Introduction to Machine Learning Course

Exercise 4 – The Virus Challenge – Clustering

Wet

Background

Scientists and engineers are working hard to develop a vaccine to end the COVID-19 epidemic. In this assignment, you will be playing the role of a data scientist for a major pharma company. You suspect that the virus might have multiple mutations. To prove your suspicions, you plan to analyze virus protein data collected from COVID-19 patients across the country, with unsupervised learning techniques. Your findings could help develop a vaccine and bring cure to mankind. The following are your main research questions:

1. Locate and describe 5 virus mutations by analyzing the virus protein data
2. Given that humanity can produce three vaccines fast, which virus mutations should be targeted?

Good luck!

Assignment

You should submit a process that starts from loading and preparing the data, and up to the completion of the tasks detailed below. This process should include the following:

1. Load the data from protein.csv file.
   Note that the data is not labeled.
2. Prepare the data.
   Focus on data imputation and outliers
3. Use unsupervised ML techniques and exploratory data analysis to accomplish the following:
   a. Identify and differentiate between 5 covid-19 virus mutations.
      Train at least two clustering models
   b. Choose one model to label the given data.
      i. Make a file with the groupings you found
      ii. Use the format in the file “clusters.csv” as a reference
   c. For each mutant virus, identify and describe its characteristics – mutation prevalence, centroid, its nearest mutant virus, etc…
   d. Identify the 5 most useful protein features in discriminating mutation groups
      i. Create a file containing the selected features names
      ii. Use the format in the file “selected_proteins.txt” as a reference

Please submit

1. The Python script file that implements the above
2. A documentation that
   a. Contains the answers of the dry part below.
   b. Briefly describes the data and explains your data preprocessing approach.
   c. Explains how you address the 4 ML tasks in section [3].
   d. Includes any significant decision you took.
   e. Discussion on when you can only provide vaccines for three virus mutants, what strategy would you use to maximize vaccine impact.
3. A file named “clusters.csv” as described in 3.b
4. A file named “selected_proteins.txt” as described in 3.d

Dry

1. Consider m i.i.d samples from a normal distribution $x_i \sim \mathcal{N}(\mu, \sigma^2)$ with unknown mean and variance.

   We proved in the tutorial that $\hat{\mu}_{MLE} = \bar{X} = \frac{1}{m}\sum_i x_i$.

   We also claimed that $\widehat{\sigma^2}_{MLE} = \frac{1}{m}\sum_i (x_i - \hat{\mu}_{MLE})^2$. Prove this estimator.

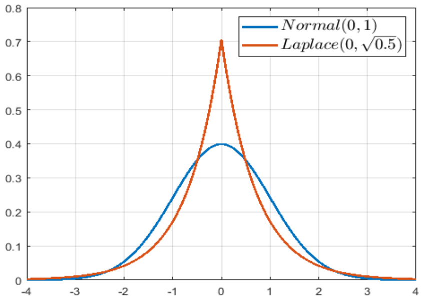
2. Consider a noisy linear model where $y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0,\ 1)$ and $x_i, w \in \mathbb{R}^d$. In tutorial 08 we showed that the LS solution with L^2 regularization (ridge) corresponds to a MAP estimator using an i.i.d Gaussian prior $w_j \sim \mathcal{N}\left(0,\ {}^{1}/_{\lambda}\right)$.

   We will now show that similarly, the LASSO regressor (L^1 regularization) corresponds to a MAP estimator under an i.i.d Laplacian prior $w_j \sim \text{Laplace}(0, b)$.

   The Laplacian pdf's is $p\left(w_j \middle| \mu, b\right) = \frac{1}{2b}\exp\left\{-\frac{|w_j - \mu|}{b}\right\}$.

   Its statistics are $\mathbb{E}\left[w_j\right] = \mu$ and $\text{Var}\left[w_j\right] = 2b^2$.

   The figure compares the pdfs of Gaussian and Laplace distributions with similar mean and variance.



   a. Explicitly write $p(\boldsymbol{w}|\mu = 0, b)$
   b. Show that $\widehat{w}_{MAP} \triangleq \underset{\boldsymbol{w}}{\text{argmax}}\, p(\boldsymbol{w}|\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m, \mu = 0, b)$ corresponds to a LASSO regressor.

      What is the suitable regularization parameter λ in terms of b?
   c. Using the above plot and what you just proved, explain intuitively why LASSO tends to yield sparser solutions in comparison to ridge regression.

3. Given a set of n observations $x_1, \ldots, x_n$
   a. Assume that the observations are Poisson distributed, i.e. $Pr(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$.

      Calculate the Maximum Likelihood Estimator (MLE) for the parameter λ.
   b. Assume that the observations were generated by K Poisson distributions
      i. Define the mixture model
      ii. Use the Expectation Maximization (EM) algorithm to estimate the parameters of the mixture model
         1. Write the likelihood of the complete data and the observed (incomplete) data. Explain the meaning of every component and every variable in each of the expressions
         2. Write and explain the expression to be calculated at the expectation step
         3. Define the F function and explain how it is related to the incomplete likelihood (no need to prove)
         4. Define and explain which parameters are calculated at the maximization step. Derive their expressions from the F function