EECE5644 Spring 2024 – Homework 2

Submit: Please submit your solutions in a single PDF file using the Canvas/Assignments page. The PDF should include all math, numerical, and visual results. Code can be appended in the PDF or a link to your online code repository can be included. If you point to an online repository, please do not edit the contents after the deadline, because graders may interpret a last-modified timestamp past the deadline as a late submission of the assignment. Only the contents of the PDF will be graded. Please do NOT link from the pdf to external documents online where results may be presented. Any material presented outside the PDF will not be considered for grading. You can use and modify code samples provided along with this handout as you wish to implement the final code needed for the homework.

This is a graded assignment and the entirety of your submission must contain only your own work. You may benefit from publicly available literature including software (not from classmates), as long as these sources are properly acknowledged in your submission. Copying math or code from each other is not allowed and will be considered as academic dishonesty. While there cannot be any written material exchange between classmates, verbal discussions to help each other are acceptable. Discussing with the instructor, the teaching assistant, and classmates at open office periods to get clarification or to eliminate doubts are acceptable.

By submitting a PDF file in response to this take home assignment you are declaring that the contents of your submission, and the associated code is your own work, except as noted in your citations to resources.

Question 1 (10%)

Consider real random variable $X \sim N(\mu, \sigma^2)$, i.e., that follows a normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Suppose we generate a dataset $\mathcal{D} = \{x_1\}$ having just *a single sample* from this distribution.

1. What is the ML estimate of the mean $\hat{\mu}$?
2. What is the (biased) ML estimate of the variance $\hat{\sigma}^2$?
3. What is the corresponding unbiased estimate of the variance?
4. Which of the two estimates makes more sense? Explain your answer.

Question 2 (20%)

Consider a categorical random variable X with K possible categories, i.e., $X \in \{1, \ldots, K\}$. For example, in the case of dice rolls, we have $K = 6$ categories, each corresponding to possible outcomes

$$\{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}.$$

Denote by

$$\theta_k = P(X = k)$$

the probability that we observe outcome k. Suppose that we collect a dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ of i.i.d. samples from this distribution.

1. Show that

   $$p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k}.$$

   where $N_k = \sum_{i=1}^{N} \mathbb{1}_{x_i = k}$ is the number of times outcome k occurs in $\mathcal{D}$.

2. Consider the Dirichlet prior distribution

   $$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

   where $\alpha_k \geq 1$, for $k = 1, \ldots, K$, are called hyperparameters.[1] Show that, under this prior, the posterior is also Dirichlet, i.e.,

   $$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{k=1}^{K} \boldsymbol{\theta}_k^{\alpha'_k - 1}$$

   for appropriately defined $\alpha'_k \geq 1$, $k = 1, \ldots, K$, and find these values.

[1] To get the full distribution, one would need to multiply the r.h.s. of this equation with a constant so that $\int_{\boldsymbol{\theta} \in S} p(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$, where $S \subseteq \mathbb{R}^K$ is the set of valid hyperparameters that sum up to 1, i.e. $S = \{\boldsymbol{\theta} \geq \mathbf{0} : \sum_{i=1}^{k} \theta_i = 0\}$. Set S is also known as the *K-dimensional simplex.*

3. Show that if $\alpha_1 = \alpha_2 = \ldots = \alpha_K = 1$, then the prior becomes the uniform distribution over valid $\boldsymbol{\theta}$ (i.e., this is an *uninformative prior*).

4. In class we discussed about the following three estimators of θ_k, $k = 1, \ldots, K$, each corresponding to ML, MAP, and the mean of full-Bayesian estimation:

$$\hat{\theta}_k^{\text{ML}} = \frac{N_k}{N} \qquad \hat{\theta}_k^{\text{MAP}} = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K} \qquad \mathbb{E}[\theta_k \mid \mathcal{D}] = \frac{N_k + \alpha_k}{N + \alpha_0} \tag{1}$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$. Suppose, as in Q1, that $\mathcal{D} = \{x_1\}$, i.e., we have only collected one sample. What would each of these estimates look like for each $k = 1, \ldots, K$ under an uninformative prior? Which of these makes most sense, and why?

Question 3 (40%)

1. Provide the formulas for the discriminant functions of the categorical (a.k.a. "multinoulli") Naïve Bayes classifier with a Dirichlet prior, when (a) parameters are learned via MAP estimation and (b) when parameters estimation is full Bayesian. How do the two differ in the case where the prior is uninformative?

2. Download the mushrooms dataset from the UCI repository:

   `https://archive.ics.uci.edu/dataset/73/mushroom`

   As described in the dataset's documentation:[2]

   > *This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one.*

   Create a random split 80% of the dataset into training set and 20% into a test set. Train a categorical Naïve Bayes classifier with different values of smoothing hyperparameter α, common amongst all features, ranging from 2^{-15} to 2^{5}. Plot the predictive performance of the trained classifier on your test set, as measured by ROC AUC, accuracy, and F1 scores. Report the parameters of the model for the α value that maximizes the AUC.

3. Repeat the same experiment with split in which you use only 1% of the dataset for training, and 99% of the dataset for testing. In this case, you should be careful about dealing with missing categories in the training set. What differences do you see?

4. Should you care more about the ROC AUC or accuracy in this dataset? Explain your answer.

To code the categorical Naïve Bayes classifier, you are encouraged to use Python's Scikit Learn class `sklearn.naive_bayes.CategoricalNB`: read its documentation! Code samples, including a program that processes the dataset from the UCI repository so that it can be handled by this class, are provided along with this handout.

[2] See file `agaricus-lepiota.names`.

Question 4 (30%)

Download the Sentence Classification dataset from the UCI repository:

`https://archive.ics.uci.edu/dataset/311/sentence+classification`

1. Create a random split of 80% of the dataset into training set and 20% into a test set. Train a multinomial naïve Bayes classifier, using a "bag of words" representation. Use again different values of smoothing hyperparameter α, common amongst all features, ranging from 2^{-15} to 2^{5}, and measure only accuracy on the test set this time. Repeat these experiment 10 times, using different random seeds to generate different random splits, and compute average accuracy and standard deviation. Plot the average accuracy a function of α, showing also confidence intervals around the average (± the standard deviation). For the α value that maximizes the average accuracy, report the words corresponding to the 5 highest parameters in each class.

You should be able to repurpose code from Q3 for most of this question. To code the multinomial naïve Bayes classifier, you are encouraged to use Python's Scikit Learn class `sklearn.naive_bayes.MultinomialNB`: read its documentation! A program that processes the dataset from the UCI repository so that it can be handled by this class is provided along with this handout. The program directly converts the above dataset into samples where each word is in a "bag of words" representation.