**Maria Varga**

**Problem Set 2 AQM**

Maximum Likelihood and Logistic Regression

# Problem 1 - Maximum Likelihood for Linear Regression

Consider the linear regression model given by

$$Y = \beta_0 + X\beta_1 + \varepsilon$$

Where errors $\varepsilon_i$are independent normally distributed with mean 0 and variance $\sigma^2$.

For each $y_i$ and a given $x_i$, if $(b_0, b_1, s^2)$ is a guess for the true parameters, we have that

$$P(y_i|x_i; b_0, b_1, s^2) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(y_i - (b_0 + b_1 x_i))^2}{2s^2}\right)$$

Conditional on the predictor X, the response variable Y is independent across observations, that is $y_1$ and $y_2$ are independent given $x_1$ and $x_2$.

As $y_i$ are independent of each other, the likelihood of observing the full dataset is

$$L(b_0, b_1, s^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(y_i - (b_0 + b_1 x_i))^2}{2s^2}\right)$$

and the log-likelihood is given by

$$Log(b_0, b_1, s^2) = \sum_{i=1}^{n} \log\left(P(y_i|x_i; b_0, b_1, s^2)\right)$$

$$= -\frac{1}{n}\log(2\pi) - nlog(s) - \frac{1}{2s^2}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

In the method of maximum likelihood, we choose the parameter values which maximize the likelihood, or, equivalently, maximize the log-likelihood. This include derivate the log-likelihood in respect of each parameters and set it to zero. After some calculus this gives us the following estimators:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\beta_0} = \bar{y} - \beta_1 \bar{x}$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

# Problem 4 – Cross-Validation Intuition

That answer would depend on many other factors not explicit in the problem such as: number of examples, computational power available, nature of variables, min and max variance across variables, how many variables we need to select, purpose of the project and so on. I would try to answer it making some additional assumptions.

Let's assume that we have a reasonable number of examples, computation power is not a problem and we have to initially select 1000 variables through step (a).

If I had to choose option 1 OR 2 (but not a combination of both) I would choose option 1. The main reason is because I don't feel safe performing variable selection dropping variables with lowest standard deviation. This, of course, depend on the nature of variable. Variables with 0 or close to 0 are in fact often useless for the model. But being strongly correlated with the response variable is often more important than having high variance. As we are excluding 90% of variables, there is a good chance of excluding variables that are strongly correlated to the response variable and, consequently, important variable for the model.

Leave-one-out cross-validation is often not the best option to go as it is always subjected to overfitting. But for highly unbalanced data set and for data set which has less than 100 instances leave one out can do a good job. It also allows you to use more of your data, so in theory gives your algorithm the best chance. However, each run is highly correlated with the others.