Math 345 - Notes
Linear Regression II
October 22, 2018

Linear Regression (extension)

Suppose we extend the linear regression model to assume that the difference between the predicted output $\mathbf{w}^T\mathbf{x}$ and the real output y occurs due to some error, or *noise* ϵ. That is,

$$y = \mathbf{w}^T\mathbf{x} + \epsilon.$$

Let us also make the assumption that given N data points $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$,

$$y_1 = \mathbf{w}^T\mathbf{x}_1 + \epsilon_1 \qquad , \ldots , \qquad y_N = \mathbf{w}^T\mathbf{x}_N + \epsilon_N,$$

with all ϵ_k independent and identically distributed ($1 \leq k \leq N$), with mean zero and variance σ^2. Typically, these errors are thought to be normally distributed.

We will use the $N \times (d+1)$-dimensional matrix X with rows $\mathbf{x}_k^T$, and the vector $\mathbf{y} = [y_1, \ldots, y_N]^T$. Let $\vec{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^T$. Then we need find $\mathbf{w}_{\text{lin}}$, the set of coefficients that lead to a least squares solution. We derived that, if X^TX is invertible, then

$$\mathbf{w}_{\text{lin}} = (X^TX)^{-1}X^T\mathbf{y}.$$

Recall that this model is probabilistic in nature, that is, each pair $(\mathbf{x}, y)$ occurs with joint probability $P(\mathbf{x}, y)$, a probability with unknown distribution. The (least squares) error resulting from approximation of the target function with a hyperplane given by coordinate vector $\mathbf{w}$, is given by the expectation

$$E_{out}(\mathbf{w}) = \mathbb{E}\left[(\mathbf{w}^T\mathbf{x} - y)^2\right] = \mathbb{E}[\epsilon^2] = Var(\epsilon) = \sigma^2.$$

The error resulting from the approximation of the hyperplane using the sample data points is:

$$E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{w}^T\mathbf{x}_k - y_k)^2.$$

We will show that the two errors are close. Consider the sample input/output:

$$\mathbf{y} = X\mathbf{w} + \vec{\epsilon},$$

and the sample predicted output, after running the Linear Regression Algorithm:

$$\begin{aligned}
\hat{\mathbf{y}} &= X\mathbf{w}_{\text{lin}} = X(X^TX)^{-1}X^T\mathbf{y} \\
&= X(X^TX)^{-1}X^T(X\mathbf{w} + \vec{\epsilon}) \\
&= X(X^TX)^{-1}(X^TX)\mathbf{w} + X(X^TX)^{-1}X^T\vec{\epsilon} \\
&= X\mathbf{w} + X(X^TX)^{-1}X^T\vec{\epsilon}.
\end{aligned}$$

Let $H = X(X^TX)^{-1}X^T$. The matrix $H = [h_{ij}]_{1\le i,j\le d+1}$ has important properties used in proving the error bounds.

Proposition 0.1. *Suppose X is an $N \times (d+1)$ matrix, with X^TX invertible. Then the matrix $H = X(X^TX)^{-1}X^T$ satisfies the following:*
(a) H is symmetric, that is $H^T = H$,

(b) trace$(H) = d+1$,

(c) $H^k = H$ for all $k \in \mathbb{Z}_+$,

(d) $(I-H)^k = I-H$ for all $k \in \mathbb{Z}_+$.

Proof. First, let us check on the dimensions of H. Since X has dimensions $N \times (d+1)$, $(X^TX)^{-1}$ has dimensions $(d+1)\times(d+1)$ and so H has dimensions $N \times N$.
(a) We check directly:

$$H^T = [X(X^TX)^{-1}X^T]^T = [X^T]^T[(X^TX)^{-1}]^TX^T = X[(X^TX)^T]^{-1}X^T = X(X^TX)^{-1}X^T.$$

(b) We will use the fact that trace$(AB) = $ trace(BA) for matrices A, B. Then

$$\text{trace}(H) = \text{trace}(X(X^TX)^{-1}X^T) = \text{trace}((X^TX)^{-1}X^TX) = \text{trace}(I_{d+1}) = d+1.$$

(c) We will show $H^2 = H$. The general case will follow via mathematical induction.

$$H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X\, I_{d+1}\,(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T.$$

(d) As in part c, we show $(I-H)^2 = I-H$, with the general case following via induction:

$$(I-H)^2 = (I-H)(I-H) = I^2 - 2H + H^2 = I - 2H + H = I - H.$$

$\square$

Now the sample output $\mathbf{y}$ and the sample predicted output $\hat{\mathbf{y}}$ are:

$$\mathbf{y} = X\mathbf{w} + \vec{\epsilon},$$

$$\hat{\mathbf{y}} = H\mathbf{y} = X\mathbf{w} + H\vec{\epsilon}.$$

The sample error is given by the vector

$$\mathbf{y} - \hat{\mathbf{y}} = \vec{\epsilon} - H\vec{\epsilon} = (I-H)\vec{\epsilon}. \tag{1}$$

Computing the error due to sampling is

$$
\begin{aligned}
E_{in}(\mathbf{w}_{\text{lin}}) &= \tfrac{1}{N}\textstyle\sum_{k=1}^{N}(\mathbf{w}_{\text{lin}}^T\mathbf{x}_k - y_k)^2 && \text{def. of } E_{in} \\
&= \tfrac{1}{N}\|X\mathbf{w}_{\text{lin}} - \mathbf{y}\|^2 && \text{re-write using matrix notation} \\
&= \tfrac{1}{N}\|\hat{\mathbf{y}} - \mathbf{y}\|^2 && \text{substitute using def. of } \hat{\mathbf{y}} \\
&= \tfrac{1}{N}\|(I-H)\vec{\epsilon}\|^2 && \text{simplification from (1)} \\
&= \tfrac{1}{N}\vec{\epsilon}^T(I-H)^T(I-H)\vec{\epsilon} && \text{def. of norm} \\
&= \tfrac{1}{N}\vec{\epsilon}^T(I-H)^2\vec{\epsilon} && \text{Proposition 0.1, part a.} \\
&= \tfrac{1}{N}\vec{\epsilon}^T(I-H)\vec{\epsilon} && \text{Proposition 0.1, part d.} \\
&= \tfrac{1}{N}\left[\|\vec{\epsilon}\|^2 - \vec{\epsilon}^T H\vec{\epsilon}\right] && \text{simplification and def. of norm}
\end{aligned}
$$

Averaging over the randomness in error $\vec{\epsilon}$,

$$
\begin{aligned}
\mathbb{E}[E_{in}(\mathbf{w}_{\text{lin}})] &= \frac{1}{N}\mathbb{E}\left[\|\vec{\epsilon}\|^2 - \vec{\epsilon}^T H\vec{\epsilon}\right] \\
&= \frac{1}{N}\mathbb{E}\left[\epsilon_1^2 + \cdots + \epsilon_N^2\right] - \\
&\quad \frac{1}{N}\mathbb{E}\left[(\epsilon_1 h_{11} + \epsilon_2 h_{21} + \cdots + \epsilon_N h_{N1})\epsilon_1 + \cdots + (\epsilon_1 h_{1N} + \epsilon_2 h_{2N} + \cdots + \epsilon_N h_{NN})\epsilon_N\right] \\
&= \frac{1}{N}(\mathbb{E}\left[\epsilon_1^2\right] + \cdots + \mathbb{E}\left[\epsilon_N^2\right]) - \frac{1}{N}(h_{11}\mathbb{E}\left[\epsilon_1^2\right] + \cdots + h_{NN}\mathbb{E}\left[\epsilon_N^2\right]) \\
&= \frac{1}{N}\left[(N\sigma^2) - \sigma^2(h_{11} + \cdots + h_{NN})\right] \\
&= \frac{1}{N}\left[N\sigma^2 - \sigma^2\text{trace}(H)\right] \\
&= \sigma^2\left(1 - \frac{d+1}{N}\right).
\end{aligned}
$$

Where in the third step we used the fact that since errors are independent and of mean zero, if $i \neq j$, $\mathbb{E}[\epsilon_i\epsilon_j] = \mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_j] = 0$, so the only non-zero terms in the sum are $\mathbb{E}[h_{ii}\epsilon_i^2]$. Therefore,

$$
\mathbb{E}[E_{in}(\mathbf{w}_{\text{lin}})] - E_{out}(\mathbf{w}_{\text{lin}}) = O\left(\frac{d}{N}\right).
$$