Machine Learning: Assigment 2

Anjali Krishna Prasad collaborated with Arpita Agrawal

October 2016

1 Question 1a

$$P(y_n|x_n;b;w) = \begin{cases} \sigma(b+w^TX_n) & \text{if } y_n = 1 \\ 1-\sigma(b+w^TX_n) & \text{if } y_n = 0 \end{cases}$$

$$\text{Where } \sigma_a = \frac{1}{1+e^{-a}} \text{ and } a = b + w^T x_n$$

$$P(y_n|x_n;b;w) = \prod_{i=1}^{n} P(Y = y_i|X = x_i)$$

$$= \sum_n \sigma(b+w^TX_n)^{y_n}[1-\sigma(b+w^TX_n)]^{1-y_n}$$

$$-logP(y_n|x_n;b;w) = -\sum_n [y_n log[\sigma(b+w^TX_n)] + (1-y_n)log[1-\sigma(b+w^TX_n)]]$$

2 Question 1b

$$\text{Append b to w } a = b + w^TX_n -> w^TX_n$$

$$\text{Derivative of } \frac{\partial\sigma(a)}{\partial a} = \frac{\partial}{\partial a}(\frac{1}{1+e^{-a}}) = \frac{-(1+e^{-a})'}{(1+e^{-a})^2} = \frac{e^{-a}}{(1+e^{-a})^2} = (\frac{1}{1+e^{-a}})(1-\frac{1}{1+e^{-a}})$$

$$\frac{\partial\sigma(a)}{\partial a} = \sigma(a)[1-\sigma(a)]$$

$$\frac{\partial log\sigma(a)}{\partial a} = 1-\sigma(a)$$

$$\frac{\partial E(w)}{\partial w} = -\sum_n y_n[1-\sigma(w^TX_n)] + (1-y_n)[\sigma(w^TX_n)]$$

$$-\sum_n X_n[y_n - y_n\sigma(w^TX_n) - \sigma(w^TX_n) + y_n\sigma(w^TX_n)]$$

$$\sum_n X_n[\sigma(w^TX_n) - y_n]$$

$$\text{update rule for w } w^{(t+1)} = w^{(t)} - \eta\sum_n X_n[\sigma(w^TX_n) - y_n]$$

$$H = \frac{\partial^2 E}{\partial ww^T} = \sum_n X_nX_n^T[1-\sigma(w^TX_n)][\sigma(w^TX_n)]$$

$X_nX_n^T$ has to be positive and $\sigma(w^TX_n); 1-\sigma(w^TX_n)$ are probabilities and cannot be < 0

Product of 3 positive terms is positive therefore $v^THv >= 0$

Thus, positive definite. Thus, the cross-entropy error function is convex, with only one global optimum.

3 question 1c

$$P(Y = k|X = x) = \frac{exp^{w_k^Tx}}{1+\sum_1^{k-1} exp^{w_t^Tx}} \text{ where k = 1....K-1}$$

$$P(Y = k|X = x) = \frac{1}{1+\sum_1^{k-1} exp^{w_t^Tx}} \text{ where k = K}$$

$w_K = 0$ The above formula can also be written as

$$P(Y = k|X = x) = \frac{exp^{w_k^Tx}}{exp^{w_K^Tx}+\sum_1^{k-1} exp^{w_t^Tx}}$$

$$P(Y = k|X = x) = \frac{exp^{w_k^Tx}}{\sum_1^{K} exp^{w_t^Tx}}$$

$$P(D) = P(y_n|x_n)$$

$$logP(D) = \sum_{j=1}^{n} logP(y_n|x_n)$$

We will change y_n to $y_{n[y_{n1}y_{n2}.....y_{nk}]^T}$ a K dimensional vector using 1 of K encoding

$$y_{ji}\begin{cases}1 & \text{if } y_j = k\\ 0 & \text{if } otherwise\end{cases}$$

$$-logP(y_n|x_n) = -\sum_{j=1}^{n} log \prod_{i=1}^{k} P(C_k|X_j)^{y_{ji}}$$

$$-logP(y_n|x_n) = -\sum_{j=1}^{n}\sum_{i=1}^{K} y_{ji}logP(C_i|X_j)$$

$$-logP(y_n|x_n) = -\sum_{j=1}^{n}\sum_{i=1}^{K} y_{ji}log\frac{exp^{w_i^Tx_j}}{\sum_1^K exp^{w_t^Tx_j}}$$

$$-logP(y_n|x_n) = -\sum_{j=1}^{n}\sum_{i=1}^{K} y_{ji}[log(exp^{w_i^Tx_j}) - log(\sum_1^K exp^{w_t^Tx_j})]$$

$$-logP(y_n|x_n) = -\sum_{j=1}^{n}\sum_{i=1}^{K} y_{ji}[(w_i^Tx_j) - log(\sum_1^K exp^{w_t^Tx_j})]$$

$$-logP(y_n|x_j) = \sum_{j=1}^{n}\sum_{i=1}^{K} y_{ji}log(\sum_1^K exp^{w_t^Tx_j}) - y_{ji}(w_i^Tx_j)$$

4 Question 1d

Differenciating partially wrt w_i negative log likelihood from above answer

$$-logP(y_n|x_n) = \sum_{j}^{n}\sum_{i=1}^{K} y_{ji}log(\sum_1^K exp^{w_t^Tx_j}) - y_{ji}(w_i^Tx_j)$$

$$-\frac{\partial logP(y_n|x_n)}{\partial w_i} = \frac{\partial}{\partial w_i}(\sum_{j=1}^{n}\sum_{i=1}^{K} y_{ji}log(\sum_1^K exp^{w_t^Tx_j}) - y_{ji}(w_i^Tx_j))$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n}\left[-y_{ji}x_j + \frac{y_{ji}exp^{w_i^Tx_j}x_j}{\sum_1^K exp^{w_t^Tx_j}}\right] = \sum_{i=1}^{K}\sum_{j=1}^{n}\left[-y_{ji}x_j + y_{ji}x_jP(Y=k|X=x)\right]$$

$$\text{update rule for w } w^{(t+1)} = w^{(t)} - \eta\sum_{i=1}^{K}\sum_{j=1}^{n}\left[-y_{ji}x_j + y_{ji}x_jP(Y=k|X=x)\right]$$

5 Question 2a

$$f(x) = \begin{cases}p_1\frac{1}{\sqrt{2\pi}\sigma_1}exp^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}} & \text{if } y_n = 1\\ p_2\frac{1}{\sqrt{2\pi}\sigma_2}exp^{\frac{-(x-\mu_2)^2}{2\sigma_2^2}} & \text{if } y_n = 2\end{cases}$$

$$\text{Likelihood function D } = \prod_{y_n=1}^{2} P(y_n)P(x_n|y_n)$$

$$logP(D) = \sum_{n} logP(x_n, y_n)$$

$$\sum_{n:y_n=1} log(p_1\frac{1}{\sqrt{2\pi}\sigma_1}exp^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}}) + \sum_{n:y_n=2} log(p_2\frac{1}{\sqrt{2\pi}\sigma_2}exp^{\frac{-(x-\mu_2)^2}{2\sigma_2^2}})$$

$$\sum_{n:y_n=1}(logp_1 + log\frac{1}{\sqrt{2\pi}\sigma_1} + \frac{-(x-\mu_1)^2}{2\sigma_1^2}) + \sum_{n:y_n=2}(logp_2 + log\frac{1}{\sqrt{2\pi}\sigma_2} + \frac{-(x-\mu_2)^2}{2\sigma_2^2})$$

We know that$p_2 = 1 - p_1$

$$\sum_{n:y_n=1}(logp_1 + log\frac{1}{\sqrt{2\pi}\sigma_1} + \frac{-(x-\mu_1)^2}{2\sigma_1^2}) + \sum_{n:y_n=2}(log(1-p_1) + log\frac{1}{\sqrt{2\pi}\sigma_2} + \frac{-(x-\mu_2)^2}{2\sigma_2^2})$$

Differenciating logP(D) wrt p_1 and equating to 0

$$\frac{\partial logP(D)}{\partial p_1} = \sum_{n:y_n=1} \frac{1}{p_1} - \sum_{n:y_n=2} \frac{1}{1-p_1} = 0$$

Let $n_1 =$ no of samples in training set with with $y_n = 1$

N is the total number of samples. $N - n_1 =$ no of samples in training set with with $y_n = 2$

$$\frac{n_1}{p_1} - \frac{N-n_1}{(1-p_1)} = 0$$

$$n_1 - p_1 n_1 - p_1 N + p_1 n_1 = 0$$

$$p_1 = \frac{n_1}{N}$$

$$p_2 = 1 - p_1$$

$$p_2 = 1 - \frac{n_1}{N}$$

$$p_2 = \frac{N-n_1}{N}$$

Let $N - n_1 = n_2$

$$p_2 = \frac{n_2}{N}$$

Differenciating logP(D) wrt $\mu_1$and equating to 0

$$\frac{\partial logP(D)}{\partial \mu_1} = \sum_{n:y_n=1} -\frac{1}{2\sigma_1^2}[2(x_n - \mu_1)](-1) = 0$$

$$\sum_{n:y_n=1} 2x_n - 2\mu_1 = 0$$

$$\mu_1 = \frac{\sum_{n:y_n=1} x_n}{n_1}$$

Similarly Differenciating logP(D) wrt $\mu_2$and equating to 0

$$\mu_2 = \frac{\sum_{n:y_n=2} x_n}{n_2}$$

Differenciating logP(D) wrt σ_1 and equating to 0

$$\frac{\partial logP(D)}{\partial \sigma_1} = \sum_{n:y_n=1} -\frac{\sqrt{2\pi}}{\sqrt{2\pi}\sigma_1} - \frac{(x_n - \mu_1)^2}{2\sigma_1^3}(-1) = 0$$

$$\sum_{n:y_n=1} \frac{\sigma_1^2 + (x_n - \mu_1)^2}{\sigma_1^3} = 0$$

$$\sigma_1^2 = \frac{\sum_{n:y_n=1}(x_n - \mu_1)^2}{n_1}$$

$$\sigma_1 = \sqrt{\frac{\sum_{n:y_n=1}(x_n - \mu_1)^2}{n_1}}$$

Similarly differenciating logP(D) wrt σ_2 and equating to 0

$$\sigma_2^2 = \frac{\sum_{n:y_n=2}(x_n - \mu_2)^2}{n_2}$$

$$\sigma_2 = \sqrt{\frac{\sum_{n:y_n=2}(x_n - \mu_2)^2}{n_2}}$$

6 Question 2b

$$P(Y = c_1|X) = \frac{P(Y = c_1)P(X|Y = c_1)}{P(Y = c_1)P(X|Y = c_1) + P(X|Y = c_2)P(Y = c_2)}$$

Dividing numerator and denominator by $P(Y = c_1)P(X|Y = c_1)$

$$P(Y = c_1|X) = \frac{1}{1 + \frac{P(X|Y=c_2)P(Y=c_2)}{P(Y=c_1)P(X|Y=c_1)}}$$

$$\frac{1}{1 + exp(ln\frac{P(Y=c_2)P(X|Y=c_2)}{P(Y=c_1)P(X|Y=c_1)})}$$

$$\frac{1}{1 + exp(ln\frac{P(Y=c_2)}{P(Y=c_1)}) * ln\frac{P(X|Y=c_2)}{P(X|Y=c_1)})}$$

Let $P(Y = c_1) = p_1; P(Y = c_2) = p_2$

$$\frac{1}{1 + exp(ln\frac{p_2}{p_1}) * ln\frac{P(X|Y=c_2)}{P(X|Y=c_1)})}$$

$$\frac{1}{1+exp[(lnp_2-lnp_1)+lnP(X|Y=c_2)-lnP(X|Y=c_1)]}$$

$$P(X|Y=c_1)=\frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}}exp\{-\frac{1}{2}(X-\mu_1)^T\sum^{-1}(X-\mu_1)\}$$

$$P(X|Y=c_2)=\frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}}exp\{-\frac{1}{2}(X-\mu_2)^T\sum^{-1}(X-\mu_2)\}$$

$$\frac{1}{1+exp[(lnp_2-lnp_1)+ln\frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}}exp^{-\frac{1}{2}(X-\mu_2)^T\sum^{-1}(X-\mu_2)}-ln\frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}}exp^{-\frac{1}{2}(X-\mu_1)^T\sum^{-1}(X-\mu_1)}}$$

Lets solve the following part of the denominator separately due to space constraint

$$ln\frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}}exp^{-\frac{1}{2}(X-\mu_2)^T\sum^{-1}(X-\mu_2)}-ln\frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}}exp^{-\frac{1}{2}(X-\mu_1)^T\sum^{-1}(X-\mu_1)}$$

$$ln\frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}}-\frac{1}{2}(X-\mu_2)^T\sum^{-1}(X-\mu_2)-ln\frac{1}{(2\pi)^{\frac{n}{2}}-|\sum|^{\frac{1}{2}}}+\frac{1}{2}(X-\mu_1)^T\sum^{-1}(X-\mu_1)$$

$$-\frac{1}{2}(X-\mu_2)^T\sum^{-1}(X-\mu_2)+\frac{1}{2}(X-\mu_1)^T\sum^{-1}(X-\mu_1)$$

Using property$(A+B)^T=A^T+B^T$ and opening the brackets, we get

$$=(\mu_2-\mu_1)^T\sum^{-1}X+\frac{1}{2}[\mu_1^T\sum^{-1}\mu_1-\mu_2^T\sum^{-1}\mu_2]$$

Getting back the entire exp term from the original equation

$$lnp_2-lnp_1+(\mu_2-\mu_1)^T\sum^{-1}X+\frac{1}{2}[\mu_1^T\sum^{-1}\mu_1-\mu_2^T\sum^{-1}\mu_2]$$

$$\text{Therefore } b=\frac{1}{2}[\mu_1^T\sum^{-1}\mu_1-\mu_2^T\sum^{-1}\mu_2]+lnp_2-lnp_1$$

$$\theta^T=[(\mu_2-\mu_1)^T\sum^{-1}]$$

7 Programming assisngment

3.1

```
Histograms available in the my program.

Pearson's Correlation for column  1  is  -0.387696987621
Pearson's Correlation for column  2  is  0.362987295831
Pearson's Correlation for column  3  is  -0.483067421758
Pearson's Correlation for column  4  is  0.203600144696
Pearson's Correlation for column  5  is  -0.424829675619
Pearson's Correlation for column  6  is  0.690923334973
Pearson's Correlation for column  7  is  -0.390179110401
Pearson's Correlation for column  8  is  0.252420566225
Pearson's Correlation for column  9  is  -0.385491814423
Pearson's Correlation for column  10  is  -0.468849385373
Pearson's Correlation for column  11  is  -0.505270756892
Pearson's Correlation for column  12  is  0.343434137151
Pearson's Correlation for column  13  is  -0.73996982063

3.2
Linear Regression
MSE for training set 20.950144508
MSE for testing set 28.4179164975

Ridge Regression
For lambda =  0.01
MSE for training set 20.9501449001
MSE for testing set 28.4182915618
For lambda =  0.1
MSE for training set 20.9501836546
MSE for testing set 28.42168497
For lambda =  1.0
MSE for training set 20.9539918317
MSE for testing set 28.4573733573

kfold cross validation:
When we dont shuffle we get
```

```
For lambda =  5.4001
MSE for training set 21.0544652787
MSE for testing set 28.6768029918
Shuffling has an effect on the set.
My values of lambda is between 0 and 1
example
winning lambda =  1.0001
MSE for training set after using winning lambda 20.9539925939
MSE for training set after using winning lambda 28.4573774989

winning lambda =  0.9001
MSE for training set after using winning lambda 20.9507674828
MSE for training set after using winning lambda 28.4332314197

3.3
a)Top four correlated columns with the target are 13 6 11 3
MSE after taking top four correlated columns with the target are:
MSE for training set 26.4066042155
MSE for testing set 31.4962025449

b)Select 4 features iteratively and select top 4: 13, 6, 11, 4
MSE_train 25.1060222464
MSE_test 34.6000723135

brute force
MSE_train for brute force:  25.1060222464
MSE_test for brute force:  34.6000723135
columns 4,6,11,13

3.4
MSE ater polynomial feature expansion on train set: 5.05978429711
MSE ater polynomial feature expansion on test set: 14.5553049727
```