CSE343/ ECE363/ ECE563: Machine Learning W2021
Assignment-1
Linear / Logistic Regression and Naive Bayes

Navneet Agarwal
2018348

April 2, 2021

Theory Questions

4)

The linear model is given by the equation :

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + +\beta_k x_{ki} + \epsilon_i \tag{1}$$

The same equation in matrix form is given as :

$$Y = X\beta + \epsilon \tag{2}$$

So, this equation can be re-written as :

$$\epsilon = Y - X\beta \tag{3}$$

Let the sum of squared error loss be represented by the symbol $L(\beta)$ then :

$$L(\beta) = \sum_{i=1}^{n} (\epsilon_i)^2$$

Since we know that :

$$\sum_{i=1}^{n} (\epsilon_i)^2 = \epsilon^T \epsilon$$

Using the above result and equation (3) we get :

$$\begin{aligned} L(\beta) &= [Y - X\beta]^T [Y - X\beta] \\ &= [Y^T - (X\beta)^T][Y - X\beta] \\ &= [Y^T - \beta^T X^T][Y - X\beta] \end{aligned}$$

$$L(\beta) = Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta \tag{4}$$

Now β has a dimension of $kX1$, X has a dimension of nXk and Y has a dimension of $nX1$
Hence, the dimension of $\beta^T X^T Y$ and $Y^T X\beta$ is $1X1$.
We also know that :

$$(\beta^T X^T Y)^T = Y^T X\beta$$

By using the information above and using equation 4 we have the sum of squared error in matrix representation as:

$$L(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta$$

In order to derive the least squares solution, we have to differentiate the matrix w.r.t to β:

$$\frac{dL(\beta)}{d\beta} = \frac{d(Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta)}{d\beta}$$

$$\frac{d}{d\beta}L(\beta) = \frac{d}{d\beta}(Y^T Y) - 2\frac{d}{d\beta}(\beta^T X^T Y) + \frac{d}{d\beta}(\beta^T X^T X\beta)$$

$$\frac{d}{d\beta}L(\beta) = 0 - 2X^T Y + 2(X^T X\beta)$$

$$\frac{d}{d\beta}L(\beta) = -2X^T Y + 2X^T X\beta$$

We need to put this derivative equal to zero for minimization:

$$\frac{d}{d\beta}L(\beta) = 0$$

$$-2X^T Y + 2X^T X\beta = 0$$

$$(X^T X\beta) = X^T Y$$

$$(X^T X\beta) = X^T Y$$

$$\beta = (X^T X)^{-1} X^T Y$$

The conditions under which closed form solution will exist are :

1. Inverse of $X^T X$ should exist.
2. The rank of matrix X should be k. (k is the number of features)
3. Number of data samples(n) should be greater than or equal to Number of features(k).

5)

We get the following equation to predict the salary :

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

$$y = \beta_0 + \beta_1 GPA + \beta_2 IQ + \beta_3 Gender + \beta_4 (GPA)(IQ) + \beta_5 (GPA)(Gender)$$

$$y = 50 + 20GPA + 0.07IQ + 35Gender + 0.01(GPA)(IQ) - 10(GPA)(Gender)$$

Since for Gender, 1 is for female and 0 is for male,

$$y = \begin{cases} 50 + 20GPA + 0.07IQ + 35(0) + 0.01(GPA)(IQ) - 10(GPA)(0), & \text{if } Gender = 0; Male \\ 50 + 20GPA + 0.07IQ + 35(1) + 0.01(GPA)(IQ) - 10(GPA)(1), & \text{if } Gender = 1; Female \end{cases}$$

$$y = \begin{cases} 50 + 20GPA + 0.07IQ + 0.01(GPA)(IQ); & \text{if } Gender = 0; Male \\ 85 + 10GPA + 0.07IQ + 0.01(GPA)(IQ); & \text{if } Gender = 1; Female \end{cases}$$

1.

We have to come with a condition where females earn more than male or vice versa, keeping the IQ and GPA as same.
Female - Male Salary = $y_{\text{female}} - y_{\text{male}}$
Female - Male Salary = $[85+10GPA+0.07IQ+0.01(GPA)(IQ)]-[50+20GPA+0.07IQ+0.01(GPA)(IQ)]$
Female - Male Salary = $35 - 10GPA$

Now, Female - Male Salary< 0 when

$$35 - 10GPA < 0$$

$$35 < 10GPA$$

$$3.5 < GPA$$

$$GPA > 3.5$$

If we assume that the maximum GPA is 4 or 5, which mostly is the case in US colleges . We can say that Male salary is greater than females when GPA is very high , (greater than 3.5) but both the male and female candidate have the same GPA and IQ.
Thus, **(C) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough** is the correct answer.

2.

$$y_{female} = 85 + 10GPA + 0.07IQ + 0.01(GPA)(IQ)$$

$$y_{female} = 85 + 10 * 3.5 + 0.07IQ + 0.01(3.5)(115)$$

$$y_{female} = 85 + 10 * 3.5 + 0.07IQ + 0.01(3.5)(115)$$

$$y_{female} = 132.075$$

Thus, the predicted salary is $\$132,075$

3.

False
The above statement is not true because we cannot judge the a feature based on the coefficients. Lets say for example that GPA/IQ interaction term (product of GPA/IQ) has a very high magnitude which when gets multiplied with a small coeffient gives a reasonable change in the value of predcited salary. It may be possible that the small coefficient is there to normalize the effect of high magnitude of this interaction term. In order to see whether this interaction term has a small effect, we will have to test a hypothesis by putting $\beta_4 = 0$. Only after testing this hypothesis and comparing the performance ,we can say whether the interaction term has small effect on the predicted salaty or not.

6)

Some definitions from the A Unified Bias-Variance Decomposition for Zero-One and Squared Loss

Training set$\{(x_1, t_1), ..., (x_n, t_n)\}$ where x_i are the data samples and t_i are the true labels.

y is the predicted value here.

Loss function L(t, y) = $(t - y)^2$ is the squared loss.

Loss function L(t, y) = $|t - y|$ is the absolute loss.

Loss function L(t, y) = 0 if y = t and $L(t, y) = 1$ otherwise, this is the zero-one loss.

The main predictions for loss function L and set of training set D in S is $y_m^{L,S} = argmin_{y'} E_D[L(y, y')]$. $y_m = y_m^{L,D}$ for our discussion

Bias $B(x) = L(y_*, y_m)$ ... (A)
It is used to measure the loss incurred by the main prediction with respect to the optimal prediction.

Variance $V(x) = E_D[L(y_m, y)]$ ... (B)
It is used to measure the average loss incurred by the realtive prediction with respect to the main prediction.

Noise $N(x) = E_t[L(t, y_*)]$ ... (C)
This is the unavoidable component of the loss.

1.

Given a dataset D in S and squared loss as $L(t, y) = (t - y)^2$ the expected loss can be written as:

$$E_{D,t}[L(t, y)] = E_{D,t}[(t - y)^2]$$

$$E_{D,t}[L(t, y)] = E_{D,t}[(t - y)^2]$$

$$E_{D,t}[L(t, y)] = E_{D,t}[t^2 + y^2 - 2ty]$$

$$E_{D,t}[L(t, y)] = E_{D,t}[t^2] + E_{D,t}[y^2] - 2E_{D,t}[ty]$$

$$E_{D,t}[L(t, y)] = E_{D,t}[t^2] + E_{D,t}[y^2] - 2E_{D,t}[t]E_{D,t}[y]$$

Since, t does not depend on the dataset D and y does not depend on variable t.

$$E_{D,t}[L(t, y)] = E_t[t^2] + E_D[y^2] - 2E_t[t]E_D[y] \quad (1)$$

Now, for a random variable x, we know that

$$E[x^2] = \overline{x}^2 + E[(x - \overline{x})^2]; E[x] = \overline{x}$$

Using the above result in equation 1 we have,

$$E_{D,t}[L(t,y)] = \overline{t}^2 + E_t[(t - \overline{t})^2] + \overline{y}^2 + E_D[(y - \overline{y})^2] - 2E_t[t]E_D[y]$$

$$E_{D,t}[L(t,y)] = \overline{t}^2 + E_t[(t - \overline{t})^2] + \overline{y}^2 + E_D[(y - \overline{y})^2] - 2E_t[t]\overline{y}$$
$$E_{D,t}[L(t,y)] = E[t]^2 + E_t[(t - E(t))^2] + \overline{y}^2 + E_D[(y - \overline{y})^2] - 2E_t[t]\overline{y}$$
$$E_{D,t}[L(t,y)] = E_t[(t - E(t))^2] + E_D[(y - \overline{y})^2] + E[t]^2 - 2E_t[t]\overline{y} + \overline{y}^2$$
$$E_{D,t}[L(t,y)] = E_t[(t - E(t))^2] + E_D[(y - \overline{y})^2] + (E[t] - y)^2 \quad (2)$$

On comparing 2 with equations A,B and C, we get:
Bias : $(E[t] - y)^2$
Variance : $E_D[(y - \overline{y})^2]$
Noise: $E_t[(t - E(t))^2]$ where $L(t, y*) = L(t, E[t]) = (t - E[t])^2$ Hence,

$$E_{D,t}[L(t,y)] = V(x) + B(x) + N(x)$$

On comparing this with the equation given in question, we can say that c1 = c2 =1.

2.

For zero-loss function, Loss function L(t, y) = 0 if y = t and $L(t, y) = 1$ or we can say that :

$$\text{Loss function L(t, y)} = \begin{cases} 0, & \text{if } y = t \\ 1, & \text{otherwise} \end{cases}$$

First we need to show that :

$$E_t[L(t,y)] = L(y*, y) + c_0 E_t[L(t, y*)] \quad (D)$$

We have the following cases for this equation:
Case 1: $y = y*$
If $c_0 = 1$ the equation will be trivially true with $y = y*$
Case 2: $y \neq y*$
There are only two classes, so if $y \neq y*$ and $t \neq y*$, this implies that $t = y$ and vice versa.
Hence $P_t(t = y) = P_t(t \neq y*)$

$$E_t[L(t,y)] = P_t(t \neq y)$$
$$E_t[L(t,y)] = 1 - P_t(t = y)$$
$$E_t[L(t,y)] = 1 - P_t(t \neq y*)$$
$$E_t[L(t,y)] = 1 - E_t[L(t, y*)]$$
$$E_t[L(t,y)] = 1 - E_t[L(t, y*)]$$
$$E_t[L(t,y)] = L(y, y*) - E_t[L(t, y*)]$$
$$E_t[L(t,y)] = L(y, y*) + c_0 E_t[L(t, y*)]$$

with c0 = -1 which is essentially equation D

Now we need to prove that:

$$E_D[L(y*, y)] = L(y*, y) + c_1 E_D[L(y_m, y)] \quad (E)$$

We have the following cases for this equation:
Case 1: $y_m = y*$

If $c_1 = 1$ the equation will be trivially true with $y_m = y*$
Case 2: $y_m \neq y*$
There are only two classes, so if $y_m \neq y*$ and $y_m \neq y*$, this implies that $y* = y$ and vice versa.
Hence $P_D(y* = y) = P_D(y_m \neq y)$

$$E_t[L(t,y)] = P_D(y* \neq y)$$
$$E_t[L(t,y)] = 1 - P_D(y* = y)$$
$$E_t[L(t,y)] = 1 - P_D(y_m \neq y)$$
$$E_t[L(t,y)] = 1 - E_D[L(y_m, y)]$$
$$E_t[L(t,y)] = L(y*, y_m) - E_D[L(y_m, y)]$$
$$E_t[L(t,y)] = L(y*, y_m) + c_1 E_D[L(y_m, y)]$$

with $c_1 = -1$ which is essentially equation E

$$E_{D,t}[L(t,y)] = E_D[E_t[L(t,y)]]$$

Using equation D,

$$E_{D,t}[L(t,y)] = E_D[L(y*, y) + c_0 E_t[L(t, y*)]]$$

Now, $L(t, y*)$ is independent of D.

$$E_{D,t}[L(t,y)] = E_D[L(y*, y)] + E_D[c_0] E_t[L(t, y*)]]$$

and

$$E_D[c_0] = P_D(y = y*) - P_D(y \neq y*)$$
$$E_D[c_0] = 2P_D(y = y*) - 1$$
$$E_D[c_0] = c_2$$

Hence, we finally obtain the equation given in question using equation E.
Note: The order of c_1 and c_2 was interchanged in our question as compared to the research paper referred below.

Reference : A Unified Bias-Variance Decomposition for Zero-One and Squared Loss