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 $\beta$ using a training set $T = \{(x^i, y^i)|i = 1, \ldots, n\}$, 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:

$$
\sum_{i=1}^{n} (y_i - \beta^T x_i)^2 = \|Y - X\beta\|^2 = \|Y - \hat{Y}\|^2
$$

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

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

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

$$
X_j = \begin{bmatrix}
x_1^j \\
x_2^j \\
\vdots \\
x_n^j
\end{bmatrix}
$$

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

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

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

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

We can equivalently write this as:

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

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 $\beta$

In other words, we have:

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

Solving this leads to the estimate

$$
\hat{\beta} = (X^T X)^{-1} X^T Y
$$
which is equivalent to:

\[
\hat{\beta} = \left( \sum_{i=1}^{n} x_i (x_i)^T \right)^{-1} \sum_{i=1}^{n} y_i x_i \\
= \left( \frac{1}{n} \sum_{i=1}^{n} x_i (x_i)^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} y_i x_i \right)
\]

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 \( \beta \) is:

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

Note that if \( x \) has 0 mean, then

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

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