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.

One option is to model price as a linear function of size, with parameters of the linear model depending on the zip code. E.g. price $= \beta_{0,z} + \beta_{1,z} \cdot \text{size} + \eta$ where we have different parameters $\beta_{0,z}, \beta_{1,z}$ for each zip code z and η is a Gaussian random variable with mean 0 and covariance σ^2. The zip code can be modeled as a discrete categorical random variable – i.e. for each zip z we have a probability $p(z)$ and $\sum_z p(z) = 1$. We set the property zip to z with probability $p(z)$. The parameters we need to learn are:

- $\beta_{0,z}, \beta_{1,z}$ for all z.
- σ^2.
- $p(z)$ for all z.

(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.

The following is a very rough model, but it's one option: Subway ridership tends to be cyclic, with more riders in the morning and evenings, going to and from work. Model x as a uniform random variable from $0, 1, \ldots, 1440$ (there are 1440 minutes in a day). Model y as $y = \sin(f \cdot x + p) + \beta + \eta$ where f, p, β are constants and η is a Gaussian random variable with mean 0 and covariance σ^2. The parameters we need to learn are:

- f, p, β
- σ^2.

(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.

Not going to write this one out fully. One option is to use a bag-of-words model like we did in class for spam, where the probability of each work depends on the genre. There are lots of ways to do ratings. For example, you could do as a Gaussian distribution with different mean for each genre. A really cool solution could also use the rating to generate the bag of words.

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.

First let p be the fraction of all data with label 1. Next let $S_0 \subseteq \{1, \ldots, n\}$ contain all indices k such that $y_k = 0$ and let $S_1 \subseteq 1\{1, \ldots, n\}$ contain all indices k such that $y_k = 1$. For each $i \in \{0, 1\}$ and $j \in 1, d$, let $\mu_{i,j} = \frac{1}{|S_i|} \sum_{k \in S_i} \vec{x}_k[j]$. Then let $\sigma_{i,j} = \frac{1}{|S_i|} \sum_{k \in S_i} (\vec{x}_k[j] - \mu_{i,j})^2$.

(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.

$p(y_{new} \mid \vec{x}_{new}) = \frac{p(\vec{x}_{new} | p(y_{new}) p(y_{new})}{p(\vec{x}_{new})}$

(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 $\sigma^2_{i,j}$.

- We compute $p(y) = p$ if $y = 1$ and $p(y) = 1 - p$ if $y = 0$
- We compute $p(\vec{x}_{new} \mid y) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2_{y,j}}} e^{-(\vec{x}_{new}[j] - \mu_{y,j})^2 / 2\sigma^2_{y,j}}$.
- Finally, we compute $p(\vec{x}_{new}, \mid i) = p \cdot p(\vec{x}_{new} \mid 1) + (1 - p) \cdot p(\vec{x}_{new} \mid 0)$.
- As a final step we use all these parts to evaluate the expression from part (b).

(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 $B \cdot \log(C \cdot p(y_{new} \mid \vec{x}_{new}))$ 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}?

- For the first part, we have: $Cp(y_{new} \mid \vec{x}_{new}) = p(y) \prod_{j=1}^{d} \frac{1}{\sigma_{y,j}} e^{-(\vec{x}_{new}[j] - \mu_{y,j})^2 / 2\sigma^2_{y,j}}$.
- For the second, we have $B \cdot \ln(C \cdot p(y_{new} = 0 \mid \vec{x}_{new})) = \log(p(y)) - \sum_{j=1}^{d} \log(\sigma_{y,j}) - \sum_{i=1}^{d} (\vec{x}_{new}[j] - \mu_{y,j})^2 / 2\sigma^2_{y,j}$

Both can be used since Cx and $B \log(Cx)$ are monotonic (order preserving) functions as long as C and B are positive.

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.

Using Bayes rule, we have that $p(\mu \mid x) \sim p(x \mid \mu)$, since we assume a uniform prior on μ. Then $p(x \mid \mu) \sim \prod_{i=1}^{n} e^{-(x_i-\mu)^2/\sigma^2}$ and $\log(p(x \mid \mu)) \sim \sum_{i=1}^{n} -(x_i-\mu)^2$. So the MAP estimate is the μ which maximizes $\sum_{i=1}^{n} -(x_i-\mu)^2$, which is the same as minimizing $\sum_{i=1}^{n}(x_i-\mu)^2$. In Homework 1 we already proved that the mean minimizes this expression.

(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.

Using Bayes rule, we have that $p(\mu \mid x) \sim p(x \mid \mu)$, since we assume a uniform prior on μ. Then $p(x \mid \mu) \sim \prod_{i=1}^{n} e^{-|x_i-\mu|/b}$ and $\log(p(x \mid \mu)) \sim \sum_{i=1}^{n} -|x_i-\mu|^2$. So the MAP estimate is the μ which minimizes $\sum_{i=1}^{n} |x_i-\mu|$, which we already proved to be the median in 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?

The MAP estimate is $\frac{\min_i(x_i)+\max_i(x_i)}{2}$, which is called the midrange. We already saw this estimate on Homework 1.