Georgia Institute of Technology

ECE 6254 - Assignment 5

Fall 2020 - v1.0

- There are 2 problems over 3 pages (including the cover page)
- The problems are not necessarily in order of difficulty.
- All problems are assigned the same weight in the overall grade.
- Each question is graded out of two points (0 for no meaningful work, 1 for partial work, 2 if correct)
- Unless otherwise specified, you should concisely indicate your reasoning and show all relevant work.
- The grade on each problem is based on our judgment of your level of understanding as reflected by what you have written. If we can't read it, we can't grade it.
- Please use a pen and not a pencil if you handwrite your solution.
- **You must submit your exam on Gradescope. Make sure you allocate time for the submission**.

Problem 1: Federalist papers

In this problem, you will explore the use of Naïve Bayes classification applied to a classic text processing problem. Specifically, one of the first usages of the Naïve Bayes approach concerned what is known as the author attribution problem. Here we will tackle a particularly famous instance: who wrote the Federalist Papers?

The *Federalist Papers* were a series of essays written in 1787–1788 meant to persuade the citizens of the State of New York to ratify the Constitution and which were published anonymously under the pseudonym "Publius." In later years the authors were revealed as Alexander Hamilton, John Jay, and James Madison. However, there is some disagreement as to who wrote which essays. Hamilton wrote a list of which essays he had authored only days before being killed in a duel with then Vice President Aaron Burr. Madison wrote his own list many years later, which is in conflict with Hamilton's list on 12 of the essays. Since by this point the two (who were once close friends) had become bitter rivals, historians have long been unsure as to the reliability of both lists.

We will try to settle this dispute using a simple Naïve Bayes classifier. You will need to download the documents which are in the file `fedpapers_split.txt` as well as some starter code in `Federalist Papers.ipynb`, both located on Canvas. The notebook loads the documents and builds a "bag of words" representation of each document. Your goal is to build a Naive Bayes classifier and determine your best guess as to who wrote each of the twelve disputed essays.

[Q1] Explain in your own words why Laplace smoothing is important for text classification.

[Q2] What priors will you use for the two classes? Justify your answer.

[Q3] Explain how you build the Naive Bayes classifier for bag of words classification.

[Q4] Submit your code. Even if correct, the code should be easily understandable to claim full credit.

[Q5] How many of the essays do you think were written by Hamilton and how many were by Madison? (Note that there is no actually verifiable correct answer here.)

Problem 2: Laplace smoothing

Consider a Naive Bayes Model in which features x_j are modeled as conditionally independent given the label y. Consider a dataset $\mathcal{D} \triangleq \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $\mathcal{X} = \{0,1\}^d$ and $\mathcal{Y} \in \{0,1\}$. Since the features are binary, $P_{x_j|y=k}$ is modeled as a Bernoulli distribution with parameter $\mu_{j,k}$, i.e., $P_{x_j|y}(1|k) = \mu_{j,k}$.

[Q1] Assume that the data is such that the first feature x_1 appears in all samples, that is $\forall i \in [\![1, N]\!]$ we have $x_{i,1} = 1$. Show that the MLE for $\mu_{1,0}$ and $\mu_{1,1}$ is

$$\hat{\mu}_{1,0} = \hat{\mu}_{1,1} = 1.$$

[Q2] At testing you receive a previously unseen vector $\mathbf{x}$ such that $x_1 = 0$. What happens when you run your classifier? Explain why one can say that the MLE resulted in overfitting of the data.

[Q3] To circumvent the problem highlighted above, we introduce a *prior* on the parameters that we try to estimate. Specifically we assume here that $\mu_{j,k}$ is a random variable with Beta(β_0, β_1) distribution. Show that the MAP estimate $\mu_{j,k}^{\text{MAP}} \triangleq \text{argmax}_{\mu_{jk}}\, p(\mu_{jk}|\{(\mathbf{x}_i : y_i = k)\})$ is given by

$$\mu_{j,k}^{\text{MAP}} = \frac{N_{j,k}^{(1)} + \beta_0 - 1}{N_k + \beta_0 + \beta_1 - 2}.$$

where $N_{j,k}^{(1)}$ is the number of times feature j takes the value 1 for a data point in class k, and N_k is the number of data points of class k.