ML Homework1

20160632 조장현

In this homework we are given a set of data composed of x and y values, which have relationship, y = **BTx** + B0 + E where E is noise. Here we can change the relationship for convenience. **B’** = [**B** B0] and **x’** = [**x** 1], so that it is of the form y = **B’Tx’** + E. In the following problems, we are to divide data set into two sets of 80% and 20%. Then get **B** and B0 through linear regression so that sum of all E’s are minimized on 80% set. Next, these **B** and B0 are applied on 20% set to get the average value of absolute value of E. So the error is calculated as

1.

|  |  |
| --- | --- |
| B0 != 0 | B0 = 0 |
| 3.755274 | 4.133775 |
| 3.41951 | 3.784075 |
| 3.875857 | 4.093042 |
| 3.587554 | 3.685145 |
| 3.584395 | 3.951871 |
| 3.605719 | 3.940512 |
| 3.74243 | 4.129507 |
| 3.669833 | 3.946408 |
| 3.578165 | 3.927182 |
| 3.514317 | 3.724565 |
| Average | Average |
| 3.633305 | 3.931608 |

Result is like the left. When B0 is not 0, average of errors is 3.63, whereas then B0 is 0 average of errors is 3.93. It is easily seen that error is smaller when B0 is not 0

For fixed B, the B0 should be average of error of training data without B0. This B0 work as a value to normalize all errors, so that their sum decrease. B0 moves the line of Y = BTX along y-axis so that it gets more optimized.

2.

|  |  |  |
| --- | --- | --- |
| large | proper | small |
| 3.0648E+189 | 3.733422 | 18.14872 |
| 3.2432E+189 | 3.413443 | 18.92868 |
| 2.0876E+189 | 3.915002 | 18.56479 |
| 2.5776E+189 | 3.666496 | 19.41636 |
| 5.48E+188 | 3.533409 | 17.25178 |
| 3.9585E+190 | 3.695467 | 18.67223 |
| 3.4573E+188 | 3.824075 | 18.90539 |
| 3.0335E+189 | 3.749033 | 19.77229 |
| 2.8577E+188 | 3.461837 | 19.14774 |
| 6.1898E+189 | 3.620974 | 18.278 |
| Average | Average | Average |
| 6.0961E+189 | 3.661316 | 18.7086 |

Left is the chart of error when step size is large, proper and small. Below there are charts of error then step size is large, proper and small. We can see that when step size is big, error becomes larger and larger, giving huge error. When it is proper, it gives error of 3.66, which is around prob 1 (when B0 != 0). In addition it will get smaller as number of iterations increase. Last when step size is small, error does get decreased but it’s too slow. But it will eventually get to optimal value.

3.

4.

Thinking of a N x N diagonal matrix Rwhich has as it’s value at xii, our formula **Y = BTX + E** changes to **Y’ = B’TX’ + E’. Y’ = RY, X’ = RX**. Then we can re-calculate **B** by **(X’TX’)-1X’TY’.** If we change **X’** to **RX** and **Y’** to **RY**, it becomes **(XTR2X)-1XTR2Y**. **R**2 is still a diagonal matrix with values ri as it’s value at xii.

For alternative interpretation, we can think of rns coming from other sides: data dependent noise or replication of data points. It noise is dependent on data so that noise is amplified by rn, it has same effect as having weighting factor. Also, if data point which give error of (yi – **B**T**x**i)2 take place rn times, it’s just the same as having weight of rn.

5.

To use the point that a continuous function that is midpoint convex is convex. I’ll show continuity and midpoint convexity.

1. Continuity

Since exp function and log functions are all continuous functions, log sum exp function is continuous.

1. Midpoint convex

By simplifying to two points, (x1, x2) and (x1’ and x2’).

By getting two to LHS and remove the log function, it becomes If we solve this and remove equal terms from each side, rest becomes It is the form same as so midpoint convexity is proven.

6.

Lagrangian of given problem is = **c**T**x** + λ(**Ax**-B)

It has to fit the condition where and

By solving this, **c**T=0 and **c**T+λ**A** = 0

g(λ) =