Machine Learning

Seminar 2 Solution

[Q1 Sample Solution]

The problem is to generalize from the samples and the mapping to be useful to estimate the output for new samples in the future.

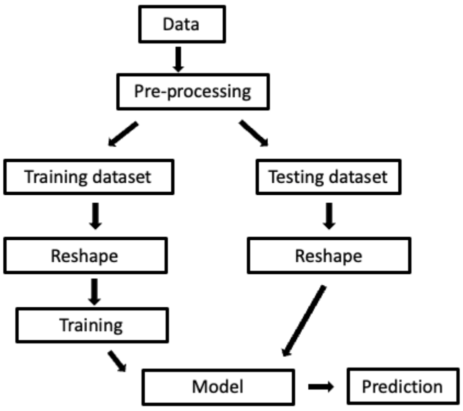
In other word, we are given input samples (x) and output samples f(x) and the problem is to estimate the function f.

[Q2 Sample Solution]

The first one is regression. Input is year. Output is sea ice extent.

The second one is classification. Input is tumour size. Output is tumour type.

[Q3 Sample Solution]



[Q4 Sample Solution]

$y = m \times x + b$

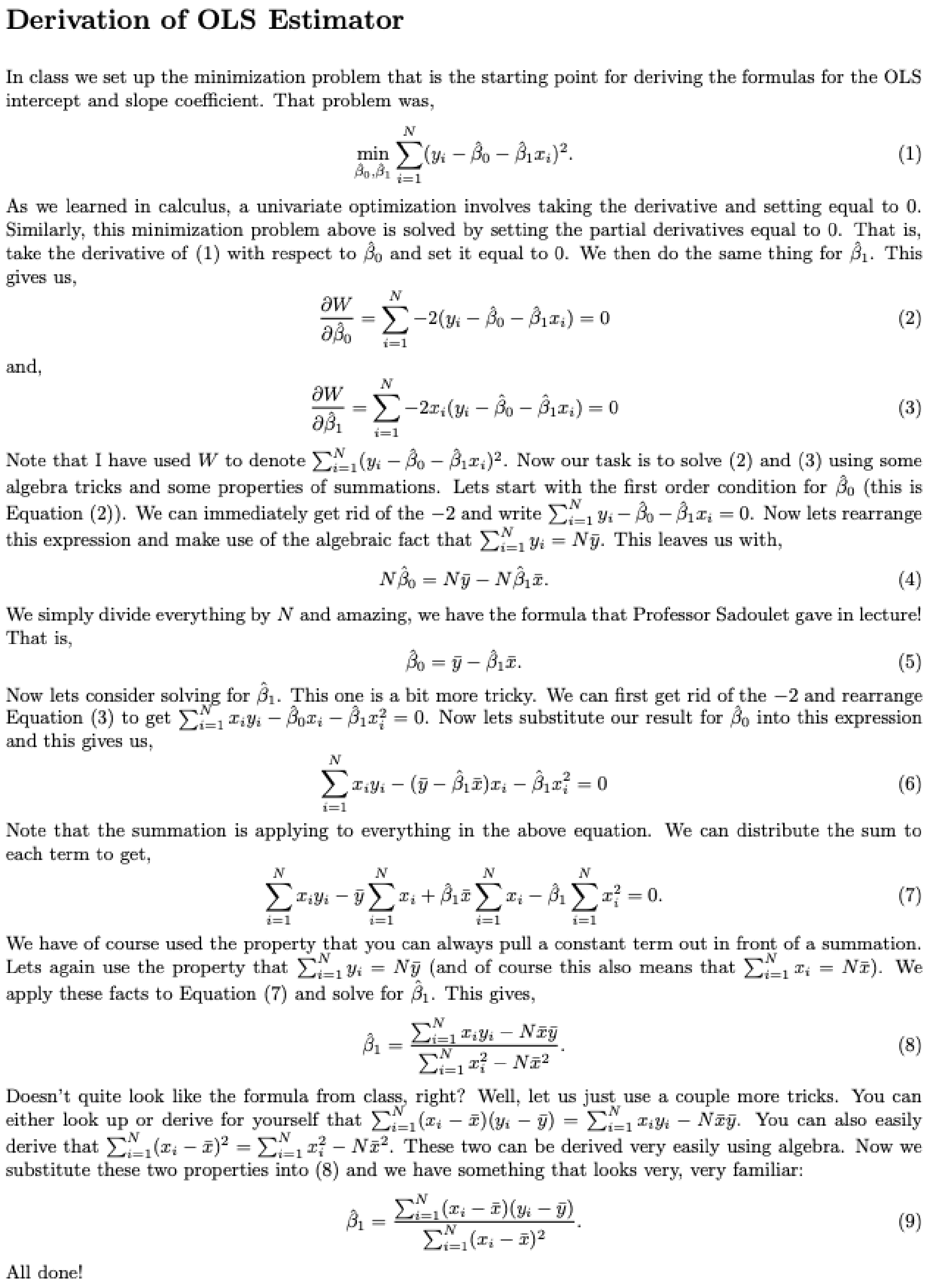
where y is the forecast variable; x is the predictor – size; m, b are parameters;

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M + e$

[Q5 Sample Solution]

Observation	Prediction	Error	Squared Error	Sum of Squared Error	Mean Squared Error
210	201	9	81	166	55.3333
190	188	2	4		
156	147	9	81		

[Q6 Sample Solution]



Derivation of OLS Estimator

In class we set up the minimization problem that is the starting point for deriving the formulas for the OLS intercept and slope coefficient. That problem was,

$$\min_{\hat\beta_0,\hat\beta_1} \sum_{i=1}^{N} (y_i - \hat\beta_0 - \hat\beta_1 x_i)^2. \tag{1}$$

As we learned in calculus, a univariate optimization involves taking the derivative and setting equal to 0. Similarly, this minimization problem above is solved by setting the partial derivatives equal to 0. That is, take the derivative of (1) with respect to $\hat\beta_0$ and set it equal to 0. We then do the same thing for $\hat\beta_1$. This gives us,

$$\frac{\partial W}{\partial \hat\beta_0} = \sum_{i=1}^{N} -2(y_i - \hat\beta_0 - \hat\beta_1 x_i) = 0 \tag{2}$$

and,

$$\frac{\partial W}{\partial \hat\beta_1} = \sum_{i=1}^{N} -2x_i(y_i - \hat\beta_0 - \hat\beta_1 x_i) = 0 \tag{3}$$

Note that I have used W to denote $\sum_{i=1}^{N}(y_i - \hat\beta_0 - \hat\beta_1 x_i)^2$. Now our task is to solve (2) and (3) using some algebra tricks and some properties of summations. Lets start with the first order condition for $\hat\beta_0$ (this is Equation (2)). We can immediately get rid of the -2 and write $\sum_{i=1}^{N} y_i - \hat\beta_0 - \hat\beta_1 x_i = 0$. Now lets rearrange this expression and make use of the algebraic fact that $\sum_{i=1}^{N} y_i = N\bar{y}$. This leaves us with,

$$N\hat\beta_0 = N\bar{y} - N\hat\beta_1\bar{x}. \tag{4}$$

We simply divide everything by N and amazing, we have the formula that Professor Sadoulet gave in lecture! That is,

$$\hat\beta_0 = \bar{y} - \hat\beta_1\bar{x}. \tag{5}$$

Now lets consider solving for $\hat\beta_1$. This one is a bit more tricky. We can first get rid of the -2 and rearrange Equation (3) to get $\sum_{i=1}^{N} x_i y_i - \hat\beta_0 x_i - \hat\beta_1 x_i^2 = 0$. Now lets substitute our result for $\hat\beta_0$ into this expression and this gives us,

$$\sum_{i=1}^{N} x_i y_i - (\bar{y} - \hat\beta_1\bar{x})x_i - \hat\beta_1 x_i^2 = 0 \tag{6}$$

Note that the summation is applying to everything in the above equation. We can distribute the sum to each term to get,

$$\sum_{i=1}^{N} x_i y_i - \bar{y}\sum_{i=1}^{N} x_i + \hat\beta_1\bar{x}\sum_{i=1}^{N} x_i - \hat\beta_1\sum_{i=1}^{N} x_i^2 = 0. \tag{7}$$

We have of course used the property that you can always pull a constant term out in front of a summation. Lets again use the property that $\sum_{i=1}^{N} y_i = N\bar{y}$ (and of course this also means that $\sum_{i=1}^{N} x_i = N\bar{x}$). We apply these facts to Equation (7) and solve for $\hat\beta_1$. This gives,

$$\hat\beta_1 = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}. \tag{8}$$

Doesn't quite look like the formula from class, right? Well, let us just use a couple more tricks. You can either look up or derive for yourself that $\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}$. You can also easily derive that $\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{N} x_i^2 - N\bar{x}^2$. These two can be derived very easily using algebra. Now we substitute these two properties into (8) and we have something that looks very, very familiar:

$$\hat\beta_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{9}$$

All done!

[Q7 Sample Solution]

$\lambda = 0$ implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model

$\lambda = \infty$ implies no feature is considered i.e., as λ closes to infinity it eliminates more and more features

The bias increases with increase in λ

variance increases with decrease in λ