbol{x}, \lambda) = \boldsymbol{x}^\top x - 2\boldsymbol{x}^\top \boldsymbol{v} + \lambda(\boldsymbol{w}^\top \boldsymbol{x} + b)$$

Finding the gradient with respect to x gives:

$$\frac{\partial L}{\partial \boldsymbol{x}} = 2\boldsymbol{x} - 2\boldsymbol{v} + \lambda \boldsymbol{w}$$

Set the gradient as zero and we can get the best solution is:

$$\boldsymbol{x}^* = \boldsymbol{v} - \frac{1}{2}\lambda \boldsymbol{w}$$

Substitute the above relation into the constraint can get the value of λ:

$$\lambda = \frac{2(\boldsymbol{w}^\top \boldsymbol{v} + \boldsymbol{b})}{\boldsymbol{w}^\top \boldsymbol{w}}$$

Therefore, the distance d of point $\boldsymbol{v}$ to the hyperplane H is:

$$d = |\lambda| \|\boldsymbol{w}\| = \frac{|\boldsymbol{w}^\top \boldsymbol{v} + \boldsymbol{b}|}{\|\boldsymbol{w}\|}$$

Problem 2

The loss function for the whole dataset is:

$$L(\boldsymbol{w}, b) = \frac{1}{2} \sum_{i=1}^{N} [(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - y_i]^2$$

In order to find the minimum of the loss function, we set derivatives wrt $\boldsymbol{w}$ and b to zero:

$$\sum_{i=1}^{N} (\boldsymbol{w}^{*\top} \boldsymbol{x}_i + b^* - y_i)\boldsymbol{x}_i = 0$$
$$\sum_{n=1}^{N} (\boldsymbol{w}^{*\top} \boldsymbol{x}_i + b^* - y_i) = 0$$

Given that:

$$\sum_{i=1}^{N} y_i = N_1 \frac{N}{N_1} - N_2 \frac{N}{N_2} = 0$$
$$\boldsymbol{m} = \frac{1}{N} \sum_{i=1}^{N} = \frac{1}{N}(N_1 \boldsymbol{m_1} + N_2 \boldsymbol{m}_2)$$

from $\sum_{n=1}^{N} (\boldsymbol{w}^{*\top} \boldsymbol{x}_i + b^* - y_i) = 0$, we can get that:

$$b^* = -\boldsymbol{w}^{*\top} \boldsymbol{m}$$

Given that:

$$\begin{aligned} S_W &= \sum_{i \in C_1} (\boldsymbol{x}_i - \boldsymbol{m}_1)(\boldsymbol{x}_i - \boldsymbol{m}_1)^\top + \sum_{i \in C_2} (\boldsymbol{x}_i - \boldsymbol{m}_2)(\boldsymbol{x}_i - \boldsymbol{m}_2)^\top \\ S_B &= (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top \\ b^* &= -\boldsymbol{w}^{*\top} \boldsymbol{m} \end{aligned}$$

from $\sum_{i=1}^{N} (\boldsymbol{w}^{*\top} \boldsymbol{x}_i + b^* - y_i)\boldsymbol{x}_i = 0$ we can get:

$$(S_W + \frac{N_1 N_2}{N} S_B)\boldsymbol{w}^* = N(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

Since $S_B \boldsymbol{w}^*$ is in the direction of $\boldsymbol{m}_1 - \boldsymbol{m_2}$, $S_W \boldsymbol{w}^*$ should also be in the direction of $\boldsymbol{m}_1 - \boldsymbol{m_2}$, that is:

$$\boldsymbol{w}^* \parallel S_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m_2})$$

It is consistent with the $\boldsymbol{w}$ in solution:

$$\max_{\omega} J_F(\omega) = \max_{\omega} \frac{\omega^T S_b \omega}{\omega^T S_w \omega}$$

Problem 2 continued on next page...

Problem 3

3.1

$$\sigma(x) + \sigma(-x) = \frac{1}{1+e^{-x}} + \frac{1}{1+e^{x}} = \frac{e^{x}}{1+e^{x}} + \frac{1}{1+e^{x}} = 1$$

3.2

$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}(1 - \frac{1}{1+e^{-x}}) = \sigma(x)(1-\sigma(x))$$

3.3

$$\begin{aligned}
\because \sigma(x) &= \frac{1}{1+e^{-x}} \\
\therefore e^{-x} &= \frac{1}{\sigma(x)} - 1 \\
\tanh(x) &= \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}} = \frac{2\sigma(x)-1}{1-2\sigma(x)+2\sigma(x)^2}
\end{aligned}$$

Bonus

Fully differentiate the loss function:

$$dL = 2(f(\boldsymbol{x}) - \hat{y}) \cdot f(\boldsymbol{x})(1 - f(\boldsymbol{x})) \cdot \mathrm{d}(\boldsymbol{w}^\top \boldsymbol{x} + b)$$

Find the partial derivatives:

$$\begin{aligned}
\because \quad & \mathrm{d}(\boldsymbol{w}^\top \boldsymbol{x} + b) = \boldsymbol{x}^\top \mathrm{d}\boldsymbol{w} + \mathrm{d}b \\
\therefore \quad & \frac{\partial L}{\partial \boldsymbol{w}} = 2(f(\boldsymbol{x}) - \hat{y}) \cdot f(\boldsymbol{x})(1 - f(\boldsymbol{x}))\boldsymbol{x} \\
& \frac{\partial L}{\partial b} = 2(f(\boldsymbol{x}) - \hat{y}) \cdot f(\boldsymbol{x})(1 - f(\boldsymbol{x}))
\end{aligned}$$

Problem 4

4.1

The 9-dimensional unit norm vector $\boldsymbol{w}^*$ is:

$$\boldsymbol{w}^* = [0.457, 0.320, 0.228, 0.089, 0.122, 0.687, 0.283, 0.247, 0.041]$$

The classification accuracy on the dataset is 96.14%.

Problem 4 continued on next page...

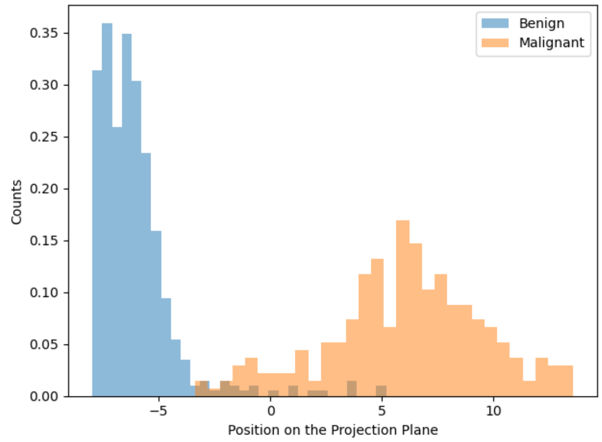

Figure 1: The histogram of the projection of dataset to the best plane

4.2

I use $\rho = 0.001$ for the regression, and iterate the regression for 10000 steps. I get a classification accuracy of 97.14%. The best weight $\boldsymbol{w}^*$ is:

$$\boldsymbol{w}^* = [0.423, 0.136, 0.203, 0.218, 0.026, 0.444, 0.225, 0.137, 0.441]$$

and the best bias b^* is:

$$b^* = -7.488$$

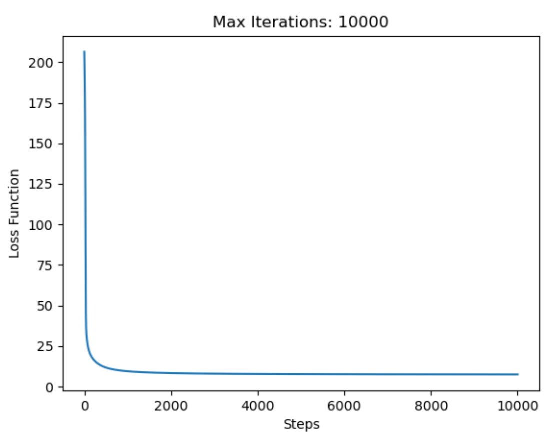

Figure 2: The loss function of the logistic regression over the iterations

4.3

The cosine between $\boldsymbol{w}_1^*$ and $\boldsymbol{w}_2^*$ is:

$$\cos < \boldsymbol{w}_1^*, \boldsymbol{w}_2^* >= \frac{\boldsymbol{w}_1^* \cdot \boldsymbol{w}_2^*}{\|\boldsymbol{w}_1^*\|\|\boldsymbol{w}_2^*\|} = 0.840$$

They have a difference in angle of 32.91°, which implies that they are similar in a way. It is because the $\boldsymbol{w}$ in both methods is projecting the feature vectors of the samples into a direction. The more indicative those features are , the higher weight they should have in both methods. Therefore, they are similar in a way. However, in logistic regression, after being projected with the $\boldsymbol{w}$, the feature vectors are acted upon by the sigmoid function to obtain the final output, so it is reseonable that these two $\boldsymbol{w}^*$ are not exactly the same.

Clump thickness is the most indicative feature that implies one gets breast cancer, because clump thickness has a relatively high weight in both $\boldsymbol{w}^*$s.