ML Homework1

20160632 조장현

In this homework we are given a set of data composed of x and y values, which have relationship, y = $\mathbf{B}^T\mathbf{x}$ + B_0 + E where E is noise. Here we can change the relationship for convenience. $\mathbf{B'}$ = [**B** B_0] and $\mathbf{x'}$ = [**x** 1], so that it is of the form y = $\mathbf{B'}^T\mathbf{x'}$ + E. In the following problems, we are to divide data set into two sets of 80% and 20%. Then get **B** and B_0 through linear regression so that sum of all E's are minimized on 80% set. Next, these **B** and B_0 are applied on 20% set to get the average value of absolute value of E. So the error is calculated as $\frac{1}{N}\sum_{i=1}^{N}|y_{data} - y_{prediction}|$

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

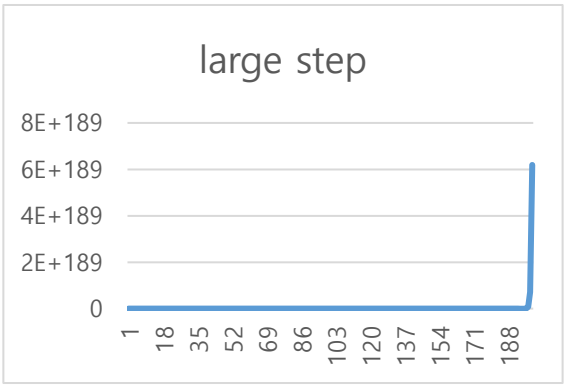
Result is like the left. When B_0 is not 0, average of errors is 3.63, whereas then B_0 is 0 average of errors is 3.93. It is easily seen that error is smaller when B_0 is not 0

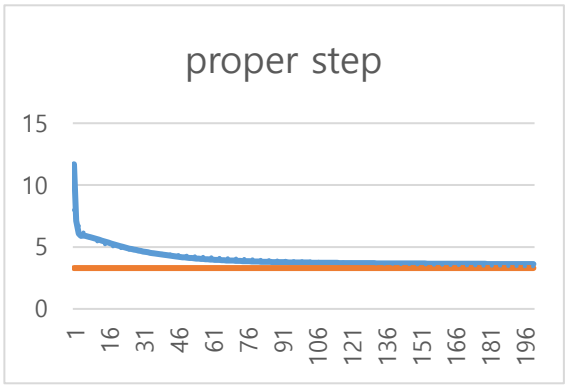
For fixed B, the B_0 should be average of error of training data without B_0. This B_0 work as a value to normalize all errors, so that their sum decrease. B_0 moves the line of $Y = B^TX$ along y-axis so that it gets more optimized.

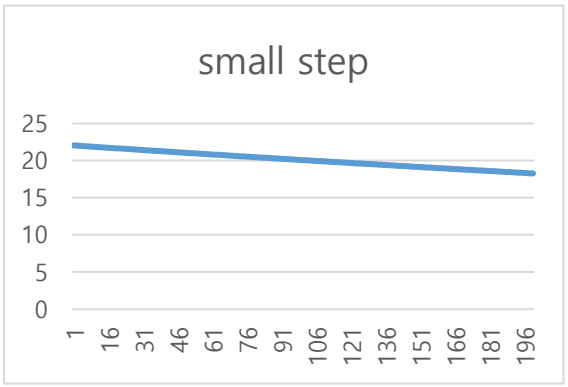
2.

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 B_0 != 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 R which has $\sqrt{r_i}$ as it's value at x_{ii}, our formula **Y** = $\mathbf{B}^T\mathbf{X}$ + **E** changes to **Y'** = $\mathbf{B'}^T\mathbf{X'}$ + **E'**. **Y'** = **RY**, **X'** = **RX**. Then we can re-calculate **B** by $\mathbf{(X'^TX')^{-1}X'^TY'}$. If we change **X'** to **RX** and **Y'** to **RY**, it becomes $\mathbf{(X^TR^2X)^{-1}X^TR^2Y}$. $\mathbf{R}^2$ is still a diagonal matrix with values r_i as it's value at x_{ii}.

For alternative interpretation, we can think of r_ns coming from other sides: data dependent noise or replication of data points. It noise is dependent on data so that noise is amplified by r_n, it has same effect as having weighting factor. Also, if data point which give error of $(y_i - \mathbf{B}^T\mathbf{x}_i)^2$ take place r_n times, it's just the same as having weight of r_n.

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.

2) Midpoint convex

   By simplifying to two points, (x_1, x_2) and (x_1' and x_2'). $\mathrm{F}\left(\frac{x_1+x_1'}{2}, \frac{x_2+x_2'}{2}\right) \leq \frac{F(x_1,x_2)+F(x_1',x_2')}{2}$

   By getting two to LHS and remove the log function, it becomes $(e^{\frac{x_1+x_1'}{2}} + e^{\frac{x_2+x_2'}{2}})^2 \leq (e^{x_1} + e^{x_2})(e^{x_1'} + e^{x_2'})$ If we solve this and remove equal terms from each side, rest becomes $2e^{\frac{x_1+x_1'+x_2+x_2'}{2}} \leq e^{x_1+x_2'} + e^{x_1'+x_2}$ It is the form same as $\mathrm{A} + \mathrm{B} \geq 2\sqrt{AB}$ so midpoint convexity is proven.

6.

Lagrangian of given problem is $\mathrm{L}(\mathbf{x}, \lambda, \mathrm{v}) = \mathbf{c}^T\mathbf{x} + \lambda(\mathbf{A}\mathbf{x}\text{-}\mathrm{B})$

It has to fit the condition where $\nabla f_0(\boldsymbol{x}^*) = 0$ and $\nabla f_0(\boldsymbol{x}^*) + \lambda\nabla f_1(\boldsymbol{x}^*) = 0$

By solving this, $\mathbf{c}^T=0$ and $\mathbf{c}^T+\lambda\mathbf{A} = 0$

$g(\lambda) = \inf_x\{\boldsymbol{c}^T\boldsymbol{x} + \lambda(\mathbf{A}\mathbf{x} - \mathrm{B})\}$