Feature Selection Via Logistic Elastic Net Regression in Genetic Cancer Research

A PROJECT REPORT
Presented to Department of Mathematics and Statistics California State University, Long Beach

In Partial Fulfillment of the Requirements for the Degree Master of Science in Mathematics Option in Statistics

Faculty Reviewer:

Kagba Suaray, Ph.D.

By Juan Carlos Apitz
M.S., 2016, California State University, Long Beach

November 20, 2015

Question 1

The pdf of the random variable Y is:

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

Since Y is iid, we obtain the likelihood as follows:

$$L(\mu) = \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

To find the MLE of μ, we proceed to maximize the log of $L(\mu)$ with respect to μ. This is equivalent to maximizing $L(\mu)$ with respect to μ. To accomplish this we can use the first and second derivatives tests from calculus to find the MLE:

$$\ln L(\mu) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2$$

$$\frac{d\ln L(\mu)}{d\mu} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} y_i - n\mu\right) \overset{set}{=} 0$$

$$\hat{\mu}_{mle} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The second derivative test confirms that $\hat{\mu}_{mle}$ is indeed a maximum:

$$\frac{d^2\ln L(\mu)}{d\mu^2} = -\frac{n}{\sigma^2} < 0$$

Question 2

To find $E\left(\bar{Y}\right)$, we use the fact that the random variable Y is distributed iid normal. Therefore we know that:

$$\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Then:

$$E\left(\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = 0$$

$$E\left(\bar{Y}-\mu\right) = 0$$

$$E\left(\bar{Y}\right) = \mu$$

For the variance, we use the definition:

$$V\left(\bar{Y}\right) = E\left(\bar{Y}-E\left(\bar{Y}\right)\right)^2$$

and given that Y is distributed iid normal we know:

$$E\left(\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}-0\right)^2 = 1$$

then:

$$E\left(\bar{Y}-\mu\right)^2 = \frac{\sigma^2}{n}$$

$$E\left(\bar{Y}-E\left(\bar{Y}\right)\right)^2 = \frac{\sigma^2}{n}$$

$$V\left(\bar{Y}\right) = \frac{\sigma^2}{n}$$

Question 3

We can use linear algebra to solve the system of normal equations:

$$\begin{pmatrix} n & \sum_i x_i & \sum_i y_i \\ \sum_i x_i & \sum_i x_i^2 & \sum_i x_i y_i \end{pmatrix}$$

Dividing the first row by n:

$$\begin{pmatrix} 1 & \bar{x} & \bar{y} \\ \sum_i X_i & \sum_i x_i^2 & \sum_i x_i y_i \end{pmatrix}$$

Multiply the first row by $\sum_i x_i$ and subtract it from the second row to get the second pivot:

$$\begin{pmatrix} 1 & \bar{x} & \bar{y} \\ 0 & \sum_i x_i^2 - \bar{x}\sum_i x_i & \sum_i x_i y_i - \bar{y}\sum_i x_i \end{pmatrix}$$

The above result implies that:

$$b_1 = \frac{\sum_i x_i y_i - \bar{y}\sum_i x_i}{\sum_i x_i^2 - \bar{x}\sum_i x_i}$$

To show that:

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

expand $\sum_i (x_i - \bar{x})(y_i - \bar{x})$:

$$\begin{aligned} &= \sum_i (x_i y_i - \bar{y} x_i - \bar{x} y_i + \bar{x}\bar{y}) \\ &= \sum_i x_i y_i - 2\bar{y}\sum_i x_i + n\bar{x}\bar{y} \\ &= \sum_i x_i y_i - 2\bar{y}\sum_i x_i + \bar{y}\sum_i x_i \\ &= \sum_i x_i y_i - \bar{y}\sum_i x_i \end{aligned} \tag{1}$$

Now expand $\sum_i (x_i - \bar{x})^2$:

$$\begin{aligned} &= \sum_i \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\ &= \sum_i x_i^2 - 2\bar{x}\sum_i x_i + n\bar{x}^2 \\ &= \sum_i x_i^2 - 2\bar{x}\sum_i x_i + \bar{x}\sum_i x_i \\ &= \sum_i x_i^2 - \bar{x}\sum_i x_i \end{aligned} \tag{2}$$

Based on result (1) and result (2), we have shown that:

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - \bar{y}\sum_i x_i}{\sum_i x_i^2 - \bar{x}\sum_i x_i} = b_1$$