New York University Tandon School of Engineering
Computer Science and Engineering

CS-UY 4563: Written Homework 3.
Due Wednesday, March 4th, 2020, 11:59pm.

Collaboration is allowed on this problem set, but solutions must be written-up individually. Please list the names of any collaborators at the top of your solution set, or write "No Collaborators" if you worked alone.

Problem 1: Practice with Probabilistic Models (15pts)

Describe a probabilistic model for each of the following data sets. There is no right answer, but your model should be reasonable and plausibly characterize the data. For each model, **separately list what parameters** which would need to be learned from past data. Also, be careful not to include in the model any data variables not explicitly listed!

(a) Each data example corresponds to an apartment and has parameters:

$$(x_1, x_2, y) = (\text{ZIP code}, \text{size}, \text{price}).$$

The "ZIP code" is for where the apartment is located, "size" is the apartment's square footage, and "price" is the current monthly rental price in dollars.

(b) Each data example corresponds to a minute in the day and has parameters:

$$(x, y) = (\text{time}, \text{number of riders}).$$

The "time" is a time of day specified in the number of minutes past midnight (e.g. 9:32am is represented as $572 = 9 \times 60 + 32$) and "number of riders" is the current number of riders on the NYC subway system.

(c) Each data example corresponds to a Netflix show and has parameters:

$$(\vec{x}_1, x_2, y) = (\text{show description}, \text{genre}, \text{rating}).$$

The "show description" is a binary bag-of-words vector corresponding to the text summary of a show, "genre" is a category like documentary, drama, romcom, historical fiction, etc, and "rating" is an average numerical user rating.

Problem 2: Gaussian Naive Bayes (15pts)

In class it was briefly mentioned that the Naive Bayes Classifier can be extended to predictor variables with continuous values (instead of just binary variables). We will derive such an approach here

Consider a data set where each example $(\vec{x}, y)$ contains a data vector $\vec{x} \in \mathbb{R}^d$ and a label $y \in \{0, 1\}$. Each y is modeled a Bernoulli random variable, which equals 1 with probability p and 0 with probability $1 - p$. To model $\vec{x}$ we have two lists of mean/variances pairs:

$$(\mu_{0,1}, \sigma^2_{0,1}), (\mu_{0,2}, \sigma^2_{0,2}), \ldots, (\mu_{0,d}, \sigma^2_{0,d}) \qquad \text{and} \qquad (\mu_{1,1}, \sigma^2_{1,1}), (\mu_{1,2}, \sigma^2_{1,2}), \ldots, (\mu_{1,d}, \sigma^2_{1,d}).$$

If y equals 0, then the j^{th} entry of $\vec{x}$ is modeled as an *independent* Gaussian (normal) random variable with mean $\mu_{0,j}$ and variance $\sigma^2_{0,j}$. Alternatively, if y equals 1, then the j^{th} entry of $\vec{x}$ is modeled as an independent Gaussian random variable with mean $\mu_{1,j}$ and variance $\sigma^2_{1,j}$.

(a) Given a training data set $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$ write down expressions for estimating all model parameters $\mu_{i,j}$ and $\sigma^2_{i,j}$ from the data.

(b) Given a new unlabled predictor vector $\vec{x}_{new}$ we would like to predict class label y_{new} using a *maximum a posterior* (MAP) estimate. In other words, we want to choose y_{new} to maximize the posterior probability $p(y_{new} \mid \vec{x}_{new})$. Write down an expression for $p(y_{new} \mid \vec{x}_{new})$ using Bayes Rule.

(c) Using your result from part (b) write down a final mathematical equation (or pseudocode) for computing $p(y_{new} = 0 \mid \vec{x}_{new})$ and $p(y_{new} = 1 \mid \vec{x}_{new})$. **Hint:** A correct answer should involved the PDF of a Gaussian random variable, and incorporate all model parameters $\mu_{i,j}$ and $\sigma^2_{i,j}$.

(d) How can your answer from part (c) be simplified if you only seek to compute $C \cdot p(y_{new} \mid \vec{x}_{new})$ for some constant C you choose? What if you only seek to compute $B \cdot \log\left(C \cdot p(y_{new} \mid \vec{x}_{new})\right)$ for some constants B, C you choose? Can either or both of these simplified expression be used in deciding on the MAP estimate for y_{new}?

Problem 3: Bayesian Central Tendency (10pts)

Let's revisit Question 3 on Written Homework 1 from a Bayesian perspective. This was the question about loss functions for measures of central tendency.

(a) Suppose we have a data set of scalar numbers $x_1, \ldots, x_n$. Assume a Bayesian probabilistic model in which the numbers are drawn from a Gaussian distribution with unknown mean μ and variance σ^2. We have no prior information on μ and σ^2: we assume all parameters are equally likely. Prove that the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a MAP estimate for the unknown parameter μ.

(b) Now assume a Bayesian probabilistic model in which the numbers are drawn from a Laplace Distribution with unknown mean μ and variance $2b^2$. Prove that the sample median is a MAP estimate for the unknown parameter μ. **Hint:** Look back at Homework 1.

(c) (**Extra Credit − 5pt**) Assume a Bayesian probabilistic model in which the numbers are drawn from a uniform distribution centered at μ and of width $2b$. I.e. each x_i is drawn uniformly from the interval $[\mu - b, \mu + b]$. Further assume that b itself is modeled as a Gaussian random variable with mean 0 and variance 1. So smaller values of b are more likely. What is a MAP estimate for μ?