Regression, The Orthogonal Case

October 22, 2006

1 Orthogonalizing x

Note that for any symmetric matrix Σ, we can write this matrix as:

$$\Sigma = R^T D R$$

where R is a rotation and D is diagonal.
For the matrix $\mathbb{E}[xx^T]$, consider doing a rotation such that after this rotation the matrix is diagonal. Hence, let us assume that

$$\mathbb{E}[xx^T] = D$$

where D is a diagonal matrix.
In other words,

$$\mathbb{E}[x_i^2] = \sigma_2^2$$

and

$$\mathbb{E}[x_i x_j] = 0$$

for $i \neq j$.

2 The Regret

The optimal β for this case is:

$$\beta_i = \mathrm{E}[yx_i]/\sigma_i^2$$

Let us define the regret of an estimate $\widehat{\beta}$ as:

$$\text{Regret}(\widehat{\beta}) = \mathbb{E}[(y - \widehat{\beta}^T x)^2] - \mathbb{E}[(y - \beta^T x)^2]$$

Remark 2.1: The regret takes the simple form:

$$\text{Regret}(\widehat{\beta}) = \sum_{i=1}^{p} (\widehat{\beta}_i - \beta_i)^2 \sigma_i^2$$

and in the *orthonormal case*, where $\sigma_i = 1$, we have:

$$\text{Regret}(\widehat{\beta}) = \sum_{i=1}^{p} (\widehat{\beta}_i - \beta_i)^2 = ||\widehat{\beta} - \beta||^2$$

Let us prove this in the orthogonal case. Let us start by writing the error as:

$$\begin{aligned}
\mathbb{E}[(y - \widehat{\beta}^T x)^2] &= \mathbb{E}[y^2] - 2\mathbb{E}[y\widehat{\beta}^T x] + \mathbb{E}[(\widehat{\beta}^T x)^2] \\
&= \mathbb{E}[y^2] - 2\widehat{\beta}^T \mathbb{E}[yx] + \widehat{\beta}^T \mathbb{E}[xx^T]\widehat{\beta} \\
&= \mathbb{E}[y^2] - 2\widehat{\beta}^T \beta + \widehat{\beta}^T \widehat{\beta}
\end{aligned}$$

Using this equation again for β gives:

$$\begin{aligned}
\mathbb{E}[(y - \beta^T x)^2] &= \mathbb{E}[y^2] - 2\beta^T \beta + \beta^T \beta \\
&= \mathbb{E}[y^2] - \beta^T \beta
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\mathbb{E}[(y - \widehat{\beta}^T x)^2] - \mathbb{E}[(y - \beta^T x)^2] &= (\mathbb{E}[y^2] - 2\widehat{\beta}^T \beta + \widehat{\beta}^T \widehat{\beta}^T) - (\mathbb{E}[y^2] - \beta^T \beta) \\
&= \widehat{\beta}^T \widehat{\beta}^T - 2\widehat{\beta}^T \beta + \beta^T \beta \\
&= ||\widehat{\beta} - \beta||^2
\end{aligned}$$

3 The Expected Regret

Let us continue working in the orthonormal case (where $\sigma_i = 1$).
With some training set T, let us consider the empirical estimator of $\widehat{\beta}$,

$$\widehat{\beta}_i = \frac{1}{n} \sum_{j=1}^{n} y^j x_i^j$$

This is related to the earlier estimator

$$\widehat{\beta} = (\frac{1}{n} \sum_j x^j (x^j)^T)^{-1} (\frac{1}{n} \sum_j y^j x^j)$$

except that we are assuming that instead of using the empirical estimate of $\frac{1}{n}\sum_j x^j (x^j)^T$, we are using the identity matrix (since we assume we know $\mathbb{E}[xx^T]$ is the identity matrix).

Remark 3.1: Under a boundedness assumptions, that y is bounded in $[-1, 1]$ (technically, we only the need the weaker condition that $\mathbb{E}[y^2] \leq 1$), we have:

$$\mathbb{E}_T[\text{Regret}(\widehat{\beta})] \leq p/n$$

where the expectation is taken with respect to the training set.

Note this equation says that we need more samples n to get a small error as the number of dimensions p increases.

To prove this, first note that

$$\mathbb{E}[\widehat{\beta}_i] = \beta_i$$

i.e. we are using an unbiased estimator of β_i. We are interested in bouding:

$$\sum_{i=1}^{p} (\widehat{\beta}_i - \beta_i)^2$$

which is just sum of variances of $\widehat{\beta}_i$. Cruedely, speaking we know the variance drops as $1/n$, and so the sum behaves as p/n.

To see this more formally, we proceed as follows:

$$\mathbb{E}(yx_i - \beta_i)^2 = \mathbb{E}[(yx_i)^2] - \beta^2 \leq \mathbb{E}[x_i^2] - \beta^2 \leq 1$$

Not this is just the variance of the estimator $\widehat{\beta}_i$, when $n = 1$. Hence, for n points, the usual result is that this drops as $1/n$, i.e.:

$$\begin{aligned}
\mathbb{E}[(\widehat{\beta}_i - \beta_i)^2] &= \mathbb{E}[((\frac{1}{n}\sum_{j=1}^{n} y^j x_i^j) - \beta_i)^2] \\
&= \mathbb{E}[(\frac{1}{n}\sum_{j=1}^{n} (y^j x_i^j - \beta_i))^2] \\
&= \frac{1}{n^2}\mathbb{E}[\sum_{j=1}^{n} (y^j x_i^j - \beta_i)^2] \\
&\leq \frac{1}{n^2}\mathbb{E}[\sum_{j=1}^{n} 1] \\
&= 1/n
\end{aligned}$$

Summing these completes the proof.