Regression and Error

CS 536 Machine Learning
Nitin Reddy Karolla

March 15, 2019

A Small Regression Example Consider regression in one dimension, with a data set $(x_i, y_i)_{i=1,\ldots,m.}$

1. Find a linear model that minimizes the training error, i.e., $\hat{w}$ and $\hat{b}$ to minimize

$$\sum_{i=1}^{m}(\hat{w}x_i + \hat{b} - y_i)^2 \tag{1}$$

To find linear model, that minimizes the training error, we need to find w and b that minimizes the cost function (F).

Let us differentiate w.r.t. w :

$$\begin{aligned}\frac{d}{dw}\sum_{i=1}^{m}(\hat{w}x_i + \hat{b} - y_i)^2 &= 2\sum_{i=1}^{m}(\hat{w}x_i + \hat{b} - y_i).x_i \\ &= 2(\hat{w}\sum_{i=1}^{m}x_i^2 + \hat{b}\sum_{i=1}^{m}x_i - \sum_{i=1}^{m}x_iy_i) \\ &= 2(m\hat{w}E[X^2] + m\hat{b}E[X] - mE[XY]) \\ &= 2m(\hat{w}E[X^2] + \hat{b}E[X] - E[XY])\end{aligned}$$

Now, differentiating w.r.t. b :

$$\begin{aligned}\frac{d}{db}\sum_{i=1}^{m}(\hat{w}x_i + \hat{b} - y_i)^2 &= 2\sum_{i=1}^{m}(\hat{w}x_i + \hat{b} - y_i) \\ &= 2(\hat{w}\sum_{i=1}^{m}x_i + \hat{b}\sum_{i=1}^{m} - \sum_{i=1}^{m}y_i) \\ &= 2(m\hat{w}E[X] + m\hat{b} - mE[Y]) \\ &= 2m(\hat{w}E[X] + \hat{b} - E[Y])\end{aligned}$$

On equating previous equation to 0 :

$$2m(\hat{w}E[X] + \hat{b} - E[Y]) = 0$$
$$\hat{b} = E[Y] - \hat{w}E[X]$$

Substituting the value of b into the first differentiation and equating it to 0, we get

$$2m(\hat{w}E[X^2] + \hat{b}E[X] - E[XY]) = 0$$
$$\hat{w}E[X^2] + (E[Y] - \hat{w}E[X])E[X] - E[XY] = 0$$
$$\hat{w}(E[X^2] - E[X]^2) + E[X]E[Y] - E[XY] = 0$$
$$\hat{w} = \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2}$$

We also know that,

$$Var(X) = E[X^2] - E[X]^2$$
$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

Therefore,

$$\hat{w} = \frac{Cov(X, Y)}{Var(X)}$$

Now, substituting back the value of w, to obtain b as :

$$\hat{b} = E[Y] - \frac{Cov(X, Y)}{Var(X)}E[X]$$

Now, the linear model that minimizes the training error has weight and bias as calculated above.

2. Assume there is some true linear model, such that $y_i = wx_i + b + \epsilon_i$, where noise variables ϵ_i are i.i.d. with $\epsilon_i N(0, 2)$. Argue that the estimators are unbiased, i.e., $E[\hat{w}] = w$ and $E[\hat{b}] = b$. What are the variances of these estimators?

We have seen from the previous question that,

$$\begin{aligned}
\hat{w} &= \frac{Cov(X, Y)}{Var(X)} \\
&= \frac{\sum_{i=1}^{m}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} - \frac{\sum_{i=1}^{m} \bar{y}(x_i - \bar{x})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m}(x_i - \bar{x})^2}
\end{aligned}$$

We have in the above equation, second term turning out to zero. Let us assign

$$\lambda_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$$

where,

$$\sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} \frac{(x_i - \bar{x})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} = 0$$
$$\sum_{i=1}^{m} \lambda_i (x_i - \bar{x}) = \sum_{i=1}^{m} \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{m}(x_i - \bar{x})^2} = 1$$

Now,

$$\begin{aligned}
\hat{w} &= \sum_{i=1}^{m} y_i \lambda_i \\
E[\hat{w}] &= E[\sum_{i=1}^{m} y_i \lambda_i] \\
&= \sum_{i=1}^{m} \lambda_i E[y_i] \\
&= \sum_{i=1}^{m} \lambda_i E[w x_i + b + \epsilon_i] \\
&= \sum_{i=1}^{m} \lambda_i (w x_i + b + E[\epsilon_i]) \\
&= (w \sum_{i=1}^{m} \lambda_i x_i + b \sum_{i=1}^{m} \lambda_i + 0) \\
&= w
\end{aligned}$$

Similarly,

$$\begin{aligned}
E[\hat{b}] &= E[\bar{y}] - E[\hat{w}]\bar{x} \\
&= b
\end{aligned}$$

Now, let us calculates the variances for the estimators.

$$
\begin{aligned}
\hat{w} &= \sum_{i=1}^{m} \lambda_i (w x_i + b + \epsilon_i) \\
&= w \sum_{i=1}^{m} \lambda_i x_i + b \sum_{i=1}^{m} \lambda_i + \sum_{i=1}^{m} \lambda_i \epsilon_i \\
&= w + mb\bar{x} + \sum_{i=1}^{m} \lambda_i \epsilon_i \\
&= constant + \sum_{i=1}^{m} \lambda_i \epsilon_i \\
Var[\hat{w}] &= \sum_{i=1}^{m} \lambda_i^2 Var[\epsilon_i] \\
&= \sigma^2 \sum_{i=1}^{m} \lambda_i^2 \\
&= \sigma^2 \sum_{i=1}^{m} (\frac{(x_i - \bar{x})}{\sum_{i=1}^{m} (x_i - \bar{x})^2})^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^{m} (x_i - \bar{x})^2}
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\hat{b} &= \frac{1}{m} \sum_{i=1}^{m} y_i - \bar{x} \sum_{i=1}^{m} \lambda_i y_i \\
&= \sum_{i=1}^{m} (\frac{1}{m} - \bar{x}\lambda_i) y_i \\
&= \sum_{i=1}^{m} (w x_i + b)(\frac{1}{m} - \bar{x}\lambda_i) + \sum_{i=1}^{m} (\frac{1}{m} - \bar{x}\lambda_i)\epsilon_i \\
&= constant + \sum_{i=1}^{m} (\frac{1}{m} - \bar{x}\lambda_i)\epsilon_i
\end{aligned}
$$

$$
\begin{aligned}
Var[\hat{b}] &= \sum_{i=1}^{m}(\frac{1}{m} - \bar{x}\lambda_i)Var[\epsilon_i] \\
&= \sigma^2 \sum_{i=1}^{m}(\frac{1}{m} - \bar{x}\lambda_i) \\
&= \sigma^2(\frac{1}{m^2} - \frac{2}{m}\bar{x}\sum_{i=1}^{m}\lambda_i + \bar{x}^2\sum_{i=1}^{m}\lambda_i^2) \\
&= \sigma^2 \frac{\sum_{i=1}^{m} x_i^2}{m\sum_{i=1}^{m}(x_i - \bar{x})^2}
\end{aligned}
$$

3. Assume that each x value was sampled from some underlying distribution with expectation E[x] and variance Var(x). Argue that in the limit, the error on $\hat{w}$ and $\hat{b}$ are approximately

$$
\begin{aligned}
Var(\hat{w}) &\approx \frac{\sigma^2}{m}\frac{1}{Var(x)} \\
Var(\hat{b}) &\approx \frac{\sigma^2}{m}\frac{E[x^2]}{Var(x)}
\end{aligned}
$$

Given the underlying distribution with expectation E[x] and variance Var(x) for x, we have

$$
\begin{aligned}
Var[\hat{w}] &= \frac{\sigma^2}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{mVar[x]}
\end{aligned}
$$

Also,

$$
\begin{aligned}
Var[\hat{b}] &= \sigma^2 \frac{\sum_{i=1}^{m} x_i^2}{m\sum_{i=1}^{m}(x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{m}\frac{E[x^2]}{Var(x)}
\end{aligned}
$$

4. Argue that re centering the data ($x\prime_i = x_i - \mu$) and doing regression on the re-centered data produces the same error on $\hat{w}$ but minimizes the error on $\hat{b}$ when $\mu = E[x]$ (which we approximate with the sample mean).

Variance of $\hat{w}$ based on re centered data:

$$\begin{aligned} Var[\hat{w\prime}] &= \frac{\sigma^2}{\sum_{i=1}^{m}(x\prime_i - \bar{x\prime})^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^{m}((x_i - \mu) - (x_i - \mu))^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^{m}(x_i - \mu)^2} \\ &= \frac{\sigma^2}{mVar[x]} \end{aligned}$$

As we can see, the variance or error of $\hat{w}$ after re centering the data is same as before.

Variance of $\hat{b}$ based on re centered data,

$$\begin{aligned} Var[\hat{b\prime}] &= \frac{\sigma^2}{m}\frac{E[x\prime^2]}{Var(x\prime)} \\ &= \sigma^2\frac{\sum_{i=1}^{m} x\prime_i^2}{m\sum_{i=1}^{m}(x\prime_i - \bar{x\prime})^2} \\ &= \sigma^2\frac{\sum_{i=1}^{m}(x_i - \mu)^2}{m\sum_{i=1}^{m}((x_i - \mu) - (x_i - \mu))^2} \\ &= \sigma^2\frac{\sum_{i=1}^{m}\frac{(x_i - \mu)^2}{m}}{mVar[x]} \\ &= \sigma^2\frac{\sum_{i=1}^{m}\frac{(x_i - E[x])^2}{m}}{mVar[x]} \\ &= \sigma^2\frac{Var[x]}{mVar[x]} \\ &= \frac{\sigma^2}{m} \end{aligned}$$

On comparing the 2 variance, we see that variance of re-centered data is smaller than

the original one for bias.

$$\begin{aligned} Var[\hat{b\prime}] - Var[\hat{b}] &= \frac{\sigma^2}{m} - \frac{\sigma^2}{m}\frac{E[x^2]}{Var(x)} \\ &= \frac{\sigma^2}{m}(1 - \frac{E[x^2]}{Var(x)}) \\ &= \frac{\sigma^2}{m}(\frac{Var(x) - E[x^2]}{Var(x)}) \\ &= \frac{\sigma^2}{m}(\frac{E[x^2] - E[x]^2 - E[x^2]}{Var(x)}) \\ &= \frac{\sigma^2}{m}(\frac{-E[x]^2}{Var(x)}) \\ &= -\frac{\sigma^2}{m}\frac{E[x]^2}{Var(x)} \end{aligned}$$

Clearly, we can see that variance of re-centered data is smaller than the original one.

5. Solution to this problem is present in the notebook file attached.

6. Intuitively, why is there no change in the estimate of the slope when the data is shifted?

   As, we saw there is no change in estimate of the slope when the data is shifted because the shift in data by a constant value only shifts the data. All the points in the data just change their positions only, but are still in same orientation or position with respect to others are earlier. In others words, the relative positions of these points does not change, so the slope of the data does not change. They just shift in space to a new location.

7. Consider augmenting the data in the usual way, going from one dimensions to two dimensions, where the first coordinate of each x is just a constant 1. Argue that taking $\Sigma = X^T X$ in the usual way, we get in the limit that

$$\Sigma \to m \begin{bmatrix} 1 & E[x] \\ E[x] & E[x^2] \end{bmatrix} \tag{2}$$

   Show that re-centering the data ($\Sigma\prime = (X\prime)^T(X\prime)$, taking $x\prime_i = x_i - \mu$), the condition number $K(\Sigma\prime)$ is minimized taking $\mu = E[x]$.

We have,

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_m \end{bmatrix}$$

$$\Sigma = X^T X$$

$$= \begin{bmatrix} 1 & 1 & . & . & 1 \\ x_1 & x_2 & . & . & x_m \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_m \end{bmatrix}$$

$$= \begin{bmatrix} 1+1+..+1 & x_1+x_2+..+x_m \\ x_1+x_2+..+x_m & x_1^2+x_2^2+..+x_m^2 \end{bmatrix}$$

$$= m \begin{bmatrix} 1 & E[x] \\ E[x] & E[x^2] \end{bmatrix}$$

Eigenvalues for a matrix A = $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is given by:

$$\lambda_1 = \frac{(a+d) - \sqrt{(a+d)^2 - 4(ad-bc)}}{2}$$

$$\lambda_2 = \frac{(a+d) + \sqrt{(a+d)^2 - 4(ad-bc)}}{2}$$

Now, calculating the eigenvalues for our matrix σ:

$$\lambda_1 = \frac{(1+E[x^2]) - \sqrt{(1+E[x^2])^2 - 4(E[x^2]-E[x]^2)}}{2}$$

$$= \frac{(1+E[x^2]) - \sqrt{(1+E[x^2])^2 - 4Var[x]}}{2}$$

$$= \frac{(1+E[x^2])}{2}(1 - \sqrt{1 - \frac{4Var[x]}{(1+E[x^2)^2}})$$

Similarly,

$$\lambda_2 = \frac{(1+E[x^2])}{2}(1 + \sqrt{1 - \frac{4Var[x]}{(1+E[x^2)^2}})$$

Now since we have eigenvalues, we can calculate the condition number for the re cen-

tered data as,

$$
\begin{aligned}
k(\Sigma\prime) &= \frac{\lambda_{max}}{\lambda_{min}} \\
&= \frac{1+\sqrt{1-\frac{4Var[x\prime]}{(1+E[x\prime^2)]^2}}}{1-\sqrt{1-\frac{4Var[x\prime]}{(1+E[x\prime^2)]^2}}} \\
&= \frac{1+\sqrt{1-\frac{4Var[x-\mu]}{(1+E[(x-\mu)^2]^2}}}{1-\sqrt{1-\frac{4Var[x-\mu]}{(1+E[(x-\mu)^2]^2}}} \\
&= \frac{1+\sqrt{1-\frac{4Var[x]}{(1+E[(x-\mu)^2]^2}}}{1-\sqrt{1-\frac{4Var[x]}{(1+E[(x-\mu)^2]^2}}}
\end{aligned}
$$

Taking $\mu = E[x]$,

$$
\begin{aligned}
&= \frac{1+\sqrt{1-\frac{4Var[x]}{(1+E[(x-E[x])^2]^2}}}{1-\sqrt{1-\frac{4Var[x]}{(1+E[(x-E[x])^2]^2}}} \\
&= \frac{1+\sqrt{1-\frac{4Var[x]}{(1+Var[x])^2}}}{1-\sqrt{1-\frac{4Var[x]}{(1+Var[x])^2}}} \\
&= \frac{1+\sqrt{1-\frac{4Var[x]}{(1+Var[x])^2}}}{1-\sqrt{1-\frac{4Var[x]}{(1+Var[x])^2}}} \\
&= \frac{1}{Var[x]}
\end{aligned}
$$