CSCI567 2013 Homework Assignment 2

Student Name	Arpit Bansal
Last 4 digits of USC ID	0979
I have collaborated with	Jia Li Udit Agrawal Yin Ray Rick Huang

1 Linear Regression

1.1 Question 1

We are given $w^{LMS} = (X^T X)^{-1} X^T y$ which is only possible when $X^T X$ is invertible. Here X^T i.e. $[x_1, x_2, \ldots, x_n]$ is a matrix of $N \times D$. So $X^T X$ will be:

$$\begin{pmatrix} \sum_n x_{n1}^2 & \sum_n x_{n1}x_{n2} & \ldots & \sum_n x_{n1}x_{nD} \\ \sum_n x_{n1}x_{n2} & \sum_n x_{n2}^2 & \ldots & \sum_n x_{n2}x_{nD} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_n x_{n1}x_{nD} & \ldots & \ldots & \sum_n x_{nD}^2 \end{pmatrix}$$

Now, if columns of X are orthogonal, then above matrix transforms into:

$$\begin{pmatrix} \sum_n x_{n1}^2 & 0 & \ldots & 0 \\ 0 & \sum_n x_{n2}^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & \ldots & \sum_n x_{nD}^2 \end{pmatrix}$$

because $\sum_n x_{nd}x_{nd'} = 0, \forall d \neq d'$. Here n iterates from 1 to N. Now using the inverse property of a diagonal matrix, we can write $(X^T X)^{-1}$ as below:

$$\begin{pmatrix} \frac{1}{\sum_n x_{n1}^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sum_n x_{n2}^2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & \ldots & \frac{1}{\sum_n x_{nD}^2} \end{pmatrix} \tag{1}$$

Now, $X^T y$, a matrix of R^D, will be:

$$\begin{pmatrix} \sum_n x_{n1}y_n \\ \sum_n x_{n2}y_n \\ \vdots \\ \sum_n x_{nD}y_n \end{pmatrix} \tag{2}$$

Now we are given $v_d = \frac{\sum_n x_{nd}y_n}{\sum_n x_{nd}^2}$ where d is R^D. So lets look at different values of v:

$$v_1 = \frac{\sum_n x_{n1}y_n}{\sum_n x_{n1}^2}, v_2 = \frac{\sum_n x_{n2}y_n}{\sum_n x_{n2}^2} \ldots v_D = \frac{\sum_n x_{nD}y_n}{\sum_n x_{nD}^2}$$

We can combine these v_d values and can represent them in matrix multiplication of a diagonal and a column matrix as below:

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_D \end{pmatrix} = \begin{pmatrix} \frac{1}{\sum_n x_{n1}^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sum_n x_{n2}^2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & \ldots & \frac{1}{\sum_n x_{nD}^2} \end{pmatrix} \begin{pmatrix} \sum_n x_{n1}y_n \\ \sum_n x_{n2}y_n \\ \vdots \\ \sum_n x_{nD}y_n \end{pmatrix} \tag{3}$$

Here also n iterate over 1 to N. Let us denote v matrix as V, diagonal matrix as P and column matrix as Q i.e. $V = PQ$. Now by comparing (1), (2) and (3), we can easily state that

$$P = (X^T X)^{-1}, Q = y \Rightarrow V = w_{LMS}$$

only if the given condition of orthogonality is satisfied.

1.2 Question 2

1.2.1

We know that for a univariate linear regression of $y = wx$, we get the w^{LMS} as $\frac{\sum_n x_n y_n}{\sum_n x_n^2}$. Now residual error e_n is defined as $y_n - w^{LMS} x_n$. So to prove that e_n and x_n are uncorrelated, we have to prove $\sum_n e_n x_n = 0$. So,

$$\sum_n e_n x_n = \sum_n (y_n - w^{LMS} x_n) x_n$$

$$\Rightarrow \sum_n e_n x_n = \sum_n y_n x_n - w^{LMS} \sum_n x_n^2$$

Now, putting the value of w^{LMS}:

$$\sum_n e_n x_n = \sum_n y_n x_n - \frac{\sum_n x_n y_n}{\sum_n x_n^2} \sum_n x_n^2 \Rightarrow 0$$

1.2.2

(a) We are given $z_3 = u_3 - \gamma_{31} z_1 - \gamma_{32} z_2$. To prove z_3 is uncorrelated with z_1,

$$\sum_n z_{3n} z_{1n} \Rightarrow \sum_n ((u_{3n} - \gamma_{31} z_{1n}) z_{1n} - \gamma_{32} z_{2n} z_{1n})$$

$$\Rightarrow \sum_n (u_{3n} - \gamma_{31} z_{1n}) z_{1n} - \gamma_{32} \sum_n (z_{2n} z_{1n})$$

Now, as we know that z_1 and z_2 are uncorrelated. So, $\sum_n z_{2n} z_{1n} = 0$. Putting this in above equation,

$$\sum_n z_{3n} z_{1n} = \sum_n (u_{3n} - \gamma_{31} z_{1n}) z_{1n}$$

Now we know from 1.2.1, $\sum_n e_n x_n = 0$. So using this and looking at the above equation, we can say that z_3 and z_1 are uncorrelated i.e. $\sum_n z_{3n} z_{1n} = 0$. This gives rise to γ_{31} which will be equal to $\frac{\sum_n u_{3n} z_{1n}}{\sum_n z_{1n}^2}$ using the concept of w^{LMS} for univariate regression. Similarily, we can prove for uncorrelation between z_3 and z_2 and get the value of γ_{32} as $\frac{\sum_n u_{3n} z_{2n}}{\sum_n z_{2n}^2}$.

(b) We can use the induction here to prove it. Suppose d=k then z_k will be,

$$z_k = u_k - \gamma_{k1}z_1 - \gamma_{k2}z_2 - \cdots - \gamma_{k(k-1)}z_{k-1}$$

We assume that z_k is uncorrelated with every $z_{k'}$ where $k^{'}\epsilon[1, k-1]$. Now we have to prove for z_{k+1} in order to prove that every z is uncorrelated with another one. Here, z_{k+1} will be,

$$z_{k+1} = u_{k+1} - \gamma_{(k+1)1}z_1 - \gamma_{(k+1)2}z_2 - \cdots - \gamma_{(k+1)k}z_k$$

Using the approach in **(a)**, we can easily show that z_{k+1} is also uncorrelated to all $z_{k'}$ where $k^{'}\epsilon[1, k]$. So if we multiply z_{k+1} with any vector z_i which is uncorrelated with any other vector, we will get,

$$\sum_n z_{(k+1)n}z_{in} = \sum_n (u_{(k+1)n} - \gamma_{(k+1)i}z_{in})z_{in}$$

Using the derivation in previous problem, we can say that z_{k+1} is also uncorrelated to z_i and we can also say that $\gamma_{(k+1)i} = \frac{\sum_n u_{(k+1)n}z_{in}}{\sum_n z_{in}^2}$.

Here our k can vary upto D. So we can say using the above induction that all pairs of z_k and $z_{k'}$ are uncorrelated as long as $1 \leq k \neq k^{'} \leq D$. Note here that, when $k = k^{'}$, we are trying to prove that a column is uncorrelated to itselfs which is not true.

1.2.3

We can now denote all these transformed column vectors by one matrix $Z = [z_1 z_2 \dots z_D]$. We can actually represent each tranformation in matrix decomosition i.e.

$$[u_1 u_2 u_3 \dots u_D]_{N\times D} = [z_1 z_2 z_3 \dots z_D]_{N\times D} \begin{pmatrix} 1 & \gamma_{21} & \gamma_{32} & \cdots & \gamma_{D(D-1)} \\ 0 & 1 & \gamma_{31} & \cdots & \gamma_{D(D-2)} \\ 0 & 0 & 1 & \cdots & \gamma_{D(D-3)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & 1 \end{pmatrix}_{D\times D}$$

Now we can see that the right hand matrix of the decomposition is upper triangular matrix and we know that upper triangular matrices are invertible. Now lets denote it as E and z matrix as Z. So now we have $X = ZE$. We know that preditcion with training set as X is $X^T X w^{LMS} = X^T y$. So now putting $X = ZE$ in this equation,

$$(ZE)^T ZE w^{LMS} = (ZE)^T y \Rightarrow E^T Z^T ZE w^{LMS} = E^T Z^T y$$

We know that E is invertible, so taking $(E^T)^-1$ on both sides,

$$Z^T ZE w^{LMS} = Z^T y$$

We also know that Z is an orthogonal matrix as we have proved in previous problems. So it is necessarily invertible. So we can write above equation as:

$$Ew^{LMS} = (Z^T Z)^{-1} Z^T y \tag{4}$$

We are using Z to predict y also. So optimal parameter vector estimation for Z will be as follow:

$$Z^T Z\beta = Z^T y \Rightarrow \beta = (Z^T Z)^{-1} Z^T y$$

Using above result in (4), we get,

$$Ew^{LMS} = \beta \Rightarrow \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_D \end{pmatrix} = \begin{pmatrix} 1 & \gamma_{21} & \gamma_{32} & \dots & \gamma_{D(D-1)} \\ 0 & 1 & \gamma_{31} & \dots & \gamma_{D(D-2)} \\ 0 & 0 & 1 & \dots & \gamma_{D(D-3)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix}$$

Now as we are interested in β_D only, we can see by multiplying above matrices that,

$$\beta_D = w_D$$

1.2.4

The time complexity for linear regression method is $O(ND^3)$. Shuffling will be done D times and transformation for whole matrix X will take $O(ND^2)$. The original closed form solution has the complexity of $O(D^3) + O(ND^2)$ where $O(D^3)$ comes from inverting the $X^T X$ and $O(ND^2)$ is the complexity of matrices multiplications. So linear regression is actually worse. However, the main advantage of this approach is scalability. We can start D tasks parallely and can compute each element of w^{LMS} independently.

1.3 Question 3

What we can do here is to use brute force for selecting the more beneficial features. One approach is first iterate over all the features and for each individual feature get the accuracy of model. Then out of these features pick the best one. Put this feature in our set of best features and remove this feature from the original space. Now, run the above process again and pick the second best feature again. This process is repeated until there is no improvement in the accuracies i.e. we got the maximal accuracy using only the features in our subset. Runtime complexity for this process is $O(n^2)$ where n is the number of total features.

2 Logistic Regression

2.1 Question 4

For generative approach, lets take the joint probability model:

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

As it is for binary classification, We can write it as below,

$$p(x, y) = p(y = 1)p(x|y = 1) + p(y = 2)p(x|y = 2)$$

$$\Rightarrow P(x, y) = \prod_{y=C_1} p_1 p(x_n|C1) + \prod_{y=C_2} p_2 p(x_n|C2)$$

So lets take log-likelihood of above distribution by putting the gaussian distribution for $p(x|y)$ as given,

$$\log p(D) = \sum_{y=C_1} \frac{-(x_n-\mu_1)^2}{2\sigma^2} + \sum_{y=C_2} \frac{-(x_n-\mu_2)^2}{2\sigma^2} + \sum_{y=C_1} \log p_1 + \sum_{y=C_2} \log p_2 - \frac{1}{2}\sum_n \log \sigma^2 - \frac{1}{2}\sum_n \log 2\pi$$

$$\Rightarrow L(D) = \frac{-1}{2\sigma^2}[N_1\mu_1^2 - 2\mu_1 \sum_n x_n + N_2\mu_2^2 - 2\mu_2 \sum_n x_n + 2\sum_n x_n^2] + \sum_{y=C_1} \log p_1 + \sum_{y=C_2} \log p_2 - \frac{N}{2}\log \sigma^2 - \frac{N}{2}\log 2\pi$$

Now differentiating over μ_1, μ_2 and σ, we get

$$\frac{\delta L(D)}{\delta \mu_1} = N_1\mu_1 - \sum_n x_n = 0 \Rightarrow \mu_1 = \frac{\sum_{n:C1} x_n}{N_1}$$

$$\frac{\delta L(D)}{\delta \mu_2} = N_2\mu_2 - \sum_n x_n = 0 \Rightarrow \mu_2 = \frac{\sum_{n:C2} x_n}{N_2}$$

$$\frac{\delta L(D)}{\delta \sigma} = \frac{\sum_{n:C_1}(x-\mu_1)^2 + \sum_{n:C_1}(x-\mu_2)^2}{\sigma^2} - N = 0 \Rightarrow \sigma^2 = \frac{\sum_{n:C_1}(x-\mu_1)^2 + \sum_{n:C_1}(x-\mu_2)^2}{N}$$

where $N = N_1 + N_2$.

2.2 Question 5

$$p(C_k|x; w_k) = \frac{e^{w_k^T x}}{\sum_{k'} e^{w_k^T x}}$$

Now lets shift w_k to $w_k + c$,

$$p(C_k|x; w_k + c) = \frac{e^{(w_k+c)^T x}}{\sum_{k'} e^{(w_k+c)^T x}} = \frac{e^{w_k^T x + cx}}{\sum_{k'} e^{w_k^T x + cx}}$$

$$\Rightarrow \frac{e^{w_k^T x} e^{cx}}{\sum_{k'} e^{w_k^T x} e^{cx}} = \frac{e^{w_k^T x} e^{cx}}{e^{cx}\sum_{k'} e^{w_k^T x}} = \frac{e^{w_k^T x}}{\sum_{k'} e^{w_k^T x}} = p(C_k|x; w_k)$$

So we have proved that shifting parameters by a constant will not change the likelihood value. So our solution is not unique here. One possible strategy is to set $w_1 = 0$ or identically say, c = -w_1. So each time we shift our solution, we are trying to do $w_k - w_1$ which will always give us value w_k. So in this case, our solution is unique. Another possible solution is $\sum_k w_k = 0$ which will also enforce the same thing as above.

2.3 Question 6

Cross entropy fucntion for multinomial logistic regression is defined as below:

$$\epsilon(w_1, w_2, \ldots, w_k) = -\sum_{k=1}^{K} \sum_{y=k} P(C_k|x_n)$$

$$\Rightarrow \epsilon(w_1, w_2, \ldots, w_k) = \sum_{k=1}^{K} \sum_{y=k} (w_k^T x) + \sum_n log(\sum_{k'=1}^{K} e^{-w_k'^T x}))$$

Now, to get Hessian matrix we have to double differentiate the above cross entropy error function with respect to parameter. Hessian matrix H is define as below:

$$H = \frac{\delta^2 \epsilon}{\delta w w^T} = \begin{pmatrix} \frac{\delta^2\epsilon}{\delta w_1 w_1^T} & \frac{\delta^2\epsilon}{\delta w_1 w_2^T} & \cdots & \frac{\delta^2\epsilon}{\delta w_1 w_k^T} \\ \frac{\delta^2\epsilon}{\delta w_2 w_1^T} & \frac{\delta^2\epsilon}{\delta w_2 w_2^T} & \cdots & \frac{\delta^2\epsilon}{\delta w_2 w_k^T} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\delta^2\epsilon}{\delta w_k w_1^T} & \cdots & \cdots & \frac{\delta^2\epsilon}{\delta w_k w_k^T} \end{pmatrix}_{KD \times KD}$$

Here K is the number of classes. We can see that each element of this Hessian matrix H is of $D \times D$ dimension. We will now consider the elements of our Hessian matrix, Diagonal elements and Non Diagonal elements, i.e. $\frac{\delta^2\epsilon}{\delta w_i w_i^T}$ and $\frac{\delta^2\epsilon}{\delta w_i w_j^T}$

Let us now consider only one parameter w_1. So by differentiating cross entropy function ϵ w.r.t. to w_1, we get

$$\frac{\delta\epsilon}{\delta w_1} = \sum_{y=1} x_n - \sum_n \frac{e^{w_1^T x}}{\sum_{k'=1}^{K} e^{-w_k'^T x}}$$

Now, we will differentiate the above equation with respect to w_1 in order to consider diagonal elements and w_k for considering non-diagonal elements. So for diagonal element,

$$\frac{\delta^2\epsilon}{\delta w_1 w_1^T} = \sum_n \frac{e^{-w_1^T x_n} \sum_{k \neq 1} e^{-w_k^T x_n}}{(\sum_{k'=1}^{K} e^{-w_k'^T x})^2} x_n x_n^T$$

Lets take $\frac{e^{-w_1^T x_n} \sum_{k \neq 1} e^{-w_k^T x_n}}{(\sum_{k'=1}^{K} e^{-w_k'^T x})^2} = c_{11}$. We can easily see here that c_{11} is positive. So by doing above computation, we can easily prove that $\forall i \epsilon (1, n)$,all the diagonal elements c_{ii} of H are positive. If we take a vector v of R^D then we can write $v^T H_{11} v$ as $\sum_n c_{11} v^T x_n x_n^T v = \sum_n c_{11} (v^T x_n)(v^T x_n)^T \geq 0$. We can do this for all the diagonal elements of H similarily. This means that all the diagonal elements of our Big Hessian Matrix are semi-positive definite.

Now lets differentiate w.r.t. w_k to consider a non-diagonal element. We get

$$\frac{\delta^2\epsilon}{\delta w_1 w_k^T} = \sum_n \frac{e^{-w_1^T x_n} e^{-w_k^T x_n}}{(\sum_{k'=1}^{K} e^{-w_k'^T x})^2} x_n x_n^T$$

Lets take $\frac{e^{-w_1^T x_n} e^{-w_k^T x_n}}{(\sum_{k'=1}^{K} e^{-w_k'^T x})^2} = c_{1k}$. We can easily see here that c_{1k} is positive. So by doing above computation, we can easily prove that $\forall i, j \epsilon (1, n)$,all the non-diagonal elements c_{ij} of H are positive. If we take a vector v of R^D then we can write $v^T H_{1k} v$ as $\sum_n c_{1k} v^T x_n x_n^T v = \sum_n c_{1k} (v^T x_n)(v^T x_n)^T \geq 0$. We can do this for all the non-diagonal elements of H similarily. This means that all the non-diagonal elements of our Big Hessian Matrix are also positive semidefinite.

Using the above observations, we can also tell that $H_{ij} = H_{ji}$ i.e., H is also *symmetric*. One another important feature of diagonal elements of our Hessian matrix is $H_{ii} = \sum_{i \neq j} H_{ij}$. This property will play an important role in proving positive semidefiniteness of H.

Now, we will approach towards our goal of proving the positive semidefiniteness of H. We know that all the elements of H are positive semidefinite.

Now lets us consider a Vector V of R^{KD}. We can break down this vector into K parts each having D dimensions.

$$V = \begin{pmatrix} V_1^T \\ V_2^T \\ \vdots \\ V_K^T \end{pmatrix}$$

By multiplying this V with our Hessian matrix H we will get,

$$V^T H V = \sum_{i=1 \cdots k; j=1 \cdots k} V_i^T H_{ij} V_j$$

So the right hand side of above equation will give us four combination of summation i.e. $V_i^T H_{ij} V_j, V_i^T H_{ij} V_i, V_j^T H_{ji} V_j, V_j^T H_{ji} V_i$. Using the property of our diagonal elements in Hessian matrix we got above, we can also state that $V_i^T H_{ii} V_i = \sum_{j \neq i} V_i H_{ij} V_i$. We also know that Hessian matrix is symmetric. We can also write all the combinations of summations as follow $V_i^T H_{ij} V_j + V_i^T H_{ij} V_i + V_j^T H_{ji} V_j + V_j^T H_{ji} V_i = (V_i + V_j)^T H_{ij} (V_i + V_j)$

So we can write the whole multiplication in following form:

$$V^T H V = \sum_{i=1 \cdots k; j=1 \cdots k} (V_i + V_j)^T H_{ij} (V_i + V_j)$$

We have already proved that all H_{ij} are positive semidefinite. So we can say that $V^T H V \geq 0$. This statement proves that our Hessian matrix H is positive semidefinite which implies that our cross entropy function is convex.