Homework Assignment # 1

Due: Friday, Sep. 30, 2022, 11:59 p.m. Mountain time
Total marks: 100

Question 1. [10 MARKS]

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be iid multivariate Gaussian random variables, with $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ for dimension $d \in \mathbb{N}$. Define sample mean estimator $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$. In this question you will reason about its properties, extending beyond the univariate sample mean you reasoned about before.

(a) [1 MARK] What is the dimension of $\mathbb{E}[\bar{\mathbf{X}}]$?

(b) [2 MARKS] What is $\mathbb{E}[\bar{\mathbf{X}}]$? Write it explicitly in terms of the given variables $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and n. Note that you do not necessarily need to use all three. Show all you steps.

(c) [3 MARKS] What is $\mathrm{Var}[\bar{\mathbf{X}}]$? Write it explicitly in terms of the given variables $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and n. Note that you do not necessarily need to use all three. Show all you steps.

(d) [2 MARKS] Now define $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij}$ where X_{ij} is the j-th element in $\mathbf{X}_i$. What is $\mathbb{E}[\bar{Y}]$? Show all you steps.

(e) [2 MARKS] Show that $\mathrm{Var}[\bar{Y}] = \frac{1}{n}\sum_{j=1}^{d}\sum_{k=1}^{d} \Sigma_{jk}$.

Question 2. [15 MARKS]

In this course, we will be dealing with data that has many features/dimensions. While many of our algorithms and concepts generalize well to these domains, new problems will arise. This is in part due to the weirdness of high dimensional objects. You will often find your intuition being challenged when dealing in larger dimensions and find that algorithms can be challenging to get working. In this question we will explore two ways our intuition can break down in higher dimensions, to help us reason about the behavior of our algorithms in high-dimensional spaces.

(a) [5 MARKS]

Consider a square with sides length = 1 and embed a circle with diameter 1 in this square. In this two-dimensional space, the area of this circle is $\pi \frac{1}{2}^2$. When we move to three dimensions, and embed a sphere with diameter 1 in a cube with sides of length 1, then the volume of the sphere is $\frac{\pi^{3/2}}{2} \frac{1}{2}^3$. In general, for a d-dimensional sphere of diameter 1 embedded in a d-dimensional cube with sides of length 1, the volume of the d-dimensional sphere is $\frac{\pi^{d/2}}{\Gamma(d/2+1)} \frac{1}{2}^d$. The gamma function Γ is a generalization of the factorial function, and for n that are positive integers, we have $\Gamma(n) = (n-1)!$.

Plot this volume in terms of increasing d. (Any online plotting software allows you to use $n!$ for $n \in \mathbb{R}$; to plot $\Gamma(x/2+1)$, use $(x/2)!$). Include this plot in your submission and describe the plot. Mainly, explain how the volume changes with increasing dimension n.

(b) [3 MARKS]

The implications from part a of this question are that (a) our intuition is not adequate to guess how volumes will behave as we increase the dimension and (b) the volume of this sphere goes to zero as d increases. But how can this be! The radius of the hypersphere continues to be the same width as the hypercube, so it touches the edges of the hypercube in all dimensions. What does this tell you about the volume of the hypercube outside of the hypersphere?

(c) [7 MARKS]

Using a computer, generate 1000 samples from a d-dimensional multivariate Gaussian with mean $\mathbf{0}$ and identity covariance matrix. For each point, compute the minimum ℓ_2 distance to the other points (find the nearest neighbor and the corresponding distance). Take the 1000 nearest-neighbor distances and average them, to get `ave-nn-distance`.

Repeat this experiment for $d \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$. What happens to `ave-nn-distance` as d increases? Include the numbers your code generates for `ave-nn-distance`, for each d.

Question 3. [75 MARKS]

Complete the programming assignment in the associated julia file called `A1.jl`.

Homework policies:

Your assignment should be submitted on eClass as a single pdf document and a zip file containing: the code (a .jl file), a .html file of the pluto notebook with all the cells run. The answers must be written legibly and scanned or must be typed (e.g., Latex). All code should be turned in when you submit your assignment. This means submitting the completed Pluto notebook, where you took the Pluto notebook with `todos` and completed them with your implementation. You are not allowed to change any of the imports in the notebook.

Because assignments are more for learning, and less for evaluation, grading will be based on coarse bins. **The grading is atypical**. For grades between (1) 80-100, we round-up to 100; (2) 60-80, we round-up to 80; (3) 40-60, we round-up to 60; and (4) **0-40, we round down to 0**. The last bin is to discourage quickly throwing together some answers to get some marks. The goal for the assignments is to help you learn the material, and completing less than 50% of the assignment is ineffective for learning.

We will not accept late assignments. Plan for this and aim to submit at least a day early. If you know you will have a problem submitting by the deadline, due to a personal issue that arises, please contact the instructor as early as possible to make a plan. If you have an emergency that prevents submission near the deadline, please contact the instructor right away. Retroactive reasons for delays are much harder to deal with in a fair way.

All assignments are individual. All the sources used for the problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see the University of Alberta Code of Student Behaviour.

Good luck!