Lecture Three

February 23, 2016

1 Features

$$y = X * \beta \tag{1}$$

Features may not be independent and what is the smallest subset of the features that can be used. Can the number of features be reduced? It is useful to reduce the feature set to remove noise from the data. It is also more elegant to do more approximations with less features.

You are able to use polynomials to fit to your training data. You can always perfectly fit to your training set, however this will over-fit to the training set and can cause more error than a reduced estimation of the training set.

If you are were to plot the degree of the polynomial with respect to the Error of the training set and the test set you will find a correlation.

1.1 subset selection

All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.

If you use zero features you guess is a horizontal line at a_0. You compute error on every length and combination of the features set.

1.2 Minimize Coefficients

$$E = (\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \tag{2}$$

$$\vec{\beta}^T\vec{\beta} = \beta_0^2 + \beta_1^2 + \cdots + \beta_n^2 \tag{3}$$

1.2.1 Ridge regression

$$\vec{\beta} = (X^TX + \lambda I)^{-1}X^T\vec{y} \tag{4}$$

Increasing the amount of λ you are able to regulate your data and prevent an invertible matrix. By adding the identity matrix to the data you are helping the data fit into the dimensionality of the space so that the matrix is invertible.

1.2.2 the Lasso

You use the same error function as in the LSE method. The addition is the a constraint on the minimization equation. The constraints define a region in space that you are allowed to have a solution. The constraint is a feasible region(convex). You use linear programming to optimize a convex region. If the minimum exists within the region then you can take the derivative and set it to zero. If the minimum is outside the feasible region then Quadratic programming is used.

$$E = (\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})\text{subject to}\lambda\sum_{j=0}^{n}|\beta_j| \leq t \tag{5}$$