Due: September 13, 2021, 12:00 PM

NOTE: You are allowed to have at most one page of answer for each problem - total of 3 pages (1 page for problems 1, 2, 3). This restriction is due to limited grader time. If you have more than 3, we will only grade the first 3 pages.

Problem 1 - Sampling

Suppose we plan to sample 10 students from 100 students, where 40 students are in CS and 60 are in EE.

- What are the probabilities of sampling exactly 4 CS and 6 EE students when sampling with replacement, or without replacement?
- Which sampling strategy (with/without replacement) is more appropriate in this scenario? Why?
- Stratified sampling can guarantee the ratio remains the same from each groups (e.g. 4 CS and 6 EE). However, it cannot be applied to all kinds of data. What are the limitations of stratified sampling?

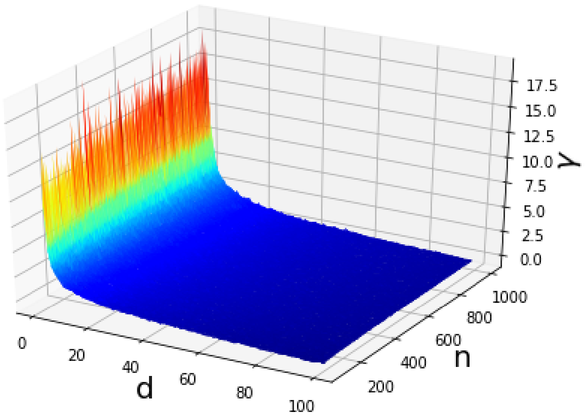

Figure 1: Problem 3 plot

Problem 2 - Curse of Dimensionality

Suppose you randomly generate n points in d dimensional space. For each point $x = (x_1, x_2, ..., x_d)$, $x_i \in [0, 1]$.

1. What is the maximum possible Euclidean distance between any two points? How about the minimum distance?
2. Let us denote the above maximum and minimum pairwise distances among these n points as $\max(d, n)$ and $\min(d, n)$. Define $\gamma(d, n) = \log \frac{\max(d,n)-\min(d,n)}{\min(d,n)}$. An actual plot of $\gamma(n, d)$ is given in Figure 1. What can you observe, especially when comparing to your derived estimates in part 1? (Note that you do not have to provide a proof)

Problem 3 - PCA

- Compute the covariance matrix for the following matrix: $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$
- After performing eigenvalue decomposition on covariance matrix C, when is the smallest eigenvalue equal to zero?
- Assume after performing eigenvalue decomposition on covariance matrix C, the first two eigenvalues are equal. Explain when this could happen.

Problem 4 - (Programming) - Data Prepossessing

Download `https://clo-pfw-prod.s3.us-west-2.amazonaws.com/data/PFW_2021_public.csv` (418MB). Practice the following prepossessing steps using any programming language (Python is recommended):

1. Remove **redundant** attributes (all the same values) and **meaningless** attributes (all values are different).
2. Combine **'date', 'month' and 'year'** into a single attribute.
3. Convert all **categorical** attributes to quoted numbers, e.g. 1 → '1', 2 → '2',. . .
4. Discretize **attribute 'how_many'** into bins of size 10, e.g. 0-10, 11-20,. . .
5. Randomly **sample 100 instances** without replacement and save them as a single .csv file, including attribute names.
6. Only submit **code and sampled data** on Blackboard.