229352 Statistical Learning for Data Science 2 **Fall 2021**

Lab 2: due Tuesday July 13

Instructor: Donlapark Ponnoprat

Note: The report must be turned in as a PDF file. If you are using the Python Notebook, go to: `File>Download as>pdf(.tex)`.

In this lab, we are going to implement the Naïve Bayes classifier with only Numpy, and we will test our model on the data of 435 US house representatives' voting on various national policies (education, resource, exports, etc.). The representatives are labeled by their associated party: republican (0) or democrat (1). Each column in the `US-representatives-votes.csv` file corresponds to each of the following attributes:

- Column 1: Class: republican=0, democrat=1
- Column 2-17: Votes on 16 different policies: no=0, yes=1

Import `US-representatives-votes.csv` and do the following tasks.

1. Set the Numpy's random seed.
2. Split the data into a training set and a test set.
3. Create an array `Count` of shape `(2,16,2)`, where `Count[l,d,j]` is the number of instances in the training set whose class is `l` and whose `d`-th feature is `j`. In other words,

$$\texttt{Count[l,d,j]} = \#(\texttt{d}\text{-th feature} = \texttt{j} \text{ and Class} = \texttt{l}).$$

4. Use `Count` to create another array `Prob` of shape `(2,16,2)`, where `Prob[l,d,j]` is given by

$$\texttt{Prob[l,d,j]} = P(\texttt{d}\text{-th feature} = \texttt{j} \mid \text{Class} = \texttt{l}) = \frac{\#(\texttt{d}\text{-th feature} = \texttt{j} \text{ and Class} = \texttt{l})}{\#(\text{Class} = \texttt{l})}.$$

   Also, the probability of each class is given by

$$P(\text{Class} = \texttt{l}) = \frac{\#(\text{Class} = \texttt{l})}{\#(\text{all training instances})}.$$

5. To verify that you did the calculations correctly, check if the sum of each row in `Prob[0]` and `Prob[1]` are equal to 1.
6. Use `Prob` to classify each representative in the test set. Report the accuracy of your predictions (Note: I suggest computing the sum of log-probabilities instead of the product of probabilities to prevent numerical underflow).