CSCI 567: Homework 1

Deepika Anand

October 3, 2016

Problem 1. (a)

Assume, $p = e^{b+(w^T)*x}$
$L = \frac{p}{1+p}^y * \frac{1}{1+p}^{1-y}$
Taking log on both sides,
$logL = \sum_{i=1}^{n}((y * log\frac{p}{1+p}) - (1-y) * log(1+p))$
$= \sum_{i=1}^{n}(y * logp - y * log(1+p) - (1-y) * log(1+p))$
$= \sum_{i=1}^{n}(y * logp - y * log(1+p) - log(1+p) + y * log(1+p))$
Replacing p with the actual value
$= \sum_{i=1}^{n} y * (b + w^T * x_i) - log(1 + e^{b+w^T*x_i})$
Give in question
$-\prod_{i=1}^{n} logP(Y = y_i|X = x_i)$
Plugging the value from above we get $-\sum_{i=1}^{n}(y_i * (b + w^T * x_i)) - log(1 + e^{b+w^T*x_i})$

Problem 1. (b)

By taking the double differentiation of the above statement we get
$x * p * x^T * (1-p) * x >= 0$ which is a convex function and hence it will converge eventually where $p = e^{b+(w^T)*x}$

Problem 1. (c)

Using softmax function

$$p(Y = c_k|X = x) = \frac{e^{w_k^T x}}{\sum_r e^{w_r^T x}} \tag{1}$$

Now since there k class and and assume the date is D dimensional

$$\prod_{i=1}^{k}\prod_{l=1}^{D}\frac{e^{w_k^T x^l}}{\sum_r e^{w_r^T x^l}} \tag{2}$$

Taking log of the equation to get the maximum likelihood

$$\sum_{i=1}^{k}\sum_{l=1}^{D} w_k^T x^l - ln\sum_r e^{w_r^T x^l} \tag{3}$$

Problem 1. (d)

Taking derivative of the equation derived above with respect to w_i

$$\sum_{l=1}^{D} I(y == k_i)x^l - \frac{x^l e^{w_i^T x^l}}{\sum_r e^{w_i^T x^l}} \tag{4}$$

I is the identity function since after differentiation presence or absence of x^l will depend whether y belongs to kth class or not

Problem 2. (a)

Since the random variable y follows gausian distribution. Therefore,

$$L = (\frac{e^{-\frac{(x_i-\mu_1)^2}{2*\sigma_1}}}{\sigma\sqrt{2*\pi}})_1^p * (\frac{e^{-\frac{(x_i-\mu_2)^2}{2*\sigma_2}}}{\sigma\sqrt{2*\pi}})_2^p \tag{5}$$

Take log on both sides to get

$$logL = p_1 * (-\sum_{i=1}^{n}(\frac{(x_i-\mu_1)^2}{2*\sigma_1}) - ln\sqrt{2*\pi} - ln\sigma_1) + p_2 * (-\sum_{i=1}^{n}(\frac{(x_i-\mu_2)^2}{2*\sigma_2}) - ln\sqrt{2*\pi} - ln\sigma_2) \tag{6}$$

To find μ_1^* which maximizes equation 6, differentiate wrt μ_1 Since all terms on right side of + is equation 6 will be constant so they will come out to zero. Left hand side terms of + will give

$$\frac{p_1}{2*\sigma_1^2} * (\sum_{i=1}^{n}(x_i - \mu_1)) = 0 \tag{7}$$

$$\mu_1^* = \frac{\sum_{i=1}^{n} I(y == 1)x_i}{\sum_{i=1}^{n} I(y == 1)} \tag{8}$$

where I that is identity function controls if outcome of random variable y belongs to class 1 or 2. If it belongs to class 2 then x is not considered hence product is zero otherwise x value is taken as it is.
Similarily for

$$\mu_2^* = \frac{\sum_{i=1}^{n} I(y == 2)x_i}{\sum_{i=1}^{n} I(y == 2)} \tag{9}$$

To find maximum value of σ_1^*, differentiate equation 6 wrt σ_1

$$\frac{dL}{d\sigma_1} = \sum_{i=1}^{n} \frac{(x_i - \mu_1)}{\sigma_1^3} - \frac{1}{\sigma_1} = 0 \tag{10}$$

$$\sigma_1^* = \frac{\sum_{i=1}^{n} I(y == 1)x_i - \mu_1}{\sum_{i=1}^{n} I(y == 1)} \tag{11}$$

$$\sigma_2^* = \frac{\sum_{i=1}^{n} I(y == 2)x_i - \mu_2}{\sum_{i=1}^{n} I(y == 2)} \tag{12}$$

To maximize p_1 that is probability of events for occurrence of $P(Y == 1)$ will be done when such that

$$p_1^* = \frac{\sum_{i=1}^{n} I(y == 1)x_i}{\sum_{i=1}^{n} I(y == 1)x_i + \sum_{i=1}^{n} I(y == 2)x_i} \tag{13}$$

$$p_2^* = \frac{\sum_{i=1}^{n} I(y == 2)x_i}{\sum_{i=1}^{n} I(y == 1)x_i + \sum_{i=1}^{n} I(y == 2)x_i} \tag{14}$$

Problem 2. (b)

$$P(Y = 1|x) = \frac{P(X|Y = 1)P(Y = 1)}{P(X|Y = 1)P(Y = 1) + P(X|Y = 2)P(Y = 2)} \tag{15}$$

This can be also written as,

$$P(Y = 1|x) = \frac{1}{1 + \frac{P(Y=2)P(X|Y=2)}{P(Y=1)P(X|Y=1)}} \tag{16}$$

Since given that multivariate distribution is followed. And adding $e^{log_e} = 1$. We can get the following

$$log(P(X|Y = 1)) = -1/2ln \sum -D/2ln\pi - 1/2(x - \mu_1)^T \sum^{-1}(x - \mu_1) \tag{17}$$

$$log(P(X|Y = 2)) = -1/2ln \sum -D/2ln\pi - 1/2(x - \mu_2)^T \sum^{-1}(x - \mu_2) \tag{18}$$

Taking $log(P(X|Y = 2)) - log(P(X|Y = 1)) = (\mu_2^T - \mu_1^T) \sum^{-1} x + \mu_2^T \sum^{-1} \mu_1 - \mu_1^T \sum^{-1} \mu_0$

Now replacing the values back in the equation found for $P(Y = 1|x)$

$$P(Y = 1|x) = \frac{1}{1 + e^{-log\frac{P(Y=1)}{P(Y=0)} - log(P(X|Y=0)) + log(P(X|Y=1))}} \tag{19}$$

Replacing $\theta = (\mu_2 - \mu_1)^T \sum^{-1}$ and $C = \mu_1^T \sum^{-1} \mu_1 - \mu_2^T \sum^{-1} \mu_2 - \frac{ln(P(Y==2))}{ln(P(Y==1))}$ in equation (9) and re arranging terms. We get $P(Y = 1|X) = \frac{1}{1+e^{-C+\theta^T x}}$

Problem 3.

3.1 (a)
Data after splitting
Size of train data : (433, 14)
Size of test data : (73, 14)

3.1 (b) Histogram(At run time)

3.1 (c) Correlation value of each feature with target (Rounded till 3 decimal places)
[-0.106, 0.208, 0.011, -0.27, -0.666, -0.643, 0.195, -0.593, -0.665, 0.19, -0.529, 0.667

3.1 (d) Normalize Data set

3.2 (a) Linear Regression on test and train data set
MSE_(tested against Test data set) 5.51989210898
MSE_(tested against Train data set) 4.82935844775

3.2 (b) Ridge Regression with lambda = 0.01, 0.1, 1

Lambda--------------------->	0.01	0.1	1.0
MSE_Test Data	5.058	2.631	0.479
MSE_Train Data	4.434	2.341	0.415

3.2 (c) Running 10 fold Cross validation with lambdas from 0.0001 to 1.0
Index fold iteration

	0.0001	0.0010	0.0100	0.1000	1.0000	10.0000
1	2.482464	2.464415	2.294240	1.266921	0.249741	1.516224e+208
2	2.559787	2.541334	2.367523	1.324194	0.232264	1.739510e+208
3	5.624606	5.586039	5.220638	2.925616	0.268493	2.190451e+208
4	2.918608	2.900923	2.734145	1.727598	0.840300	1.546092e+208
5	3.440323	3.418787	3.215264	1.961945	0.612245	1.753030e+208
6	2.654651	2.642130	2.523555	1.779358	0.961945	1.115681e+208
7	1.662144	1.647112	1.506403	0.709193	0.153959	1.767553e+208
8	20.147736	19.963289	18.234491	8.325222	0.853048	2.875456e+209
9	8.363246	8.280602	7.510484	3.290240	0.180236	3.217027e+209
10	10.642447	10.557693	9.758847	4.952334	0.284052	7.548728e+208

3.2 (d) Cross Validation over various values of lambda

	0.0001	0.0010	0.0100	0.1000	1.0000	10.0000
0	6.049601	6.000233	5.536559	2.826262	0.463628	8.010209e+208

```
3.3 (a) MSE after selecting 4 maximum correlated features all at once
MSE = 1.82656701222


3.3 (b) Feature selection one at a time
MSE after first run =  1.09576238549
MSE after Second run =  1.89233146728
MSE after Third run =  2.40756708072
MSE after Fourth run =  2.63206193463


3.3 (c)
MSE after selecting 4 columns with maximum Mutual information =  2.90209297064


3.3(d) Random feature selection:
Minimum MSE is =  1.58792365425  for feature set =  4,5,7,8
Indices are 0 based.


3.4 Polynomial expansion
New Data set feature matrix dimensions :  (433, 105) #One of the column is target itself
MSE_(after polynomial expansion) 3.56284120444
```