Regression Assignment

by Sagar Jain

Question 1

Consider regression in one dimension, with a data set {(xi, yi)}i=1,...,m. Find a linear model that minimizes the training error, i.e.,

$$\dot{w} \quad \text{and} \quad \dot{b}$$

to minimize

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

We can write the above equation as

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

Lets expand this expression, we get:

$$SquaredError_{\cdot}line = (y_1 - (\dot{w}x_1 + \dot{b}))^2 + (y_2 - (\dot{w}x_2 + \dot{b}))^2 + (y_3 - (\dot{w}x_3 + \dot{b}))^2 + \ldots\ldots + (y_m$$

$$\begin{aligned} = \quad & y_1^2 - 2y_1(\dot{w}x_1 + \dot{b}) + (\dot{w}x_1 + \dot{b})^2 \\ & + y_2^2 - 2y_2(\dot{w}x_2 + \dot{b}) + (\dot{w}x_2 + \dot{b})^2 \\ & . \\ & . \\ & . \\ & . \\ & + y_m^2 - 2y_m(\dot{w}x_m + \dot{b}) + (\dot{w}x_m + \dot{b})^2 \end{aligned}$$

$$\begin{aligned} = \quad & y_1^2 - 2y_1\dot{w}x_1 - 2y_1\dot{b} + \dot{w}^2x_1^2 + 2\dot{w}x_1\dot{b} + \dot{b}^2 \\ & + y_2^2 - 2y_2\dot{w}x_2 - 2y_2\dot{b} + \dot{w}^2x_2^2 + 2\dot{w}x_2\dot{b} + \dot{b}^2 \\ & . \\ & . \\ & . \\ & . \\ & + y_m^2 - 2y_m\dot{w}x_m - 2y_m\dot{b} + \dot{w}^2x_m^2 + 2\dot{w}x_m\dot{b} + \dot{b}^2 \end{aligned}$$

$$\begin{aligned} = \quad & (y_1^2 + y_2^2 + y_3^2 + \ldots\ldots + y_m^2) - 2w(y_1x_1 + y_2x_2 + \ldots\ldots\ldots + y_mx_m) - 2b(y_1 + y_2 + \ldots + y_m) \\ & + mb^2 \end{aligned}$$

$$= \quad m\bar{y^2} - 2wm\bar{xy} - 2bm\bar{y} + w^2m\bar{x^2} + 2wbm\bar{x} + mb^2$$

Now inorder to minimise the

$$\dot{w} \quad and \quad \dot{b}$$

we need to differentiate the above equation w.r.t to w and b and equate them to zero. that is

$$\frac{\partial SE}{\partial w} = 0 \quad and \quad \frac{\partial SE}{\partial b} = 0$$

Differentiating the squarederror equation, we get

$$-2m\bar{x}\bar{y} + 2m\bar{x^2}w + 2bm\bar{x} = 0 \quad and \quad -2m\bar{y} + 2mw\bar{x} + 2bm = 0$$
$$= -\bar{x}\bar{y} + w\bar{x^2} + b\bar{x} = 0 \quad and \quad -\bar{y} + w\bar{x} + b = 0$$
$$w\bar{x^2} + b\bar{x} = \bar{x}\bar{y} \qquad \dots\dots\dots\dots(1)$$
$$and \quad w\bar{x} + b = \bar{y} \qquad \dots\dots\dots\dots(2)$$

By looking at the above equation, we can conclude that point

$$(\bar{x}, \bar{y}) \quad and \quad (\frac{\bar{x}^2}{\bar{x}}, \frac{\bar{x}\bar{y}}{\bar{x}}) \quad lies \quad on \quad the \quad best \quad line \quad y = wx + b$$

Solving equation 1 and 2, we get the values of w and b which will minimise the mean squared error.

$$\hat{w} = \frac{\bar{x}\bar{y} - \bar{x}\bar{y}}{\bar{x}^2 - \bar{x^2}} \quad and \quad \hat{b} = y - (\frac{\bar{x}\bar{y} - \bar{x}\bar{y}}{\bar{x}^2 - \bar{x^2}})\bar{x}$$

Question 2

Assume there is some true linear model, such that

$$y_i = wx_i + b + \epsilon$$

, where noise variables E are i.i.d.

with

$$\epsilon \sim N(0, \sigma^2)$$

. Argue that the estimators are unbiased, i.e., \mathbb{E}[\hat{w}]= $wand$\mathbb{E}[\hat{b}]$=b

What are the variances of these estimators?

Now $\hat{w}$ can be written as $\frac{c_{XY}}{s_x^2}$ where $s_x^2 = \bar{x}^2 - \bar{x^2}$ and $c_{XY} = \bar{x}\bar{y} - \bar{x}\bar{y}$

We'll start with the slope, $\hat{w}$

$$\hat{w} = \frac{c_{XY}}{s_x^2}$$
$$= \frac{\frac{1}{m}\sum_{i=1}^m x_i(y_i - \bar{x}\bar{y}}{s_x^2}$$
$$= \frac{\frac{1}{m}\sum_{i=1}^m x_i(b + w_i x_i + \epsilon_i) - \bar{x}(b + w_i\bar{x} + \bar{\epsilon})}{s_x^2}$$
$$= \frac{b\bar{x} + w\bar{x^2} + \frac{1}{m}\sum_{i=1}^m x_i\epsilon_i - \bar{x}b - w\bar{x^2} - \bar{x}\bar{\epsilon}}{s_x^2}$$
$$= \frac{ws_x^2 + \frac{1}{m}\sum_{i=1}^m x_i\epsilon_i - \bar{x}\bar{\epsilon}}{s_x^2}$$
$$= w + \frac{\frac{1}{m}\sum_{i=1}^m x_i\epsilon_i - \bar{x}b - w\bar{x^2} - \bar{x}\bar{\epsilon}}{s_x^2}$$

since $\bar{x}\bar{\epsilon} = n^{-1}\sum_i \bar{x}\epsilon_i$

$$\hat{w} = w + \frac{\frac{1}{m}\sum_{i=1}^m x_i\epsilon_i - \bar{x}\bar{\epsilon}}{s_x^2}$$

This representation of the slope estimate shows that it is equal to the trueb slope (w) plus something which depends on the noise terms (the ϵ_i, and their sample average $\bar{\epsilon}$).

Expected value and bias:

Recall that $\mathbb{E}[\epsilon_i|X_i]$=0 , so

$$\frac{1}{m}\sum_{i=1}^m (x_i - \bar{x})\mathbb{E}[\epsilon_i] = 0$$

Thus, $\mathbb{E}[\hat{w}] = w$

Since the bias of an estimator is the difference between its expected value and the truth, $\hat{w}$ is an unbiased estimator of the optimal slope.

Turning to the intercept,

$$\mathbb{E}[\hat{b}] = \mathbb{E}[\bar{Y} - \hat{w}\bar{X}]$$
$$= b + w\bar{X} - \mathbb{E}[\hat{w}]\bar{X}$$
$$= b + w\bar{X} - w\bar{X}$$
$$= b$$

so it is also unbiased

Variance and Standard Error

$$\mathbb{Var}[\hat{w}] = \mathbb{Var}[w + \frac{\frac{1}{m}\sum_{i=1}^{m} x_i\epsilon_i - \bar{x}\bar{\epsilon}}{s_x^2}]$$
$$= \mathbb{Var}[\frac{\frac{1}{m}\sum_{i=1}^{m} x_i\epsilon_i - \bar{x}\bar{\epsilon}}{s_x^2}]$$
$$= \frac{\frac{1}{n^2}\sum_{i=1}^{m}(x_i - \bar{x})^2\mathbb{Var}[\epsilon_i]}{(s_x^2)^2}$$
$$= \frac{\frac{\sigma^2}{m}s_x^2}{(s_x^2)^2}$$
$$= \frac{\sigma^2}{ms_x^2}$$

Hence,

$$\mathbb{Var}[\hat{w}] \quad is \quad approximately \quad equal \quad to \quad \frac{\sigma^2}{m\mathbb{Var}(x)}$$

In words, this says that the variance of the slope estimate goes up as the noise around the regression line σ^2 gets bigger, and goes down as we have more observations (m), which are further spread out along the horizontal axis s_x^2; it should not be surprising that it's easier to work out the slope of a line from many, well-separated points on the line than from a few points smushed together

Similarly for calculating the variance for $\hat{b}$

$$\mathbb{Var}[\hat{b}] = \mathbb{Var}[\bar{y}] + \bar{x}^2\mathbb{Var}[\hat{w}] - 2\bar{x}\mathbb{Cov}(\bar{y}, w) \qquad \ldots\ldots eq(3)$$

On calculating $\mathbb{Cov}(\bar{y}, w)$ we get 0 value.

Putting all the values in the above equation, we get

$$\mathbb{Var}[\hat{b}] = \sigma^2(\frac{1}{m} + \frac{\bar{x}^2}{(s_x^2)^2})$$

Hence,

$$\mathbb{Var}[\hat{b}] \quad is \quad approximately \quad equal \quad to \quad \frac{\sigma^2\mathbb{E}[x^2]}{m\mathbb{Var}(x)}$$

Question4

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

We have calculated the variances of both estimators and we can observe that

Variance of w is independent of the shift of data from one place to the origin as

$$\mathbb{Var}[\hat{w}] = \frac{\sigma^2}{m\mathbb{Var}(x)}$$

in above equation Var(x) will not get changed even if we will shift the data from one point to another point. Therefore, we can conclude that doing regression on the re-centered data produces the same error on $\hat{w}$.

While in the case of variance of b, which is

$$\mathbb{V}\mathrm{ar}[\hat{b}] = \frac{\sigma^2 \mathbb{E}[x^2]}{m\mathbb{V}\mathrm{ar}(x)}$$

depends on the $\mathbb{E}[x^2]$

Therefore, we can conclude that doing regression on the re-centered data produces the minimise error on $\hat{b}$.

Additional Observation

Also we can observe while calculating the values of w and b that,

$$\hat{w} = \frac{\bar{x}\bar{y} - \bar{xy}}{\bar{x}^2 - \bar{x^2}} \quad and \quad \hat{b} = y - (\frac{\bar{x}\bar{y} - \bar{xy}}{\bar{x}^2 - \bar{x^2}})\bar{x}$$

So when we look at the eqn. for w we see that the absolute value of x had less dependence on w, so if we shift the value of x the difference in the numerator and denominator will change with same amount and when divided will provide the orignal fraction.

Where as in the case of b, we see that its directly proportional to $-x$. Hence any shift in x will directly affect the value of b. Also it depends on the value of $\bar{x}$. And the value of $\bar{x}$ is minimum when the values of x are taken around the mean.

Question 5

Verify this numerically in the following way: Taking $m = 200, w = 1, b = 5, \sigma^2 = 0.1$

- Repeatedly perform the following numerical experiment: generate $x_1, \ldots, x_m$ ~ $Unif(100, 102)$, $y_i = wx_i + b + \epsilon_i$ (with ϵ_i as a normal, mean 0, variance σ^2), and $x_i' = x_i - 101$, compute $\hat{w}, \hat{b}$ based on the $\{(x_i, y_i)\}$ data, and $\hat{w}', \hat{b}'$ based on the $\{(x_i', y_i)\}$ data.
- Do this 1000 times, and estimate the expected value and variance of $\hat{w}, \hat{w}', \hat{b}, \hat{b}'$. Do these results make sense? Do these results agree with the above limiting expressions?

In [3]:

```
#importing libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as st
```

In [4]:

```
#initializing global variables
m = 200
w = 1
b = 5

mean = 0
sigma_sq = 0.1

#function to create dataset
def create_dataset():
    epsilon = np.random.normal(mean, np.sqrt(sigma_sq), 1)

    x = np.random.uniform(low=100, high=102, size=m)
    y = (w * x) + b + epsilon

    x_dash = x - 101

    return x, y, x_dash
```

In [5]:

```
def compute_coefficients(x, y):

    x_bar = np.mean(x)
    y_bar = np.mean(y)

    x_bar_y_bar = x_bar * y_bar

    xy = x * y
    xy_bar = np.mean(xy)

    x_sq = x ** 2
    x_sq_bar = np.mean(x_sq)

    x_bar_sq = x_bar ** 2

    w_hat = (x_bar_y_bar - xy_bar) / (x_bar_sq - x_sq_bar)

    b_hat = y_bar - (w_hat * x_bar)

    return w_hat, b_hat
```

In [6]:

```
#simulate the expriment 1000 times
simulation_count = 1000

w_hat = []
w_hat_dash = []

b_hat = []
b_hat_dash = []

for count in range(0, simulation_count):
    x, y, x_dash = create_dataset()

    w, b = compute_coefficients(x, y)
    w_hat.append(w)
    b_hat.append(b)

    w_dash, b_dash = compute_coefficients(x_dash, y)
    w_hat_dash.append(w_dash)
    b_hat_dash.append(b_dash)
```

In [7]:

```
#Estimate the expected value and variance of w_hat
print('Expected value of W_hat: ',np.mean(w_hat))
print('Variance of W_hat: ',np.var(w_hat))
```

```
Expected value of W_hat:  0.999999999908474
Variance of W_hat:  5.400003811314988e-21
```

In [8]:

```
#Estimate the expected value and variance of w_hat_dash
print('Expected value of W_hat_dash: ',np.mean(w_hat_dash))
print('Variance of W_hat_dash: ',np.var(w_hat_dash))
```

```
Expected value of W_hat_dash:  0.9999999999086403
Variance of W_hat_dash:  5.402766126968039e-21
```

In [10]:

```
#Estimate the expected value and variance of b_hat
print('Expected value of b_hat: ',np.mean(b_hat))
print('Variance of b_hat: ',np.var(b_hat))
```

```
Expected value of b_hat:  5.025041398157789
Variance of b_hat:  5.223345644933421
```

In [11]:

```
#Estimate the expected value and variance of b_hat_dash
print('Expected value of b_hat_dash: ',np.mean(b_hat_dash))
print('Variance of b_hat_dash: ',np.var(b_hat_dash))
```

```
Expected value of b_hat_dash:  106.02504138891366
Variance of b_hat_dash:  5.22334566551959
```

Inference

Yes, the results make sense. The value of x is shifted to x' , as expected the value of intercept is also shifted by the same amount, as seen in above simulation. It does agree with the above limiting equation in the previous question.

Question 6

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

Answer

The linear model $y = wx + b$ gives a straight line which try to minimize the mean squared error (MSE) of distance between point and the line. When the data is shifted on the x-axis, its coordinates for y-axis remains the same. As the data points have not change their relative position with each other, a similar shifted line will provide a line which minimizes the MSE of the data. This shifted line thus has the same slope but a shifted intercept parameter.

Question 7

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

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

Show that re-centering the data ($\Sigma = (X')^T(X')$, taking $x_i' = x_i - \mu$), the condition number $\kappa(\Sigma')$ is minimized taking $\mu = \mathbb{E}[x]$.

Answer

When we transform the data from 1-D to 2-D we consider the matrix of data X. So in this case taking the square of the initial matrix X, we get $\Sigma = X^T X$.

So when we compute individual value of x of the Σ matrix such as $x_{1,1}, x_{1,2} \dots, x_{m,m}$

For the diagonal values we get

$$\begin{aligned} \frac{1}{m}\Sigma_{1,1} &= \frac{1}{m}\sum_{i=0}^{m} X_1^i X_1^i = \mathbb{E}[X_1, X_1] = \mathbb{E}[X_1^2] \\ \frac{1}{m}\Sigma_{1,2} &= \frac{1}{m}\sum_{i=0}^{m} X_1^i X_2^i = \mathbb{E}[X_1, X_2] \end{aligned}$$

So on computing for all values from 0 to m then we get Σ in the form,

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

Recentering the data is usually done for the purpose of preconditioning the data.

We considered the data points x1 = (100, 50), x2 = (100, 52), x3 = (101, 51). We calculated the largest eigenvalue of $X^T X$ which comes out to be 38, 000, and the smallest comes out to be 1.6, giving a condition number of κ(Σ) ≈ 22, 000.

In this case, note that the average data point $\bar{x}$ is (100.333, 51) If we will ‘center’ the data by subtracting off this mean, we get x1 = (−1/3, −1), x2 = (−1/3, 1), x3 = (2/3, 0).

Building the data matrix out of this re-centered data we get that $(X^{`})^T X^{`}$ has eigenvalues of 2 and 2/3, with a total condition number of κ(Σ0) = 3.

This represents a massive improvement in the relative error in various directions, with nothing more complicated than re-centering the data to have mean 0. Re-centering like this ensures that the principal components of the data really capture the fundamental geometry of the data, rather than simply where the cloud is sitting in space.