

Regression

October 19, 2006

1 Linear Prediction

In regression, the output space \mathcal{Y} is just \mathcal{R} and input space \mathcal{X} is a \mathcal{R}^p . So here x^i is p -dimensional.

At the point x we form a linear predictor \hat{y} :

$$\hat{y} = \beta^T x$$

Our task is to fit β using a training set $T = \{(x^i, y^i) | i = 1, \dots, n\}$, such that we minimize our *generalization error* (our error on new samples).

2 Least Squares Fitting

We will form an estimate $\hat{\beta}$ by minimizing the error:

$$\sum_{i=1}^n (y^i - \beta^T x^i)^2$$

With the training set, it is convenient to define the vector Y as

$$Y = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{bmatrix}$$

so $Y \in \mathcal{R}^n$. Also, define the matrix X as

$$X = \begin{bmatrix} (x^1)^T \\ (x^2)^T \\ \vdots \\ (x^n)^T \end{bmatrix}$$

so X is a $n \times p$ matrix. The rows of X are just the points from the training set. This matrix is often referred to as the *design matrix*.

Hence, we can write the error as:

$$\sum_{i=1}^n (y^i - \widehat{\beta}^T x^i)^2 = \|Y - X\beta\|^2 = \|Y - \widehat{Y}\|^2$$

where $\|v\|^2$ denotes the norm $\sum_{i=1}^n (v^i)^2$ and we have defined our prediction vector as:

$$\widehat{Y} = X\widehat{\beta}$$

One can take derivatives to solve for this equation.

There is another simple geometric method for solving this equation. Let X_j be the j -th column in X , so

$$X_j = \begin{bmatrix} x_j^1 \\ x_j^2 \\ \vdots \\ x_j^n \end{bmatrix}$$

so this is the vector of the j -th feature value (over the training set) and X_j is an element of \mathcal{R}^n . Our prediction \widehat{Y} lies in the span of X_1 to X_p .

Hence, the best \widehat{Y} is the the projection of Y into the space spanned by $S = \{X_1, X_2, \dots, X_p\}$. If \widehat{Y} is this projection, then the orthogonal component is $Y - \widehat{Y}$ and it must be orthogonal to S , i.e.

$$X_j^T (Y - \widehat{Y}) = 0$$

for all j . These equations are known as the *normal equations*.

We can equivalently write this as:

$$X^T (Y - \widehat{Y}) = 0$$

where now the left hand side is a p dimensional vector and the 0 on the right hand side is really 0 in all p dimensions.

2.1 Estimating β

In other words, we have:

$$X^T (Y - X\widehat{\beta}) = 0$$

Solving this leads to the estimate

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

which is equivalent to:

$$\begin{aligned}\widehat{\beta} &= \left(\sum_{i=1}^n x^i (x^i)^T\right)^{-1} \sum_{i=1}^n y^i x^i \\ &= \left(\frac{1}{n} \sum_{i=1}^n x^i (x^i)^T\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n y^i x^i\right)\end{aligned}$$

where the last quantity is stated in terms of normalized quantities.

If we consider n to be large, this essentially shows that the optimal value of β is:

$$\beta = \mathbb{E}[xx^T]^{-1} \mathbb{E}[yx]$$

Note that if x has 0 mean, then

$$\text{cov}(x) = \mathbb{E}[xx^T]$$

i.e. $\mathbb{E}[xx^T]$ is the covariance matrix.