Homework 1
Probability, MLE, MAP, KNN and Naive Bayes[1]

CMU 10-701: Introduction to Machine Learning (Fall 2020)

piazza.com/cmu/fall2020/10701/home
OUT: Wednesday, Sep 9th, 2020

DUE: Wednesday, Sep 23rd, 2020, 11:59pm
TAs: Jie Jiao, Clay Yoo

START HERE: Instructions

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 2.1"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only. See the Academic Integrity Section on the course site for more information: https://www.cs.cmu.edu/~epxing/Class/10701-20/about.html
- **Late Submission Policy:** See the late submission policy here: https://www.cs.cmu.edu/~epxing/Class/10701-20/about.html
- **Submitting your work:**
  - **Gradescope:** There will be two submission slots for this homework on Gradescope: Written and Programming.
    For the written problems such as short answer, multiple choice, derivations, proofs, or plots, we will be using the written submission slot. Please use the provided template. The best way to format your homework is by using the Latex template released in the handout and writing your solutions in Latex. However submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. Each derivation/proof should be completed in the boxes provided below the question, **you should not change the sizes of these boxes** as Gradescope is expecting your solved homework PDF to match the template on Gradescope. If you find you need more space than the box provides you should consider cutting your solution down to its relevant parts, if you see no way to do this it please add an additional page a the end of the homework and guide us there with a 'See page xx for the rest of the solution'.

[1]Compiled on Thursday 10th September, 2020 at 12:28

You are also required to upload your code, which you wrote to solve the final question of this homework, to the Programming submission slot. Your code may be ran by TAs so please make sure it is in a workable state.

Regrade requests can be made after the homework grades are released, however this gives the TA the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted.

For multiple choice or select all that apply questions, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeXusers, use ■ and ●for shaded boxes and circles, and don't change anything else.

1 Probability Review [10pts]

A group of travellers find themselves lost in a cave. They come upon 3 tunnels A, B, C. Both tunnels A and B are closed loops that do not lead to an exit and in fact lead right back to the entrance of the 3 tunnels. Tunnel C is the tunnel which leads to the exit. If they go through tunnel A, then it takes 2 days to go through the tunnel. If they go through tunnel B, then it take 1 day to go through the tunnel. If they go through tunnel C, then they immediately leave the cave. Suppose the travellers choose tunnels A, B and C with constant probability 0.3, 0.5, 0.2 every time. (For the following questions please round your answer up to 4 digits.)

1. **[4 pts]** Suppose we record down the travellers choices into a sequence (e.g., $ABBA\ldots C$). What is the probability that the pattern AAB appears in the sequence before any BAA appears?
   Note: You should also count cases where AAB appears in the sequence and BAA does not.

2. **[2 pts]** What is the expected number of days that the travellers will be lost in the cave?

3. [**4 pts**] What is the variance of days that the travellers will be lost in the cave? (Hint: To compute $Var(T)$ for a random variable T, you can either compute $\mathrm{E}[T^2]$ first and then $Var(T)$ or directly compute the variance using the law of total variance.)

2 MLE and MAP [20pts]

2.1 MLE with Exponential Family [5 pts]

Exponential family distribution has the form $P(x|\theta^*) = h(x)\exp(\theta^*\,\phi(x) - A(\theta^*))$. It might look unfamiliar but in fact many well-known distributions including Gaussian, Bernoulli, Geometric and Laplace distributions belong to this family[2]. Suppose we are given n i.i.d samples $X_n = \{x_1, x_2, \ldots, x_n\}$ drawn from the distribution $P(x|\theta^*)$, derive the Maximum Likelihood Estimator $\widehat{\theta}_{\text{MLE}}$ for this true parameter θ^*. Here A and A' are some functions that you can assume are invertible.

[2]To see the parameter setting for each of these distributions, which makes them become special cases of exponential distributions you can check https://en.wikipedia.org/wiki/Exponential_family#Table_of_distributions.

2.2 MLE and MAP with Weibull Distribution [15 pts]

1. **[5 pts]** The Weibull distribution has the form

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \quad x \geq 0.$$

with the parameters $k > 0, \lambda > 0$. When $k = 1$, this is an exponential distribution and when $k = 2$, this is a Rayleigh distribution. For our purposes say k is known. We obtain n i.i.d. data points $x_1, x_2, \ldots, x_n$ from the Weibull distribution. Find the MLE estimate $\hat{\lambda}$.

2. **[8 pts]** Let $t = \lambda^k$. Suppose t has a prior distribution in the form of inverse-gamma with probability density function

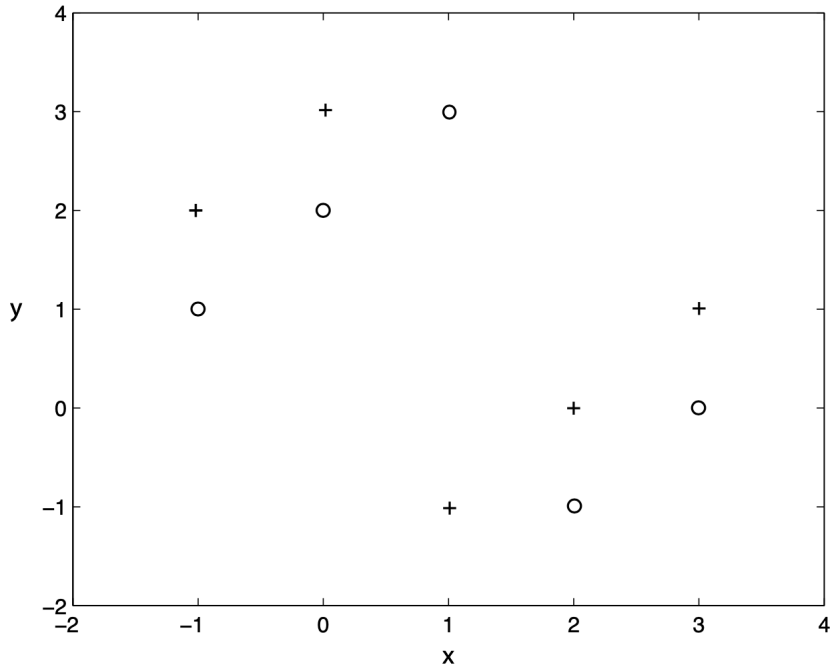
$$f(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{t}\right)^{\alpha+1} e^{-\beta/t}.$$

The parameters $\alpha > 0, \beta > 0$ are both known. Find the posterior distribution of t given x and the MAP estimate $\tilde{\lambda}$.

3. **[2 pts]** Assume $\sum_{i=1}^{n} x_i^k \to \infty$ as $n \to \infty$ for Weibull distribution. Compare the MLE ($\hat{\lambda}$) and the MAP ($\tilde{\lambda}$) as $n \to \infty$ and describe your findings.

3 K-Nearest Neighbors [10 Points]

1. [**2pt**] Consider K-NN using Euclidean distance on the following data set (each point belongs to one of two classes: + and ∘).



(a) [**1pt**] What is the leave one out cross validation error when using 1-NN?

(b) [**1pt**] Which of the following values of k leads to the minimum leave-one-out cross validation error: 3, 5 or 9? What is the error for that k? (If there is a tie, please elaborate)

2. **[2pt]** Consider k-fold cross-validation. Let's consider the tradeoffs of larger or smaller k (the number of folds). Please select one of the multiple choice options.

   With a higher number of folds, the estimated error will be, on average,

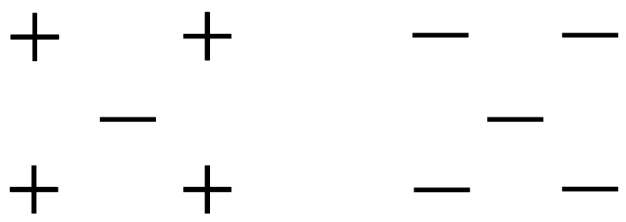
   ◯ Higher

   ◯ Lower

   ◯ Same

   ◯ Can't tell

3. **[1pt]** For the following dataset, circle the classifier which has larger Leave-One-Out Cross- validation error.



   ◯ 1-NN

   ◯ 3-NN

4. [**5pt**] KNN Black Box

   (a) [**3pt**] In a KNN classification problem, assume that the distance measure is not explicitly specified to you. Instead, you are given a "black box" where you input a set of instances P1, P2, .. Pn and a new example Q, and the black box outputs the nearest neighbor of Q, say Pi and its corresponding class label Ci. Is it possible to construct a k-NN classification algorithm (w.r.t the unknown distance metrics) based on this black box alone? If so, how and if not, why not?

   (b) [**2pt**] If the black box returns the j nearest neighbors (and their corresponding class labels) instead of the single most nearest neighbor (assume j != k) , is it possible to construct a k-NN classification algorithm based on the black box ? If so how, and if not why not ?

4 Naive Bayes [20 Points]

Suppose we let $X = (x^1, x^2, .., .x^n)$ denote the features, and $y \in \{0, 1\}$ denote the label. Note that in any generative model approach, we model the conditional label distribution $P(y|X)$ via the conditional distribution of features given the label $P(X|y)$:

$$P(y|X) \propto P(X|y)P(y) \tag{1}$$

1. **[1pt]** Rewrite the conditional distribution in (1) under the Naïve Bayes assumption that the features are conditionally independent given the label.

2. **[4pt]** Suppose that each feature x^i takes values in the set $\{1, 2, ..., K\}$. Further, suppose that the label distribution is Bernoulli, and the feature distribution conditioned on the label is multinomial. Please give detailed step by step derivations for the following questions.

   (a) **[1pt]** What is the total number of parameters of the model under the Naïve Bayes assumption?

   (b) **[1pt]** What is the total number of parameters of the model without the Naïve Bayes assumption?

(c) [**2pt**] Suppose we change the set of values that y takes, so that $y \in \{0, 1, ..., M-1\}$. How would your answers change in both cases (with/out Naïve Bayes assumption)?

3. **[6pt]** Suppose each feature x^i takes values in the set $\{0, 1, \ldots, K-1\}$. Suppose the label distribution is Bernoulli with $P(y = 1) = \pi$, and the distribution for a given feature x^i conditioned on the label is $P(x^i = j | y = c) = \alpha_{i,c,j}$, for $i = 1, 2, ..., n$, $j = 0, 1, \ldots, K-1$ and $c = 0, 1$. Given N observations $\{(X_{(\ell)}, y_{(\ell)})\}_{\ell=1}^{N}$, derive the MLE estimators of π and $\alpha_{i,c,j}$ under the Naïve Bayes assumption.

4. **[9pt]** Suppose each feature is real-valued, with $x^i \in \mathbb{R}$, and $P(x^i|y=c) \sim \mathcal{N}(\mu_{i,c}, 1)$ for $i = 1, 2, ..., n$ and $c = 0, 1$. Solve the following problems under the Naïve Bayes assumption.

   (a) **[3pt]** Given N observations $\{(X_{(\ell)}, y_{(\ell)})\}_{\ell=1}^{N}$, derive the MLE estimator of $\mu_{i,c}$.

(b) [**6pt**] Show that the decision boundary $\{(x^1, x^2, .., .x^n) : P(y = 0|x^1, x^2, .., .x^n) = P(y = 1|x^1, x^2, .., .x^n)\}$ is linear in $x^1, x^2, .., .x^n$.

5 Programming Exercise (20 points)

Note: Your code for all of the programming exercises including this one should be submitted to the corresponding Programming submission slot on Gradescope. Feel free to use any programming language, as long as your TAs can read your code. Turn in your code in a single .tar ball that might contain multiple source code files. While visualizations and written answers should still be submitted to Gradescope Written as a part of the rest of the homework. In your code, **please use comments to point out primary functions that compute the answers to each question.**

In this problem, you will implement the Naive Bayes (NB) algorithm on a pre-processed dataset that contains both **discrete** and **continuous** covariates. Recall from class that Naive Bayes classifiers assume the attributes $x^1, x^2, \ldots$ are conditionally independent of each other given the class label y, and that their prediction can be written as $\hat{y} = \text{argmax}_y P(y|X)$, where:

$$P(y|X = (x^1, \ldots, x^n)) \propto P(X, y) = P(X|y) \cdot P(y) = P(y) \cdot \prod_i P(x^i|y) \tag{2}$$

Consider the case where there are C classes, so that $y \in C$, and N different attributes.

- For a discrete attribute i that takes M_i different values, the distribution $P(x^i|y = c)$ can be modeled by parameters $\alpha_{i,c,1}, \alpha_{i,c,2}, \ldots, \alpha_{i,c,M_i}$, with $\sum_{j=1}^{M_i} \alpha_{i,c,j} = \sum_{j=1}^{M_i} P(x^i = j|y = c) = 1$.
  Important: Do NOT use smoothing. Assume $\log(0) = \lim_{x \to 0} \log x = -\infty$.
- For a continuous attribute i, **in this question**, we can assume the conditional distribution is Gaussian; i.e. $P(x^i|y = c) = \mathcal{N}(\mu_{i,c}, \sigma^2_{i,c}) \approx \frac{1}{\sqrt{2\pi(\sigma^2_{i,c}+\varepsilon)}} \exp(-\frac{(x^i-\mu_{i,c})^2}{2(\sigma^2_{i,c}+\varepsilon)})$, where $\mu_{i,c}$ and $\sigma^2_{i,c}$ are the mean and variance for attribute i given class c, respectively. In your implementation, you should estimate $\mu_{i,c}$ via the sample mean and $\sigma^2_{i,c}$ via the sample variance.
  Important: Meanwhile, take $\varepsilon = 10^{-9}$, which is a small value just to ensure the variance is not 0.

You now need to implement a Naive Bayes algorithm that predicts whether a person makes over \$50K a year, based on various attributes about this person (e.g., age, education, sex, etc.). You can find the detailed description of the attributes, and download the data at

https://archive.ics.uci.edu/ml/datasets/adult.

You will need 2 files:

- `adult.data`[3]: Each line is a training data sample, with attributes listed in the same order as on the website and delimited by commas. For instance, the first entry of each line is **age**. The last entry of each line gives the correct label (>50K, ≤50K). There should be 32,561 training data samples.

[3] https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

- `adult.test`[4]: Same format as `adult.data`, but only used in evaluation of the model (i.e. testing), so you shouldn't use the label for training your NB classifier. There should be 16,281 testing data samples.

Important: You should ignore (but do not delete the lines) all incomplete data lines, which contains "?" as values for certain attributes in the line.

Important: Because $P(y)\prod_i P(x^i|y)$ can get extremely small, you should use log-posterior for your computations:

$$\log\left[P(y)\prod_i P(x^i|y)\right] = \log P(y) + \sum_i \log P(x^i|y)$$

5.1 Report Parameters

For questions below, report only up to **4 significant digits** after the decimal points.

1. [**2 points**] Report the prior probability of each class.

   '≤50K':

   '>50K':

2. [**8 points**] For each class c and for each attribute i in [education-num, marital-status, race, capital-gain] print & report the following:

   - If the attribute is discrete, report the value of $\alpha_{i,c,j}$ for every possible value j, **in the same order as on the website** (e.g., for attribute "sex", you should report the α for "Female" first, then "Male"). Clearly mark what the attribute is and what is the value of j.
   - If the attribute is continuous, report the value of $\mu_{i,c}$ and $\sigma_{i,c}$.

   (The values given below for age and workclass are what is expected. You should use these values to check correctness of your programming):

   Class "$> 50K$":

   - age: mean=43.9591, var=105.4513
   - workclass: Private=0.64944, Self-emp-not-inc=0.0950, Self-emp-inc=0.0799, Federal-gov=0.0486, Local-gov=0.0811, State-gov=0.0458, Without-pay=0.0, Never-worked=0.0,

   Class "$<= 50K$":

   - age: mean=36.6081, var=181.2883
   - workclass: Private=0.7685, Self-emp-not-inc=0.0787, Self-emp-inc=0.0209, Federal-gov=0.0255, Local-gov=0.0643, State-gov=0.0412, Without-pay=0.0006, Never-worked=0.0,

[4] https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test

(a) Class "$> 50K$":

- education-num:

  Mean=

  Variance=

- marital-status:

  Married-civ-spouse=

  Divorced=

  Never-married=

  Separated=

  Widowed=

  Married-spouse-absent=

  Married-AF-spouse=

- race:

  White=

  Asian-Pac-Islander=

  Amer-Indian-Eskimo=

  Other=

  Black=

- capital-gain:

  Mean=

  Variance=

(b) Class "<= $50K$":

- education-num:

  Mean=

  Variance=

- marital-status:

  Married-civ-spouse=

  Divorced=

  Never-married=

  Separated=

  Widowed=

  Married-spouse-absent=

  Married-AF-spouse=

- race:

  White=

  Asian-Pac-Islander=

  Amer-Indian-Eskimo=

  Other=

  Black=

- capital-gain:

  Mean=

  Variance=

3. **[2 points]** Report the log-posterior values (i.e. $\log[P(X|y)P(y)]$) for the first 10 test data (in the same order as the data), each rounding to 4 decimal places (have 4 numbers after decimal points, for example, 12.3456). Ignore the lines which contain "?" and report the values with the corresponding line numbers.

5.2 Evaluation

1. **[1 point]** Evaluate the trained model on the training data. What is the training accuracy of your NB model? Round your answer to 4 decimal places.

2. **[1 point]** Evaluate the trained model on the testing data. What is the testing accuracy of your NB model? Round your answer to 4 decimal places.

3. [**6 points**] Instead of training the NB using all training data, train only with the first n data by following these steps:

   (a) Select the first n data points including lines with "?" and call this your training data.

   (b) Remove lines with "?" from your training data (so you have $n - n'$ rows where n' rows contain "?").

   (c) Train on the $n - n'$ data and test on the entire testing data.

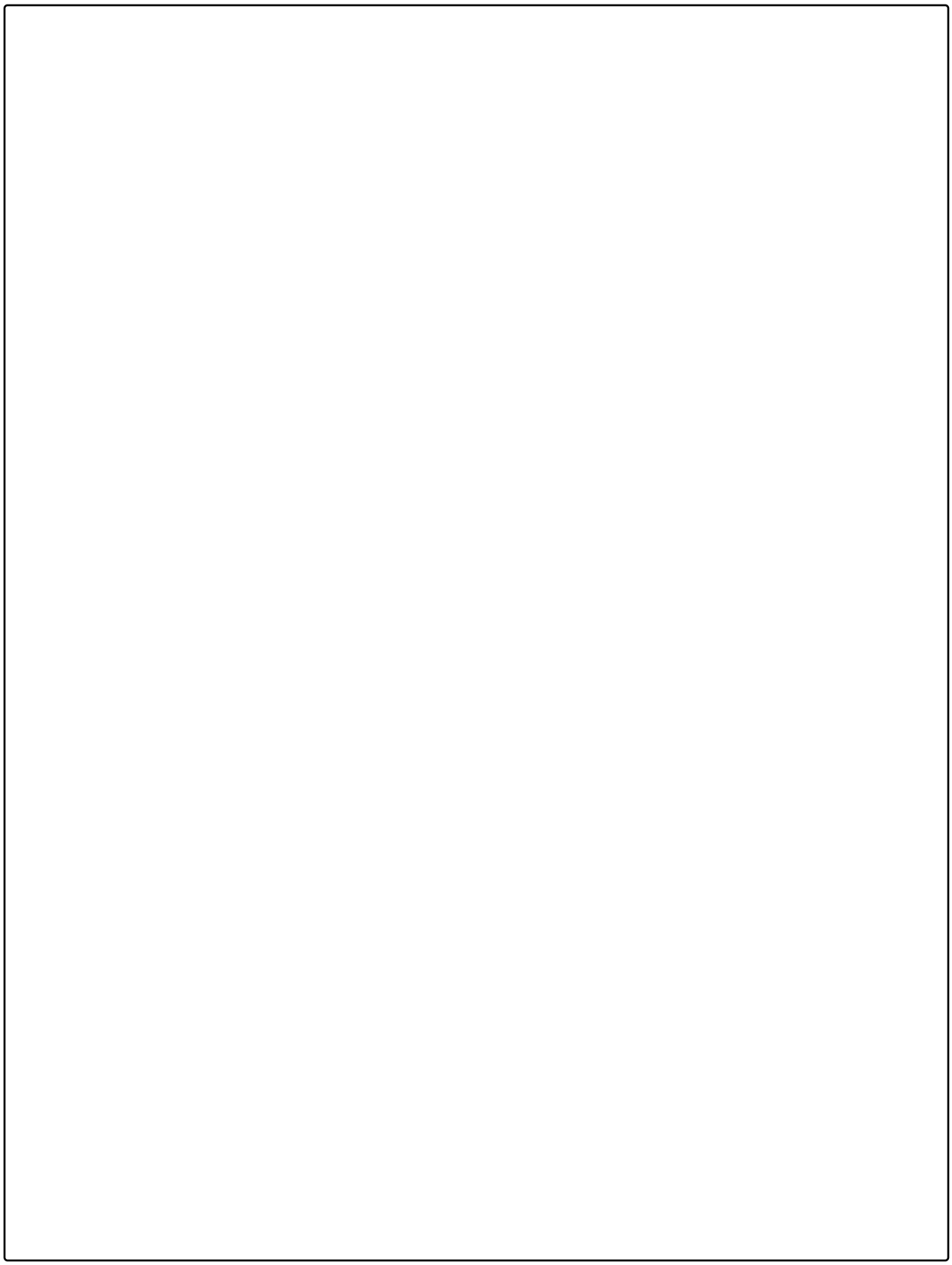
   (d) Repeat step (a) - (c) for $n = \{2^i \text{ for } i = 5, 6, 7, \ldots, 13\}$ (i.e. $n = 32, \ldots, 8192$)

   (e) Report training accuracy over the n samples and testing accuracy over all of the test data.

   (f) Plot training and testing accuracies calculated in (e) vs. # of training data.

   (**Important:** Use "$\leq$50K" as a label if $P_{leq} > P_{gr}$ else ">50K" to break ties.)

   What do you observe? At what values of n do testing accuracy and training accuracy attain their maximums, respectively? **In general**, what would you expect to happen if we use only a few (say $n < 3$) training data for Naive Bayes? Explain briefly (hint: we did not use smoothing). Please put your solutions the box on the next page.



6 Collaboration Questions

1. (a) Did you receive any help whatsoever from anyone in solving this assignment?

   (b) If you answered 'yes', give full details (e.g. "Jane Doe explained to me what is asked in Question 3.4")

2. (a) Did you give any help whatsoever to anyone in solving this assignment?

   (b) If you answered 'yes', give full details (e.g. "I pointed Joe Smith to section 2.3 since he didn't know how to proceed with Question 2")

3. (a) Did you find or come across code that implements any part of this assignment?

   (b) If you answered 'yes', give full details (book & page, URL & location within the page, etc.).