ECE-GY 6143 / Spring 2020

Homework 1

Please scan and upload your assignments on or before February 6, 2020.

- You are encouraged to discuss ideas with each other; but
- you **must acknowledge** your collaborator, and
- you **must compose your own** writeup and/or code independently.
- We **strongly** encourage answers to theory questions in Latex, and answers to coding questions in Python (Jupyter notebooks).
- Maximum score: 50 points.

---

1. **(10 points)** Let $\{x_1, x_2, \ldots, x_n\}$ be a set of points in d-dimensional space. Suppose we wish to produce a single point estimate $\mu \in \mathbb{R}^d$ that minimizes the squared-error:

$$\|x_1 - \mu\|_2^2 + \|x_2 - \mu\|_2^2 + \ldots + \|x_n - \mu\|_2^2$$

   Find a closed form expression for μ and prove that your answer is correct.

2. **(10 points)** Not all norms behave the same; for instance, the ℓ_1-norm of a vector can be dramatically different from the ℓ_2-norm, especially in high dimensions. Prove the following norm inequalities for d-dimensional vectors, starting from the definitions provided in class and lecture notes. (Use any algebraic technique/result you like, as long as you cite it.)

   a. $\|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$

   b. $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\|x\|_\infty$

   c. $\|x\|_\infty \leq \|x\|_1 \leq d\|x\|_1$

3. **(10 points)** When we think of a Gaussian distribution (a bell-curve) in 1, 2, or 3 dimensions, the picture that comes to mind is a "blob" with a lot of mass near the origin and exponential decay away from the origin. However, the picture is very different in higher dimensions (and illustrates the counter-intuitive nature of high-dimensional data analysis). In short, we will show that *Gaussian distributions are like soap bubbles*: most of the mass is concentrated near a shell of a given radius, and is empty everywhere else.

   a. Fix $d = 3$ and generate 10,000 random samples from the standard multi-variate Gaussian distribution defined in $\mathbb{R}^d$.

   b. Compute and plot the histogram of Euclidean norms of your samples. Also calculate the average and standard deviation of the norms.

   c. Increase d on a coarsely spaced log scale all the way up to $d = 1000$ (say $d = 50, 100, 200, 500, 1000$), and repeat parts (a) and (b). Plot the variation of the average and the standard deviation of Euclidean norm of the samples with increasing d.

   d. What can you conclude from your plot from part (c)?

   e. **Bonus, not for grade.** Mathematically justify your conclusion using a formal proof. You are free to use any familiar laws of probability, algebra, or geometry.

4. **(20 points)** The goal of this problem is to implement a very simple text retrieval system. Given (as input) a database of documents as well as a query document (all provided in an attached .zip file), write a program, in a language of your choice, to find the document in the database that is the best match to the query. Specifically:

   a. Write a small parser to read each document and convert it into a vector of words.

   b. Compute tf-idf values for each word in every document as well as the query.

   c. Compute the cosine similarity between tf-idf vectors of each document and the query.

   d. Report the document with the maximum similarity value.

5. **(optional)** How much time (in hours) did you spend working on this assignment?