Homework 1

Zongdi Qiu
zq9ms

February 24, 2022

Question 1

When we know the true distribution of the two classes. We have been in the ideal state (when the sample data is large enough) the distribution of the two types of data. There will be some overlap between the two distributions. This means that both class of data will locate in same location. It is precisely because of the overlap of the two types of data that the selection of the classification boundary is more difficult. That is because we cannot find a perfect hyperplane to completely separate the two types of data According to the properties of the Bayesian predictor, for a binary classification problem, the location of the decision boundary should satisfy the properties of

$$p(+1|x) = p(-1|x) = \frac{1}{2}$$

This hyperplane divides space into two parts. One of the two parts must satisfy: the properties of

$$p(+1|x) < p(-1|x)$$

and the other part must satisfy

$$p(+1|x) > p(-1|x)$$

These two conditions can ensure the number of misclassified samples must be less than the number of correctly classified samples in ideal situation. If we move this dividing line, not only will the previous irreducible errors not be reduced, but more errors will be introduced. For example, as shown in the figure below, when we shift the dividing line to the left we get the new classification error shown as the orange region. In practice, according to the real distribution of data, there will be more samples belonging to -1 category in the orange area, while the new decision boundary classifies this part as 1. This will undoubtedly increase classification errors. Therefore, Bayesian method is the best predictor of decision boundary guaranteed by the properties of both sides of the boundary

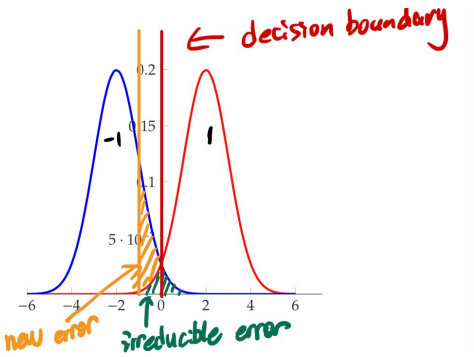

Figure 1: Question1

To be more precisely:
So if we're looking for the best way to divide between data that's classified as -1 and data that's classified as 1. We need to reduce expect errors. h(x) is the classifier and h(x) = -1,1

$$\begin{aligned} min(E[|h(x)-y||x]) &= \frac{1}{2} min[P(y=-1|x)*(h(x)+1)+P(y=1|x)(1-h(x))] \\ &= \frac{1}{2} min[P(y=-1|x)*h(x)+P(y=1|x)+p(y=-1|x)-P(y=1|x)*h(x)] \\ &\doteq min[h(x)*(P(y=-1|x)-P(y=1|x))] \end{aligned}$$

In order to minimize this term, it's easy to get:

$$h(x) = \begin{cases} -1 & 0 < P(y=-1|x) - P(y=1|x) \\ 1 & P(y=-1|x) - P(y=1|x) < 0 \end{cases}$$

This h(x) is equivalent to Bayes Predictor

Question 2 In my jupyter notebook

Question 3 In my jupyter notebook

Question 4

$$\begin{aligned} L(hw,s) &= \frac{1}{m}\sum_i^m log(1+e^{-y_i<w,x_i>}) \\ \frac{d}{dw}L(hw,s) &= \frac{1}{m}\sum_i^m \frac{1}{1+e^{-y_i<w,x_i>}}\frac{d}{dw}(1+e^{-y_i<w,x_i>}) \\ &= \frac{1}{m}\sum_i^m \frac{1}{1+e^{-y_i<w,x_i>}}\frac{d}{dw}e^{-y_i<w,x_i>} \\ &= \frac{1}{m}\sum_i^m \frac{1}{1+e^{-y_i<w,x_i>}} * -e^{-y_i<w,x_i>}\frac{d}{dw}(-y_i<w,x_i>) \\ &= \frac{1}{m}\sum_i^m \frac{-e^{y_i<w,x_i>}}{1+e^{-y_i<w,x_i>}} * (-y_i x_i) \end{aligned}$$

Question 5 Here we are going to minimize $L_{l2}(hw,s)$

$$
\begin{aligned}
minL_{l2}(hw,s) &= min\sum_{i}^{m}(h_w(x_i)-y_i)^2+\lambda||w||_2^2 \\
&= min\sum_{i}^{m}[(y_i-x_iw)^T(y_i-x_iw)+\lambda||w||_2^2] \\
&= min\sum_{i}^{m}(y_i^T-w^Tx_i^T)(y_i-x_iw)+\lambda w^Tw \\
&= min\sum_{i}^{m}y_i^Ty_i-y_i^Tx_iw-y_iw^Tx_i^T+x_i^Tx_iw^Tw+\lambda w^Tw \\
\frac{dL_{l2}(hw,s)}{dw} &= \sum_{i}^{m}y^Tx_i-y_ix_i^T+x_i^Tx_iw+\lambda w^T \\
set\frac{dL_{l2}(hw,s)}{dw} &= 0 \\
\sum_{i}^{m}w^T(\lambda+x_iT) &= \sum_{i}^{m}2y_i^Tx_i \\
w &= \sum_{i}^{m}(\lambda+x_i^Tx_i)x_i^Ty \\
that \quad is \quad w &= (A+\lambda I)^{-1}b
\end{aligned}
$$