HW1

Start Assignment

Due Nov 25 by 11:59p.m. **Points** 100 **Submitting** a file upload

Problem 1 (Bayesean reasoning and estimation theory)

A patient comes to a clinic complaining of running nose, cough and fever. The doctor in the clinic knows that 100% of people with COVID show these symptoms. On the other hand, 100% of people with common cold also show these symptoms. The doctor also knows that the statistical prevalence of COVID in the population is 1 % and of common cold is 20%

1.1. What is the best estimate of the probability of the patient having COVID-19?

1. 2. To be sure, the doctor does a test on COVID-19, which is known to have 5% false positives, and 0% false negatives. The test comes out positive. What is the new estimate for the probability that the patient has COVID-19?

1.3 How does this estimate change when the test is repeated twice? How many times does one need to repeat the test to get the certainty within 1%?

1.4. Worried that his prevalence estimate might be wrong, the doctor goes out and administers the test to a random sample of 1000 people, out of which 100 come out positive. What is the best estimate (and its accuracy - as discussed in class) of the true COVID prevalence in the population? What if only 10 come out positive?

Problem 2. (More on estimate accuracy) In the Republic of the Great North, the population is split 50/50 between two political parties - the Peelers and the Pinchers.

The Peelers remove the skin from a banana by peeling from the top end, and the Pinchers remove it by pinching at the bottom end.

A sociologist takes a sample of 100 Peelers and finds that the mean height in the sample is 1.98 m and the sampled variance is 30 cm^2. In a similar sample of Pinchers, the average height is 2.02 m with the same sampled variance.

2.1. What is the expected accuracy of these measurements (as discussed in class)?

2.2 What can the sociologist extrapolate about the average height differences between Peelers and Pinchers?

3. 3 What can the sociologist conclude about the effects of the dietary habits on height?

What additional measurements can be done to make such inference? What other factors might the sociologist consider?

Problem 3. (Clustering and dimensionality reduction).

Unless otherwise indicated, for this problem use any resources from **Examples and Python nbs from Schwab-Pankaj review (https://physics.bu.edu/~pankajm/MLnotebooks.html) https://scikit-learn.org/stable/ (https://scikit-learn.org/stable/)** (lots of examples and tools) and **https://keras.io/ (https://keras.io/)**

1. Write your own python code for K-means algorithm. Download the MNIST dataset and cluster the digits using K-means and another clustering algorithm of your choice. Compare the results. Are there differences in clustering?

2. Use t-SNE, PCA to embed the MNIST data set in a lower dimensional space (say, 2 dimensions). Eyeballing the results, how well is the cluster structure preserved? Are there substantial differences between the methods?

(Bonus 10 pt): use spectral/diffusion map methods

3. Perform clustering of MNIST digits using a method of your choice for this lower dimensional dataset. How well is the cluster structure conserved in the lower dimensional space? Which embedding is better?

Problem 4 (Bonus 20 pt).

1. Write your own code for t-SNE. Test that it works using MNIST dataset (either the full or reduced one)

2. Modify the code using a different expression for the neighbourhood function q_{ij} . Instead of using the usual one, use the same function that is used for $p_{i|j}$ (with the same σ)

3. Use the new code to embed the MNIST database in two dimensions. Does it work now? If it does not - why?

4. Can you find another function for q_{ij} that would make it work?