# CSC 580 Homework 3

Due: 11/16 (Wed) 5pm

**Instructions:**

- If you use math symbols, please define it clearly before you use it (unless they are standard from the lecture).
- You must provide the derivation for obtaining the answer and full source code for whatever problem you use programming. Please email your source codes to `csc580homeworks@gmail.com`.
- Please use the problem & subproblem numbering of this document; do not recreate or renumber them.
- Submit your homework on time to gradescope. NO LATE DAYS, NO LATE SUBMISSIONS ACCEPTED.
- The submission must be one single PDF file (use Acrobat Pro from the UA software library if you need to merge multiple PDFs).
- Please include your answers to all questions in your submission to Gradescope. (Do not store your answers in your source codes or Jupyter notebooks - I will not look at them by default.)
  - You can use word processing software like Microsoft Word or LaTeX.
  - You can also hand-write your answers and then scan it. If you use your phone camera, I recommend using TurboScan (smartphone app) or similar ones to avoid looking slanted or showing the background.
  - Watch the video and follow the instruction: `https://youtu.be/KMPoby5g_nE` .
- Collaboration policy: do not discuss answers with your classmates. You can discuss HW for the clarification or any math/programming issues at a high-level. If that is the case, please mention who you've talked to in your submission. Declaring your collaborators will not result in deduction of points; instead, failure to declare your collaborators counts as academic integrity violation.

**Problem 1. Probabilistic Reasoning.**
(a) Denote background evidence by event $E$. Suppose $X, Y$ are two other events. Prove the conditional version of Bayes' rule:

$$P(X \mid Y, E) = \frac{P(Y \mid X, E)P(X \mid E)}{P(Y \mid E)}$$

(b) Consider the following Bayesian network (picture by Lawrence Saul):

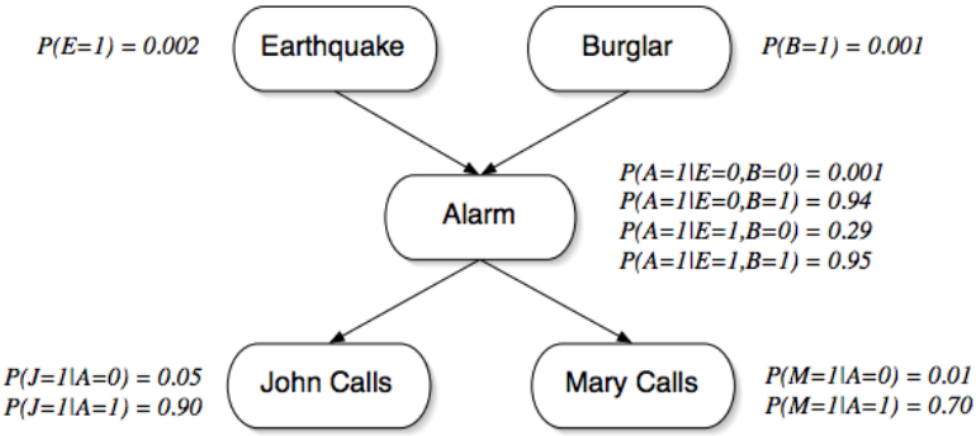

Figure 1: A Bayesian network for a house

(i) Using Bayes' rule, calculate $P(E = 1 \mid A = 1)$. Is it larger than $P(E = 1)$? Does it make intuitive sense?

(ii) Using Bayes' rule, calculate $P(E = 1 \mid A = 1, B = 1)$. Is it larger than $P(E = 1 \mid A = 1)$? Use this as an example to demonstrate the "explain away" phenomenon discussed in class.

(iii) Is $E \perp\!\!\!\perp B \mid M$? Justify your answer. Does your answer match your intuition?

(iv) Calculate the joint distribution of $(J, M)$. Is $J \perp\!\!\!\perp M$? Is $J \perp\!\!\!\perp M \mid A$? Justify your answers.

**Problem 2. Maximum Likelihood Estimation.**

(a) Let $(n_1, \dots, n_K) \sim \mathsf{Multinomial}(n, p)$ where $p \in \Delta^{K-1}$ (recall that $\Delta^{K-1}$ denotes the $K$-dimensional probability simplex). We'd like to estimate $p$ using this *single observation*.

(i) Write down the maximum log likelihood optimization problem (it is okay to omit terms that do not matter w.r.t. the optimization problem). Don't forget to specify the constraints.

(ii) Write down the Lagrangian of the optimization problem you have in (i).

(iii) Solve (ii) to find the MLE solution $\hat{p}$. (**hint**: we did something similar to this in the naive Bayes model lecture.)

(b) Suppose you model a dataset of $n$ iid $D$-dimensional sensor measurements $S = (x_1, \dots, x_n)$ (where each $x_i \in \mathbb{R}^D$) using a spherical Gaussian distribution, where $\mu \in \mathbb{R}, \sigma > 0$ are the distribution parameters:

$$p(x; \mu, \sigma) = N(x; \mu, \sigma^2 I_D) = (\frac{1}{2\pi\sigma^2})^{\frac{D}{2}} \exp(-\frac{\|x - \mu\|_2^2}{2\sigma^2})$$

(i) What is the maximum likelihood estimator of this model given $S$?

(ii) As discussed in Piazza, there are some errors in the CIML book Eqs. (16.14-16.16) and (16.20-16.22). Based on your results in (i), can you write down their respective correct formulae? Justify your answer.

**Problem 3. Language Identification with Naïve Bayes**

Implement a character-based Naive Bayes classifier that classifies a document as English, Japanese, or Spanish - all written with the 26 lower case characters and space.

The dataset is languageID.tgz and can be found in our Piazza page. You need to unpack it. This dataset consists of 60 documents in English, Japanese, and Spanish. The correct class label is the first character of the filename: $y \in \{E, J, S\}$.

We will be using a character-based multinomial naïve Bayes model. You need to view each document as a bag of characters, including space (we say 'bag' because we ignore the order). We have made sure that there are only 27 different types of printable characters (a to z, and space) – there may be additional control characters such as new-line, please ignore those. Your vocabulary will be these 27 character types.

Here is the model. Let $n_i$ be the length of the $i$-th document (same as the total number of characters in the document including the space character). For $i \in [n] := \{1, ..., n\}$,

- Generate $y_i \in \{e, j, s\}$ from $\mathsf{Categorical}(\pi)$ where $\pi \in \Delta^2$ (i.e., $\pi_1 = \mathbb{P}(y_i = E), \pi_2 = \mathbb{P}(y_i = J), \pi_3 = \mathbb{P}(y_i = S)$).
- Generate $\forall j \in [n_i],\;\; x_{i,j} \sim \mathsf{Categorical}(\theta_{y_i})$ where $\theta_y \in \Delta^{26}, \forall y \in \{E, J, S\}$.

**Background on smoothing**: When estimating a multinomial parameter, add-$\epsilon$ smoothing is a popular technique. This amounts to performing the MLE, i.e., count the occurrences and normalize it, assuming that we have $\epsilon > 0$ additional observations for each outcome (note: $\epsilon$ does not have to be an integer). For example, if $n_1, \ldots, n_K \sim \mathsf{Multinomial}(n; p)$, then we estimate $p$ by

$$\hat{p} = \frac{\epsilon + n_i}{\sum_{l=1}^{K}(\epsilon + n_l)} \;.$$

This helps avoiding the issue of assigning zero probability for test data points.

(a) Use files [y]0.txt to [y]9.txt where $y \in \{E, J, S\}$ in each language as the training data. Estimate the prior probabilities $\pi$ with add-1 smoothing and print them. (Hint: Store all probabilities here and below in $\log()$ internally to avoid underflow. This also means you need to do arithmetic in log-space. But answer questions with probability, not log probability.)

(b) Using the same training data, estimate the class conditional distribution for English (i.e., $\theta_E$) using add-1 smoothing. Ensure that the components of the vector $\theta_E$ is ordered with the following order: $(a, \ldots, z, \mathsf{space})$. Write down the formula for add-1 smoothing in this case. Print $\theta_E$ which is a vector with 27 elements. Do the same for $\theta_J$ and $\theta_S$.

(c) Treat e10.txt as a test document $x$. Represent $x$ as a count vector $c(x) \in \mathbb{N}^{27}_{\geq 0}$. This is called a bag-of-words vector (it is actually bag of characters, here, but bag-of-words is a standard terminology in the field of natural language processing). Print the bag-of-words vector $c(x)$.

(d) Let $\theta_{y,i}$ be the $i$-th component of $\theta_y$. Write down mathematically how you will compute $\hat{p}(x \mid y)$ for $y = \{E, J, S\}$ with our estimated parameters. Here, we used $\hat{p}$ to denote that it is evaluated using the estimated probability. Then, compute and show the following three: $\hat{p}(x \mid y = E), \hat{p}(x \mid y = J), \hat{p}(x \mid y = S)$.

(e) Write down mathematically the posterior $\hat{p}(y \mid x)$ using Bayes rule and your estimated prior and likelihood. Show the three values: $\hat{p}(y = E \mid x), \hat{p}(y = J \mid x), \hat{p}(y = S \mid x)$. Show the predicted class label of $x$ based on your estimated model.

(f) Evaluate the performance of your classifier on the test set (files [y]10.txt to [y]19.txt in three languages). Present the performance using a confusion matrix. A confusion matrix summarizes the types of errors your classifier makes, as shown in the table below. The columns are the true language a document is in, and the rows are the classified outcome of that document. The cells are the number of test documents in that situation. For example, the cell with row = English and column = Spanish contains the number of test documents that are really Spanish, but misclassified as English by your classifier.

| | English | Spanish | Japanese |
|---|---|---|---|
| English | | | |
| Spanish | | | |
| Japanese | | | |

(g) Repeat the same experiment as (f), but this time with training and test examples induced by loading only the first 5 rows of the respective documents. Report the new confusion matrix.

**Problem 4. Principal Component Analysis**

Download three.txt and eight.txt, which can be found in our Piazza page. Each has 200 handwritten digits. Each line is for a digit, vectorized from a 16x16 gray scale image.

(a) Each line has 256 numbers: they are pixel values (0=black, 255=white) vectorized from the image as the first column (top down), the second column, and so on. Visualize using python the two gray scale images corresponding to the first line in three.txt and the first line in eight.txt.

(b) Put the two data files together (threes first, eights next) to form a $n \times d$ matrix $X$ where $n = 400$ digits and $d = 256$ pixels. The $i$-th row of $X$ is $x_i^\top$, where $x_i \in \mathbb{R}^d$ is the $i$-th image in the combined data set. Compute the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Visualize $\bar{x}$ as a 16x16 gray scale image.

(c) Center $X$ using $\bar{x}$ above. Then form the sample covariance matrix $S = \frac{X^\top X}{n-1}$. Show the 5x5 submatrix $S(1 \ldots 5, 1 \ldots 5)$.

(d) Use appropriate software/library to compute the two largest eigenvalues $\lambda_1 \geq \lambda_2$ and the corresponding eigenvectors $v_1, v_2$ of $S$. For example, in python one can use `scipy.sparse.linalg.eigs`. Show the value of $\lambda_1, \lambda_2$. Visualize $v_1, v_2$ as two 16x16 gray scale images. Hint: you may need to scale the values to be in the valid range of grayscale ([0, 255] or [0,1] depending on which function you use). You can shift and scale them in order to show a better picture. It is best if you can show an accompany 'colorbar' that maps gray scale to values.

(e) Now we project (the centered) $X$ down to the two PCA directions. Let $V = [v_1, v_2]$ be the $d \times 2$ matrix. The projection is simply $XV$. (To be precise, these are the coefficients along the principal directions, not the projection itself.) Show the resulting two coordinates for the first line in three.txt and the first line in eight.txt, respectively.

(f) Report the average reconstruction error $\frac{1}{n}\sum_{i=1}^n \|x_i V V^\top - x_i\|^2$, where $x_i \in \mathbb{R}^{1 \times d}$ is the $i$-th row of the centered data matrix $X$.

(g) Now plot the 2D point cloud of the 400 digits after projection. For visual interest, color points in three.txt red and points in eight.txt blue. But keep in mind that PCA is an unsupervised learning method and it does not know such class labels.

**Problem 5: Project check-in.**
Please answer the following in a point-by-point manner.

(a) Respond to my feedback on your project proposal. (Of course, if some points in my feedback does not make sense to you, please point them out – I am happy to discuss more.)

(b) (Answer it around Nov. 14) Where are you in terms of your project progress:

- What have you achieved?
- What difficulties have you encountered?
- What remains to be done? List all your todo items and give your deadline for each of them. (The last homework will be shorter, to give you more time for your project.) Make sure you allocate 1-2 weeks for writing up the project report.
- What difficulties do you anticipate, and what are your contingency plans?