STAT 383C: Statistical Modeling I **Fall 2015**

Lecture 12 — September 3

Lecturer: Purnamrita Sarkar *Scribe: Maria Shovgun*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

12.1 How collinearity effect ridge and lasso

Let us us assume that we have m copies of the same feature (similar to add $m-1$ copies).

12.1.1 Ridge

$$\begin{aligned}
&\min\left[(y-X\beta)^T(y-X\beta)+\lambda\beta^T\beta\right]\\
&\hat{\beta}=\left(X^TX+\lambda\mathbf{I}\right)\\
&a=\frac{X^Ty}{X^TX+\lambda}
\end{aligned} \tag{12.1}$$

The same for m copies.

$$\hat{\beta}'=\left(x^Tx\cdot J+\lambda\mathbf{I}\right)^{-1}x^Ty \tag{12.2}$$

Using

$$\begin{aligned}
X^TX&=(x^Tx)\mathbf{J}\\
X^Ty&=(x^Ty)1_m
\end{aligned} \tag{12.3}$$

Eq. **??** becomes

$$\hat{\beta}'=(x^Ty)(c\mathbf{J}+\lambda\mathbf{I})^{-1}_{1_m} \tag{12.4}$$

Now, using the following expressions

$$\begin{aligned}
\mathbf{J}&=1_m\cdot 1_m^T=(m,1_m)\\
c\mathbf{J}&=(cm,1_m)\\
c\mathbf{J}+\lambda\mathbf{I}&=(cm+\lambda,1_m)
\end{aligned} \tag{12.5}$$

Eq. **??** becomes

$$\hat{\beta}'=(x^Ty)\frac{1_m}{cm+\lambda} \tag{12.6}$$

Therefore

$$\hat{\beta}'=ax\frac{c+\lambda}{cm+\lambda}=ax\frac{1+\frac{\lambda}{X^TX}}{m+\frac{\lambda}{X^TX}}\approx\frac{a}{m} \tag{12.7}$$

The last equality holds for small λ.

Now let's solve the HW question.

$$x_{i1} = x_{i2} = x_{i3} = ... = x_{im} = x \tag{12.8}$$

Then

$$\begin{aligned}
&\sum_i \left(y_i - x_i\beta_1 - \sum_{j\neq i} X_{ij}\beta_j \right)^2 + \lambda \sum_j \beta_j^2 \\
&= \sum_i \left(y_i - \sum_j X_{i1}\left(\beta'_{11} + \beta'_{12} + ... + \beta'_{1m}\right) - \sum_j X_{ij}\beta_1 \right)^2 + \lambda \sum_{j=1}^{m} \left(\beta'_{1j}\right)^2 + \sum_{j=2}^{p} \left(\beta'_j\right)^2 \\
&= \left\{ \sum_i \left(y_i - \sum_j X_{i1}\left(\beta'_{11} + \beta'_{12} + ... + \beta'_{1m}\right) - \sum_j X_{ij}\beta_1 \right)^2 + \lambda \sum_{j=1}^{p} \left(\beta'_j\right)^2 \right\} + \sum_{j=1}^{m} \left(\beta'_{j1}\right)^2 - \beta_1'^2
\end{aligned} \tag{12.9}$$

When n is large, the first part of the optimization dominates, and so we can make the following approximate argument. Say we keep the first part pinned at its original optimal value,

$$\beta'_{11} + \beta'_{12} + ... + \beta'_{1m} = a \tag{12.10}$$

Minimizing

$$\min \sum_{k=1}^{m} \left(\beta'_{ij}\right)^2 \tag{12.11}$$

Consequently, all β'_{ij} should be the same, such that

$$\sum_j \beta'_{ij} = a \tag{12.12}$$

12.2 Logistic regression

12.2.1 Multiclass logistic regression

b for every class.

$$P\left(y = k|x_j\right) \propto e^{\beta_m^T x_i} = \frac{e^{\beta_k^T x}}{\sum_i e^{\beta_i^T x_i}} \tag{12.13}$$

Assuming

$$\beta'_k = 0_p \tag{12.14}$$

and applying the transformation

$$\beta'_j = \beta_j - \beta_k, j \neq k \tag{12.15}$$

Eq. **??** becomes

$$P\left(y=k|x_j\right)=\frac{1}{\sum_c e^{(\beta_c-\beta_k)^T x_i}} \tag{12.16}$$

Discriminative, because does not care about x distribution (generated). It cares only about how $\left(Y|x\right)$ was generated.

12.2.2 Generative analogued (Linear discriminant analysis)

$$X_i|y_i=K\sim N(\mu_k,\Sigma_k), \tag{12.17}$$

where μ and Σ are known. how to decide which class involves any particular points (Point classification).

$$P\left(y_i=k\right)=\pi_k \tag{12.18}$$

$$\begin{aligned} &P\left(Y=1|x_\xi\right)\propto P\left(x|Y=1\right)P\left(Y=1\right)\\ &=\frac{1}{(2\pi)^{p/2}\left|\Sigma\right|}\exp\left(-\frac{(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}{2}\right)\pi_1 \end{aligned} \tag{12.19}$$

Decision rule. Classify as 1 if

$$P\left(y_i=1|x\right)\geq P\left(Y=0|x\right) \tag{12.20}$$

Similarly,

$$\log P\left(y_i=1|x\right)\geq \log P\left(Y=0|x\right) \tag{12.21}$$

$$\begin{aligned} &-\frac{1}{2}\log\left|\Sigma_1\right|-\frac{(x-\mu_1)^T\Sigma_1^{-1}(x-\mu 1)}{2}+\log\pi_1\geq\\ &-\frac{1}{2}\log\left|\Sigma_2\right|-\frac{(x-\mu_2)^T\Sigma_2^{-1}(x-\mu 2)}{2}+\log\pi_2 \end{aligned} \tag{12.22}$$

If

$$\Sigma_1=\Sigma_2, \tag{12.23}$$

then we get a linear decision boundary.

$$-\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)+(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)}{2}\geq\log\frac{\pi_2}{\pi_1} \tag{12.24}$$

$$2\mu_1^T\Sigma^{-1}(x-\mu_1)-2\mu_2^T\Sigma^{-1}(x-\mu_2)+\frac{\mu_1^T\Sigma^{-1}\mu_1-\mu_2^T\Sigma^{-1}\mu_2}{2}\geq\log\frac{\pi_2}{\pi_1} \tag{12.25}$$

$$-\left(x-\frac{\mu_1+\mu_2}{2}\right)^T\Sigma^{-1}(\mu_2-\mu_1)\geq\log\frac{\pi_2}{\pi_1} \tag{12.26}$$