95-828 Machine Learning for Problem Solving: Homework 1

H. John Heinz III College
Carnegie Mellon University

- There are 7 questions on this assignment. Four *conceptual* (short answers) and three *applied* (implementation) that involve coding. For the applied questions, you may write your code either in an R Markdown file with a .rmd extension or in an IPython notebook with a .ipynb extension.
- Deliverables for R users: Submit a zip file to Canvas, which includes 1) a R markdown file and 2) a knitted PDF file, and submit a hard copy of the PDF file in class. In your submitted PDF file and the hard copy, do NOT include your codes unless otherwise stated in a question. To suppress code while creating the PDF submission, you can use `echo=False` in each code block in R.
- Deliverables for Python users: Submit a zip file to Canvas, which includes 1) an IPython notebook file and 2) an exported HTML file (File ⇒ Download as ⇒ HTML), and submit a printout of the HTML file in class. In your submitted HTML file and the hard copy, do NOT include your codes unless otherwise stated in a question. To suppress code while creating the report for submission, you can copy the code available here in a cell in your IPython notebook and executing it.
- The assignment is due at 4:30 PM (beginning of class) on **Feb 20, 2017**. If you are taking late day(s), please submit your homework (both write-up and codes) on Canvas to mark the time of submission and then submit a hard copy of write-up next time in class. Please note down the number of late days you used as well as the total number of late days remaining on top of the first page of your submission.
- Do not forget to put both your name and andrew ID on *every* page of your submission.
- If you have any questions about the HW, please use Piazza or come to office hours and recitations.
- You may *discuss* the questions with fellow students, *however* you must always write your own solutions and must acknowledge whom you discussed the question with. Do not copy from other sources, share your work with others, or search for solutions on the web. Plagiarism will be penalized according to university rules.

1 Conceptual: ML Problem Setup [8 points]

Online music and movie stores such as Spotify and Netflix have been developing rapidly, and more and more people choose to consume digital entertainment products through these channels. With availability of rich historical data, machine learning can be applied for many tasks and help the operations. For each of the 3 tasks below, specify what type of machine learning problem it is (supervised or unsupervised; and classification, regression, etc.). Explain your reasoning briefly in 1-2 sentences each.

a. (2 pts) Predicting the sales of an album.

b. (2 pts) Organizing the movies for better browsing experience.

c. (2 pts) Estimating the probability that a certain customer will watch a certain movie in January.

d. (2 pts) Estimating the probability that the online store contains albums of a given total duration.

2 Conceptual: Exploratory Data Analysis [20 points]

Answer each of the questions below. Your answer should be concise and consist of only a few phrases or sentences. If you would like to write a long paragraph, try to split it into bullet points, and limit your answer in each bullet point to no more than two sentences. Answers in long paragraphs (more than 100 words) will receive no credit.

a. (2 pts) Give an example (other than the ones shown in the lecture slides) for each of the following types of attributes: nominal, ordinal, interval, and ratio. Would it be suitable to calculate the following quantities for these attributes?: frequency distribution, median, mean, and coefficient of variation. Why (not)?

b. (2 pts) List two problems that might be caused by discarding data records with missing data.

c. (3 pts) Duplicate data:

   1. (1 pt) What is the main issue with doing data analytics using *exact-duplicate* records?
   2. (2 pts) Now instead imagine a retail store database (such as Macy's) with *near-duplicate* records where we have records with different names for a brand (e.g., 'Williams Sonoma' and 'Williams–Sonoma'). This is called the entity resolution problem. List one application which would be negatively affected by analysis on such a database. (Hint: database with *near-duplicate* would treat one entity as multiple separate entities.)

d. (2 pts) Describe the difference between a linear model, a generalized linear model (GLM), and a generalized additive model (GAM), and give an example equation for each of the models.

e. (4 pts) Given dataset $X = \{4, 5, 5, 6, 12, 14, 15, 15, 16, 17\}$, use Parzen windows to estimate the density $p(x)$ at $y = 3, 10, 15$. Use a boxcar kernel with $\Delta = 2$.

f. (7 pts) True or False. For the ones you answer 'False', briefly explain why it is false.

   1. For categorical attributes, more fine-grained data is always more desirable.
   2. Semantic accuracy is more difficult to check than syntactic accuracy.
   3. In kernel density estimation, the choice of the kernel bandwidth is often more important than the choice of the kernel function.
   4. In kernel density estimation, a bin width that is too large will under-smooth the estimate and yield a spiky estimate.
   5. If Spearman's rho of two variables X and Y equals to 1, then X and Y have perfect positive linear correlation.
   6. $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + X_3^{\beta_3}$, where y is a numerical outcome, Xs are features and βs are parameters, is a linear model.
   7. Principal Component Analysis can be performed using one of two different objectives: Maximum variance subspace, and minimum reconstruction error. The solutions to these two objectives are similar in most cases, but they are not identical.

3 Conceptual: Linear Regression [10 points]

What did housing prices look like in the 'good old days'? The median sales prices for new single-family houses are given in the accompanying table for the years 1972 through 1979.[1] Letting y denote the median sales price and x the year (using integers 1, 2, . . . , 8)

Year	Median Sales Price ($\times$1000)
1972(1)	\$ 27.6
1973(2)	\$ 32.5
1974(3)	\$ 35.9
1975(3)	\$ 39.3
1976(5)	\$ 44.2
1977(6)	\$ 48.8
1978(7)	\$ 55.7
1979(8)	\$ 62.9

a. (2 pts) Fit the linear model $Y = \beta_0 + \beta_1 x + \epsilon$: Report the coefficient estimates β_0 and β_1. (Do not use simple linear regression command in statistical software for this question. Do it by hand and show your calculations. You will be asked to implement linear regression using statistical software in Problem 6.)

b. (1 pts) Based on the model you fit in part a, what is your estimated value of the median sales price for year 1980?

c. (2 pts) Calculate the RSS (Residual Sum of Squares) and estimate the residual variance σ^2.

d. (1 pts) Estimate the standard error for $\hat{\beta}_1$.

e. (1 pts) Is there sufficient evidence to indicate that the median sales price for new single-family houses increased over the period from 1972 through 1979 at the 0.01 level of significance?

f. (1 pts) Give a 95% confidence interval for β_1.

g. (1 pts) Which of the following have closed-form analytical solutions?:

(I) Unregularized Linear Regression,
(II) Ridge Regression (i.e. linear regression with l_2 regularization),
(III) Lasso Regression (i.e. linear regression with l_1 regularization)

h. (1 pts) Consider a dataset which has millions of features. Describe two challenges one would face in calculating the closed-form multiple linear regression solution $\beta^* = (X^T X)^{-1} X^T y$. (Hint: Consider the time and space complexities.)

4 Conceptual: Naive Bayes [20 points]

4.1 MLE for Naive Bayes [12 pts]

One of the most common applications of machine learning is classification i.e. separating instances into various classes. A very popular special case is binary classification i.e. classifying data into *two* classes. In this problem, we consider the problem of classifying text articles into two classes. Specifically, given a collection of documents, each coming from either the serious European magazine *The Economist*, or from the not-so-serious American magazine *The Onion*, our goal is to learn a classifier that can distinguish between articles from each magazine.

[1] Source: Adapted from *Time*, 23 July 1979, p. 67.

The set of all words in the data is called the *vocabulary* and we let V be the number of words in the vocabulary. For each article, we produced a feature vector $X = \langle X_1, \ldots, X_V \rangle$, where X_i is equal to 1 if the i^{th} word appears in the article and 0 otherwise. Each article is also accompanied by a class label of either 1 for The Economist or 2 for The Onion.

When we apply the Naive Bayes classification algorithm, we make two assumptions about the data: first, we assume that our data is drawn i.i.d. from a joint probability distribution over the possible feature vectors X and the corresponding class labels Y; second, we assume for each pair of features X_i and X_j with $i \neq j$ that X_i is conditionally independent of X_j given the class label Y (recall that this is the Naive Bayes assumption). Under these assumptions, a natural classification rule is as follows: Given a new input X, predict the most probable class label $\hat{Y}$ given X. Formally,

$$\hat{Y} = \operatorname*{argmax}_y P(Y = y|X).$$

a. (3 pts) Prove that the classification objective can be rewritten as

$$\hat{Y} = \operatorname*{argmax}_y \left(\prod_{w=1}^{V} P(X_w|Y = y) \right) P(Y = y).$$

b. (3 pts) How many parameters are needed to represent the distribution $P(X|Y = y)$ when using the Naive Bayes assumption? How many are needed if we do not use the Naive Bayes assumption? Based on this difference, in which cases is there a big gain from making this assumption?

Of course, since we don't know the true joint distribution over feature vectors X and class labels Y, we need to estimate the probabilities $P(X|Y = y)$ and $P(Y = y)$ from the training data. For each word index $w \in \{1, \ldots, V\}$ and class label $y \in \{1, 2\}$, the distribution of X_w given $Y = y$ is a Bernoulli distribution with parameter θ_{yw}. In other words, there is some unknown value θ_{yw} such that

$$P(X_w = 1|Y = y) = \theta_{yw} \quad \text{and} \quad P(X_w = 0|Y = y) = 1 - \theta_{yw}.$$

c. (3 pts) Derive the MLE estimate of θ_{yw}. Please do not provide just the final answer but derive the maximum liklihood estimate by writing out and maximizing the log-likelihood.

Similarly, the distribution of Y (when we consider it alone) is a Bernoulli distribution (except taking values 1 and 2 instead of 0 and 1) with parameter ρ. In other words, there is some unknown ρ such that

$$P(Y = 1) = \rho \quad \text{and} \quad P(Y = 2) = 1 - \rho.$$

d. (3 pts) Derive the MLE estimate of ρ. Please do not provide just the final answer but derive the maximum liklihood estimate by writing out and maximizing the log-likelihood.

4.2 Naive Bayes Example Calculations [8 pts]

Consider the following data. It has 4 features $\mathbf{X} = (X_1, X_2, X_3, X_4)$ and 3 labels $(+1, 0, -1)$. Assume that the probabilities $p(\mathbf{X} = (X_1, X_2, X_3, X_4)|Y = y)$ and $p(Y = y)$ are both Bernoulli distributions. Answer the questions that follow under the Naïve Bayes assumption.

a. (1 pt) Compute the MLE for the prior probabilities $p(Y = +1), p(Y = 0), p(Y = -1)$.

b. (4 pts) Compute the Maximum Likelihood Estimate (MLE) for $p(X_i = 1|Y = y), \forall i \in [1, 4]$ and $\forall y \in \{+1, 0, -1\}$. Use the table below to populate corresponding values.

X_1	X_2	X_3	X_4	Y
1	1	0	1	+1
0	1	1	0	+1
1	0	1	1	0
0	1	1	1	0
0	1	0	0	-1
1	0	0	1	-1
0	0	1	1	-1

	$Y = +1$	$Y = 0$	$Y = -1$
$X_1 = 1$			
$X_2 = 1$			
$X_3 = 1$			
$X_4 = 1$			

c. (3 pts) Use the values computed in the above two parts to classify the data point ($X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1$) as either belonging to class $+1, 0$ or -1.

5 Applied: Exploratory Data Analysis and KDE [16 points]

Employee turnover is a key problem faced by many organizations. When good people leave, it usually costs the organization substantial time and other resources to find a replacement. Therefore, many organizations try to keep the churn rate at a low level. Imagine a company who now wants to understand its employee churn situation. Its HR (Human Resources) department gives you some data of their employees, and asks you to do exploratory data analysis and to predict employee churn.

You are free to choose any statistics software to analyze the data. In your answer, please include both the snippets of your code as well as the outputs.

Download the data `'termination.csv'` from Canvas and load it into R (or any other software). Use it to answer the following questions:

a. (2 pts) Display a summary of the data (i.e. min, max, mean and quartiles for each variable). In the summary statics, is there any meaningless quantity?

b. (2 pts) The data include 10 years (2006 - 2015) of records for both active and terminated employees. `Status Year` field shows the year of data, and `Status` field shows the employment status – ACTIVE or TERMINATED in the corresponding status year. The company is interested in what proportion of the staff are leaving. Compute: 1) the percent of terminated employees out of all employees for each year; 2) average termination rate over the 10 years?

c. (3 pts) In addition to the proportion of terminated employees, the company wants to know more about different types of termination. Give a stacked bar chart (see for e.g. http://ggplot2.tidyverse.org/reference/geom_bar.html) of terminates, where x-axis is status year, y-axis is number of terminated employees, and different colors in a bar show different termination reasons ('termreason_desc' field in the data). What do you observe in this plot?

d. (2 pts) Does `Age` affect termination? Draw (2) Box-plots of `Age` for active and terminated employees separately. What does the box-plot tell you?

e. (7 pts) Does `Length of Service` affect termination? Plot Kernel Density Estimation of `Length of Service` for active and terminated employees separately in one plot (use separate colors) with the following steps and **print your codes for this part**:

A (1 pt) Uniformly sample points between [0, 30] with interval 0.1.

B Write a kernel density estimation function that takes x (a point as sampled above) as input:

(a) (1 pt) Calculate the distance from x to each of the **observed** `Length of Service` for *active* employees.

(b) (1 pt) Feed the differences into the standard Gaussian kernel (with mean 0, and standard deviation 1) to obtain density estimate from each kernel.

(c) (1 pt) Sum the density estimate from all the kernels and normalize it by the number of observed data. Return the estimate.

C (1 pt) Compute the kernel density estimate for each of the sampled points in A and plot the final density estimate.

D (1 pt) Repeat step B and C, this time for *terminated* employees. Plot the two density estimates (one for active and the other for terminated employees) on the same graph for easy comparison. (1pt) What can you deduce from the plot? State briefly.

6 Applied: Linear Regression [8 points]

Labor economists are generally interested in how wages are correlated with other factors. In this question, you will explore the relationship of wage and several covariates, and use linear models to do prediction and classification.

Download the data `'Wage.csv'` from Canvas and load it. If you use R, you can load the data using the commands

```
library(ISLR); attach(Wage)
```

You may want to first take a look at the data by summarizing it (i.e. through summary statistics).

Hint: Take a look at `glmnet` package in R or `statsmodels` package in python, and use `summary` function to show regression results.

a. (2 pts) Regress `Wage` on `Age`. Show the regression result and draw a scatter plot of the data with the fitted line).

b. (1 pts) Regress `Wage` on `Age`, `Jobclass` as well as their interaction. How would you interpret the result?

c. (2 pts) Regress `Wage` on a fourth-degree polynomial in `Age`. Again, show the regression result and draw a scatter plot of the data with the fitted line). What can you learn from the result?

d. (1 pts) Now regress `Wage` on all available variables (`year, age, maritl, race, education, jobclass, health, health_ins`; note that there is no variation in sex and region, therefore the two attributes are not included) and all the 2-way interactions. Show the regression result.

e. (2 pts) The model estimated above has many dependent variables, and may lead to the *over-fitting* problem. To avoid overfitting, run Lasso regression to enforce regularization. Use $\lambda = 0.04$ in the Lasso regression. (Note that in general we will pick this hyperparameter via cross-validation, to be discussed in class soon, but for now we provide you with a suitable λ to use.) Compare the coefficients from the Lasso regression to those from the linear regression in part e) by showing a scatter plot of points (x,

y) where x is coefficient from the (unregularized) linear regression and y is corresponding coefficient from the Lasso regression.

7 Applied: Naive Bayes [18 points]

In this question you will implement a Naive Bayes classifier for a text classification problem. You will be given a collection of text articles, each coming from either the serious European magazine *The Economist*, or from the not-so-serious American magazine *The Onion*. The goal is to learn a classifier that can distinguish between articles from each magazine.

Download the data `'Q7_naive_bayes_data.zip'` from Canvas. We have pre-processed the articles so that they are easier to use in your experiments. We extracted the set of all words that occur in any of the articles, known as the vocabulary. The dataset is provided in the form of five files which can be loaded using appropriate commands in your programming language to yield the following variables:

- `Vocabulary` is a V dimensional list that that contains every word appearing in the documents. When we refer to the j^{th} word, we mean `Vocabulary(j)`.
- `XTrain` is a $n \times V$ dimensional matrix describing the n documents used for training your Naive Bayes classifier. The entry `XTrain(i,j)` is 1 if word j appears in the i^{th} training document and 0 otherwise.
- `yTrain` is a $n \times 1$ dimensional matrix containing the class labels for the training documents. `yTrain(i,1)` is 1 if the i^{th} document belongs to The Economist and 2 if it belongs to The Onion.
- Finally, `XTest` and `yTest` are the same as `XTrain` and `yTrain`, except instead of having n rows, they have m rows. This is the data you will test your classifier on and it should not be used for training.

You may want to first take a look at the data by summarizing it (i.e. through summary statistics). Provide answers to the following questions in your submitted report:

a. (5 pts) Fit a Naive Bayes classifier to the provided data (with no Laplacian smoothing). Produce a confusion matrix for the Naive Bayes classifier. Plot the matrix as a heatmap. (In R, you may import the `klaR` package and use the `NaiveBayes()` command with `usekernel = TRUE`. In Python, you may use the `scikit-learn` package implementation of Naive Bayes.)

b. (4 pts) Calculate and report the precision and recall considering the articles from *The Onion* as the positive class. (Note that precision is given as $precision = \frac{tp}{tp+fp}$ and recall is given as $recall = \frac{tp}{tp+fn}$ where tp=true positives, tn=true negatives, fp=false positives, and fn=false negatives.)

c. (2 pts) Calculate and report the precision and recall considering the articles from *The Economist* as the positive class.

d. (1 pts) What is the misclassification rate of Naive Bayes on this problem?

e. (3 pts) What is the misclassification rate of Naive Bayes when you use a Laplacian smoothing of 1.0?

f. (3 pts) What is the true class of the 45^{th} observation? What is its predicted class? What are the estimated posterior probabilities for the 45^{th} observation according to Naive Bayes?