hw2

February 15, 2023

1 Q1: Verify OLS

Apply your OLS to each of the datasets in the google drive here: `https://drive.google.com/drive/u/0/folders/1oIcj6jpwUwHU3oK0zJS0Aywup1cGelUD`. Compare with the $\hat{a}, \hat{b}$ in the corresponding model file. Do not upload your results, but let us know if there are significant differences.

2 Q2: Create your synthetic datasets

Follow the procedure below:

1. $a = 2, b = -3$
2. FOR dataset index $i = 1 \dots 100$
3. FOR $j = 1 \dots n$
4. draw $x_j^{(i)} \sim unif[-1, 1]$
5. draw $\epsilon_j^{(i)} \sim N(0, \sigma^2)$
6. $y_j^{(i)} = ax_j^{(i)} + b + \epsilon_j^{(i)}$
7. END FOR j
8. Collect these n noisy training points into the ith dataset $D^{(i)} = (x_1^{(i)}, y_1^{(i)}) \dots (x_n^{(i)}, y_n^{(i)})$
9. END FOR i

For this question, please use $n = 10$, $\sigma^2 = 1$. That is, you will be creating 100 separate (but related by the same underlying a, b) datasets, each has 10 data points. The noise is moderate. Save these datasets for yourself.

Note: in reality, you will see only one training set for machine learning problems. What we are simulating here is to give you a peek into 100 "alternative universes", each universe has one randomized dataset of size n. We will see how the model learned in each universe relate next. Keep in mind, though, in practice you will not have such clairvoyant vision.

3 Q3: Run OLS on these 100 datasets

Run your OLS on each dataset. That is, you will "learn" parameters $\hat{a}^{(i)}, \hat{b}^{(i)} = OLS(D^{(i)})$ for $i = 1 \dots 100$. The distribution of these 100 pairs of parameters gives us a good sense of how one typical training set of size n will behave. To visualize such distribution, produce two plots:

1. In the first plot show 100 lines, one for each $\hat{a}^{(i)}x + \hat{b}^{(i)}$. On top of them, also show the ground truth line $ax + b$ (make sure you can distinguish this line from the other 100 lines). You do not need to show any data points.

2. In the second plot, show each parameter pair $(\hat{a}^{(i)}, \hat{b}^{(i)})$ as a point in the 2D parameter space. This will produce a point cloud with 100 points. Also show the true parameter (a, b) as another point.

4 Q4: Change the dataset size n

Repeat Q2 and Q3, but with $n = 100$. This simulates "big data". You will generate two more plots.

Repeat Q2 and Q3 again, but this time with $n = 2$. This simulates "not enough training data". You will generate two more plots.

For best visual effect, keep the same axis range for the same type of plots.

Think why n affects the distribution of $\hat{a}, \hat{b}$.

5 Q5: Change the noise level

Reset $n = 10$. Repeat Q2 and Q3, this time with $\sigma^2 = 0.01$. This simulates low noise level. You will generate two more plots.

Repeat Q2 and Q3 but increase noise to $\sigma^2 = 100$. This simulates low noise level. You will generate two more plots.

6 Hand in

Create a subdirectory with your name in google drive `https://drive.google.com/drive/u/0/folders/1aKP_fw2RxeM-XtSOYRaba2DmNc44nGh-`. Upload 2 plots for Q3, 4 plots for Q4, 4 plots for Q5 to your subdirectory.