Math 156 Term Project 1

Deadline: Saturday April 18th 11:59pm

Problem 1

Explain in your own words the benefit of maximizing the log likelihood compared to directly maximizing the likelihood function. You will have to write a few complete (English) sentences along with some supplementary mathematical formulas to answer this question. (20 points)

Problem 2

(a) Give an example of continuous data, i.e. a data set for which we can use a continuous random variable to model. (5 points)
(b) Give an example of discrete data, i.e. a data set for which we need to use a discrete random variable. (5 points)
(c) Prove that $\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$ for both continuous and discrete random variables. Recall the definition of variance of $f(x)$ is defined by

$$\text{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2].$$

(Feel free to use measure theory to give a unifying proof for both.) (10 points)

Problem 3

Suppose we have a data set $\mathcal{D} = (\mathbf{x}, \mathbf{t}), \mathbf{x} = (x_1, \cdots, x_N), \mathbf{t} = (t_1, \cdots, t_N)$ generated by

$$t_i = f(x_i) + \text{noise},$$

where $f : \mathbb{R} \to \mathbb{R}$ is an unknown function of interest. N is the number of data points given. We can assume the noise is some Gaussian noise with mean zero, i.e. each noise is independent of each other and $\sim \mathcal{N}(0, \beta^{-1})$ for some unknown parameter $\beta > 0$. Figure 1 shows the data provided with $N = 20$. We would like to use this data set to approximate f using a function of the form

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j,$$

where $M \geq 2$ (degree of the polynomial) and $\mathbf{w} = (w_0, w_1, \cdots, w_M) \in \mathbb{R}^{M+1}$ (polynomial coefficients) are parameters to be determined by you. Use everything we taught in class (and reference to textbook §1.1 and §1.2.5) to finish the following tasks.

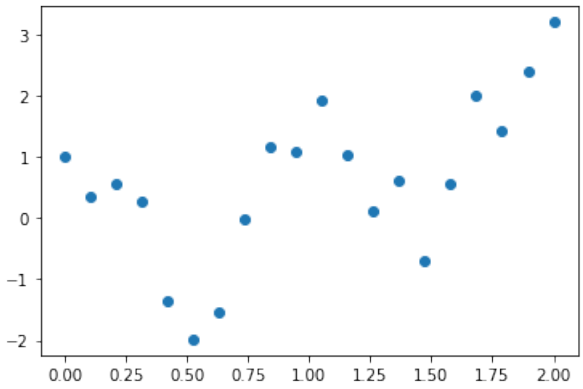

Figure 1: Scatter plot of `fitting_N20.csv`.

(a) Formulate the problem into a minimization or maximization problem. What is your cost function? (Feel free to add regularizer if you deem suitable.) Your answer to this question should be handwritten math equations or typed in LaTeX. (Hint: they are provided on textbook p.5, p.29, or p.30.) (10 points)

(b) Read the data into your program from the csv file `fitting_N20.csv` and create a plot similar to Figure 1. (5 points)

(c) Solve the problem numerically, that is, write a program that takes the data in `fitting_N20.csv` and compute the corresponding optimal coefficients $\mathbf{w}^* = (w_0^*, w_1^*, w_2^*, \cdots, w_M^*)$. Write down your plan for solving this problem in math equations or pseudocode (or type in LaTeX) for partial credit. (25 points)

(d) Plot your fitting function $y(x, \mathbf{w})$ as a curve alongside the data. The following graph in Figure 2 is an example for you. (10 points)

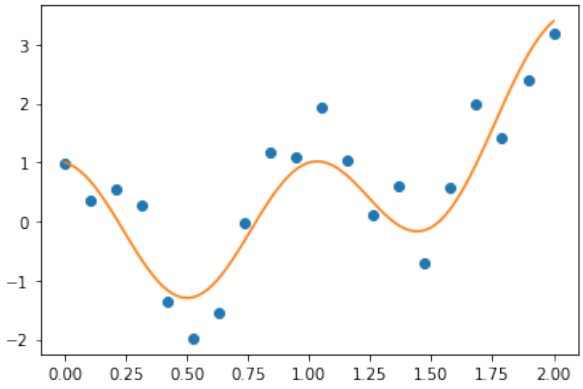

Figure 2: Sample plot for task (d).

(e) What is your choice of M? Why? You can performed the above tasks with a few different choices of M, or give a convincing heuristic argument. (20 points)

(f) **(Bonus)** You can find `fitting_N50.csv` on the course website. It is generated with the same process as `fitting_N20.csv`, except with more data point in the same interval $[0, 2]$. See Figure 3. Note that this data set is obtained with the same kind of random noise ($\sim \mathcal{N}(0, \beta^{-1})$). Use this however you wish to generate more information and explain how a larger data set helps with your work. (up to 20 points)

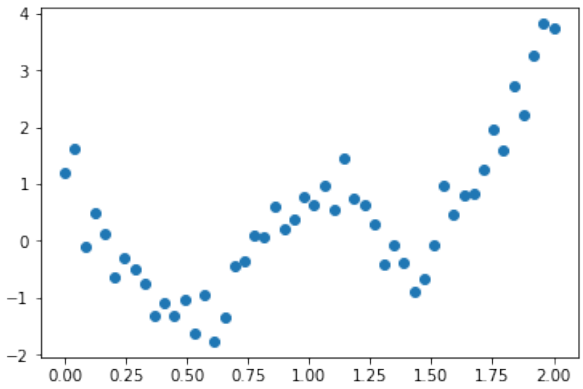

Figure 3: Scatter plot for `fitting_N50.csv`.

Important submission note: You can use any programming language to accomplish the coding tasks. The university now provides MATLAB access to all UCLA students, see https://www.it.ucla.edu/news/matlab-software-now-available. Other high-level programming languages suitable for our tasks include Python, R, Julia, Octave, and more. The instructor and the TA guarantee assistance with MATLAB programming and can potentially help you trouble shoot your code in other programming languages (not guaranteed). If you wish to email us asking for programming assistance, please attach your code, highlight the part in question, and articulate your problems. Please attach your code at the end of your submission and document a list of your collaborators and external resource. This document is created with LaTeX and you can find the source .tex file on CCLE. For more LaTeX help, check out this LaTeX tutorial link, or take a look at my template .tex file.

The following table is the data provided in `fitting_N20.csv` for your visual inspection.

i	x_i	t_i
1	0.000000	0.991459
2	0.105263	0.360328
3	0.210526	0.558448
4	0.315789	0.265560
5	0.421053	-1.364200
6	0.526316	-1.983883
7	0.631579	-1.551820
8	0.736842	-0.020161
9	0.842105	1.164831
10	0.947368	1.090539
11	1.052632	1.925967
12	1.157895	1.031809
13	1.263158	0.099923
14	1.368421	0.608555
15	1.473684	-0.701440
16	1.578947	0.566558
17	1.684211	1.998774
18	1.789474	1.423031
19	1.894737	2.386509
20	2.000000	3.199598