1.

The Federalist papers,[1] authored by Alexander Hamilton, John Jay, and James Madison, consist a series of 85 papers published between October 1787 and April 1788 under the pseudonym PUBLIUS to convince the people of New York to ratify the US constitution. The authorship of some of the papers is in dispute. In particular, the authorship of 12 of the papers is disputed, while Hamilton and Madison later published their lists of authors of the rest, although even those lists have discrepancies. One can use Machine Learning algorithms to classify the disputed papers using papers with known authors. Later in this course, we will learn how one can convert text to a vector of numerical features, but you do not need that to solve this problem. (15 pts)

(a) How many data points does the training set contain? How many data points are in the test set?

(b) Formulate the above problem as a multiclass problem. How many classes does ther outcome Y have? What are those classes?

(c) Some experts believe that some of the papers are collaborative efforts of two or sometimes all three of Hamilton, Jay, and Madison. Capturing multi-author papers can be done by formulating the problem as a multi-label problem. Explain what the labels are, if the labels are binary, and how the algorithm should label a paper that was solely written by Jay, a paper that was written by Hamilton and Madison, and a papert that was the collaborative work of Hamilton, Jay, and Madison.

2.

In a weird simulated world, we have three types of creatures. Each creature has between 1 to 100 legs, 1 to 100 teeth, and 1 to 100 noses. The fraction of creatures type-1, type-2, and type-3 are respectively $l/(l+t+n)$, $t/(l+t+n)$, and $n/(l+t+n)$, where l, t, n are respectively the number of legs, teeth, and noses of a creature.

(a) If a creature has 10 teeth, 25 legs, and 30 noses, what is your best guess about the type of the creature?

(b) What type of supervised learning problem are you solving in this question? Explain.

3.

In this problem, we show that logistic regression for binary classification can be formulated using the following parametric form for the class conditional probability, when the labels for the negative and the positive class are respectively -1 and 1, i.e. $Y \in \{-1, 1\}$:

$$\Pr(Y = y^{(i)} | X = x^{(i)}) = \frac{1}{1 + \exp(-y^{(i)}[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p])} \quad (1)$$

where $x^{(i)} = (x_1, x_2, \ldots, x_p)$ is the vector of p features of the i^{th} sample and $y^{(i)} \in \{-1, 1\}$ is its label, and $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters to be found using maximum likelihood estimation.

Show that Eq. (1) is a valid alternative formulation of logistic regression by (30 pts)

(a) Showing that $\Pr(Y = 1 | X = x^{(i)}) + \Pr(Y = -1 | X = x^{(i)}) = 1$.

(b) Showing that the form $\Pr(Y = y^{(i)} | X = x^{(i)})$ in Eq. (1) (which is obviously between 0 and 1) can approach 0 and 1. Hint: Assume $z^{(i)} = -y^{(i)}[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p])$. What happens when $z^{(i)} \to +\infty$ or $z^{(i)} \to -\infty$.

(c) Finding the decision boundary between classes $Y = -1$ and $Y = 1$ as a function of $(x_1, x_2, \ldots, x_p)$. For what values of $z^{(i)}$ a new test data point is classified in positive class $Y = 1$?

4.

Consider a logistic regression problem in which there are no features, which means that:

$$\Pr(Y = 1) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

Assume that we have m data points with label $Y = 1$ and n data points with label $Y = 0$ (remember that features are irrelevant).

(a) Write down the likelihood function $l(\beta_0)$.

(b) Find the Maximum Likelihood estimate $\hat{\beta}_0$ for this data set. [**Hint**: maximize $\log_e l(\beta_0)$].

(c) Determine conditions under which this simple classifier classifies data points into $Y = 1$ or $Y = 0$.

5.

Consider the novel logistic regression method for binary classification with two features $\mathbf{X} = (X_1, X_2)$, formulated by

$$p(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_1^2+\beta_3 X_2}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_1^2+\beta_3 X_2}}$$

Assume that using a data set of 200 observations from both classes, we obtained the following results:

	Coefficient	Standard Error
β_0	- 1	0.2
β_1	2	1
β_2	-1	0.02
β_3	1	0.6

(a) Determine the equation for the decision boundary for this classifier.

(b) Sketch the decision boundary for this classifier and clearly show the regions for the positive and the negative class.

(c) Because the coefficients are calculated using Maximum Likelihood Estimation from a large enough dataset, one can assume that $Z = \frac{\widehat{\beta}_i - \beta_i}{\mathrm{SE}(\widehat{\beta}_i)}$ is standard normal. Using this fact, determine which of the predictors can be eliminated from the model. You can consider the significance level to be $\alpha = 0.05$.

6.

Consider multinomial regression for multiclass classification with three features $\mathbf{X} = (X_1, X_2, X_3)$, formulated by

$$p_k(\mathbf{X}) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\beta_{2k}X_2+\beta_{3k}X_3}}{e^{\beta_{01}+\beta_{11}X_1+\beta_{21}X_2+\beta_{31}X_3} + e^{\beta_{02}+\beta_{12}X_1+\beta_{22}X_2+\beta_{32}X_3} + e^{\beta_{03}+\beta_{13}X_1+\beta_{23}X_2+\beta_{33}X_3}},$$
$$k \in \{1, 2, 3\}$$

Assume that using a data set of 498 observations from three classes, we obtained the following results:

Coefficient	Value	Standard Error
β_{01}	1	
β_{11}	-2	s_1
β_{21}	-1	s_2
β_{31}	1.5	s_3
β_{02}	0	
β_{12}	0	
β_{22}	-2.5	s_4
β_{32}	0	
β_{03}	0	
β_{13}	0	
β_{23}	0	
β_{33}	2	s_5

maximum

(a) Determine the value for standard errors s_1, s_2, s_3, s_4, s_5 so that their corresponding coefficients are statistically significant at level $\alpha = 0.05$. Assume all other coefficients are statistically significant.

(b) In what class will the classifier classify $\mathbf{X}^* = (0, 0, -1)$?

(c) Find the equation of the decision boundary between classes 1,2 and the decision boundary between classes 1,3 and the decision boundary between classes 2,3.

7.

In an unusual logistic regression problem,

$$\Pr(Y = 1|X_1, X_2) = \frac{X_1^2 X_2^2}{1 + X_1^2 X_2^2}$$

log

(a) Write down the model for the odds of $Y = 1$ given X_1 and X_2.

(b) Determine the equation for the decision boundary between classes $Y = 1$ and $Y = 0$ and sketch it in the X_1, X_2 plane. Show the regions for each class.

8.

Consider the following dataset:

Index	X_1	X_2	Y
1	0	0	1
2	1	0	2
3	0	1	3
4	1	1	2

We wish to fit a multinomial regression model to this dataset.

(a) Write down the likelihood function for this dataset. Use one parameter vector $(\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}), k = 1, 2, 3$ for each class.

(b) Does this problem have a unique set of β's as its solution?

9.

Consider multinomial regression for multiclass classification with three features $\mathbf{X} = (X_1, X_2, X_3)$, formulated by

$$p_k(\mathbf{X}) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\beta_{2k}X_2+\beta_{3k}X_3}}{1 + e^{\beta_{01}+\beta_{11}X_1+\beta_{21}X_2+\beta_{31}X_3} + e^{\beta_{02}+\beta_{12}X_1+\beta_{22}X_2+\beta_{32}X_3}}, \quad k \in \{1, 2\}$$

where the classes are determined by $k \in \{1, 2, 3\}$

Assume that using a data set of 210 observations from three classes, we obtained the following results:

Coefficient	Value
β_{01}	1
β_{11}	-2
β_{21}	-1
β_{31}	1
β_{02}	0
β_{12}	0
β_{22}	1
β_{32}	1

Assume that the coefficients are all statistically significant.

(a) In what class will the classifier classify $\mathbf{X}^* = (1, 0, -1)$?

(b) Explain why despite having three classes, we formulated multinomial regression using ONLY TWO sets of parameters, $(\beta_{01}, \beta_{11}, \beta_{21}, \beta_{31})$ and $(\beta_{02}, \beta_{12}, \beta_{22}, \beta_{32})$, specific to classes 1 and 2, respectively?

10. The Bayes' Rule: The Farmer and The Librarian

11.

(Iterative Bayesian Inference) Hunter says she is itchy. There is a test for Allergy to Cats, but this test is not always right:

For people that really do have the allergy, the test says "Yes" 80% of the time

For people that do not have the allergy, the test says "Yes" 10% of the time ("false positive")

If 10% of the population have the allergy, and Hunter's test says "Yes". The doctor calculates the probability that she really has allergy. But the doctor decides to make sure that the diagnosis is right, so the test is done again, and it says yes again. The doctor uses the probability of Hunter having allergy from the last test as prior probability, and calculates the probability of Hunter having Allergy, given the test says "Yes." What is the doctor's diagnosis?

12.

You are studying the behavior of Pokémon trainers catching Pikachu, and would like to create a model that predicts whether or not a trainer is able to catch a Pikachu based only on the length of time they spend out in the tall grass. You have recorded the time it takes for each trainer to either catch a Pikachu in the tall grass, or leave the tall grass without catching one. A Pokémon trainer is categorized into one of two classes: Catch or Miss. Assume that the number of hours is normally distributed for Pokémon trainers who are able to catch a Pikachu and those who are not. The recorded times (in hours) are as follows:

- **Catch** (Class 1): [1, 3, 5]
- **Miss** (Class 2): [6, 8]

[By Emily Chen]

(a) Calculate the class means $\hat{\mu}_1$ and $\hat{\mu}_2$ and class variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ used by LDA. Express your answer in fraction form and simplify your answer as much as possible.

(b) Calculate the combined variance $\hat{\sigma}^2$ (pooled variance) used by LDA. Express your answer in fraction form and simplify your answer as much as possible.

(c) Calculate the LDA decision boundary. Express your answer in fraction form and simplify your answer as much as possible. Do not simplify logarithms.

13.

Consider the following multiclass dataset:

Index	X_1	X_2	Y
1	0	-1	1
2	-2	0	2
3	-2	3	2
4	12	1	3
5	-5	6	2
6	1	-9	1
7	19	-10	1
8	0	15	3
9	12	-4	1

For this dataset, assume that X_1 and X_2 are normally distributed in each class and are independent in each class (i.e., the Naïve Bayes' assumption).

(a) Calculate the linear discriminant scores $\delta_k(x_1, x_2)$ for $k = 1, 2, 3$, assuming that $\pi_1 = \pi_2 = \pi_3$. Assume each feature has the same variance in all classes. Simplify them as much as you can.

(b) Determine to which class the data point $(x_1, x_2) = (0, 1)$ will be classified

Remember that for this problem, you DO NOT need formulas for multivariable normal distributions and covariance matrices.

14.

Assume that in a binary classification problem with one feature X, the distribution of X in class $k = 1$ is

$$f_1(x) = \frac{1}{2} \exp\left(-\frac{x}{2}\right), x \geq 0$$

and the distribution of X in class $k = 2$ is

$$f_2(x) = \frac{1}{4} x \exp\left(-\frac{x}{2}\right), x \geq 0$$

(a) Derive the discriminant functions $\delta_1(x)$ and $\delta_2(x)$ assuming the prior class probabilities satisfy $\pi_1 = \pi_2$.

(b) Find the decision boundary between the two classes and determine to what class $x = 3$ is classified.

15.

In a classification problem with two classes and two features, the joint distribution of the features in each class is:

$$f_k(x_1, x_2) = \frac{1}{2\pi\sqrt{(1-k/4)}} \exp\left[-\frac{z}{2(1-k/4)}\right], \quad k = 1, 2$$

where

$$z = (x_1 - k)^2 - \sqrt{k}(x_1 - k)(x_2 - k^2) + (x_2 - k^2)^2$$

(a) Assuming that the prior probability of class one is twice the prior probability of class two, in what class is the point $(X_1, X_2) = (1, 5)$ is classified?

(b) The marginal distributions of features in each class can be calculated from the joint distributions, and are:

$$f_k(x_1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x_1 - k)^2}{2}\right]$$
$$f_k(x_2) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x_2 - k^2)^2}{2}\right]$$

The Naïve Bayes assumption clearly does not hold in this problem. However, classify $(X_1, X_2) = (1, 5)$ pretending the Naïve Bayes assumption holds and compare the results with part 2a.

16.

Assume that in a classification problem with one feature, the conditional distribution of the feature in class $k \in \{1, 2, 3\}$ is $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2})$, where $\mu_k = \sigma_k = k, k \in \{1, 2, 3\}$. Also, assume that the prior probabilities of each class are $\pi_k = \frac{1}{3}, k \in \{1, 2, 3\}$. Build a multinomial regression model that yields the same decision boundaries as the Bayesian Quadratic Discriminant Analysis method for this three-class problem; that is, show $\Pr(Y = k|X = x)$ for all $k \in \{1, 2, 3\}$. Note that we want each class to have a vector of parameters associated with it, i.e. three parameter vectors. Hence, the posterior probabilities will be in the form:

$$\Pr(Y = k|X = x) = \frac{e^{g_k(x)}}{e^{g_1(x)} + e^{g_2(x)} + e^{g_3(x)}}$$

17.

The following dataset was collected to classify people who evade taxes:

Tax ID	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	122 K	No
2	No	Married	77 K	No
3	No	Married	106 K	No
4	No	Single	88 K	Yes
5	Yes	Divorced	210 K	No
6	No	Single	72 K	No
7	Yes	Married	117 K	No
8	No	Married	60 K	No
9	No	Divorced	90 K	Yes
10	No	Single	85 K	Yes

Considering relevant features in the table (only one feature is not relevant), assume that the features are *conditionally independent*. Determine to what class the Naïve Bayes' classifier assigns the test point (Yes, Married, 70K). Assume Gaussianity for continuous features and use Laplace correction for discrete features.

18.

Assume that in a binary classification problem with one feature X, the distribution of X in class $k = 1$ is

$$f_1(x) = \frac{x}{\sigma_1^2} \exp\left(\frac{-x^2}{2\sigma_1^2}\right), x \geq 0$$

and the distribution of X in class $k = 2$ is

$$f_2(x) = \frac{1}{x\sqrt{2\pi}\sigma_2} \exp\left(\frac{-(\ln x - \mu_2)^2}{2\sigma_2^2}\right), x \geq 0$$

(a) Are there any conditions under which the discriminant function is a linear function of x?

(b) If $\sigma_1 = \sigma_2 = 1$, $\mu_2 = 10$, and $\pi_1 = \pi_2 = 0.5$, in what class will $x = 10$ be classified?

19.

In a four-class classification problem, performing multinomial regression on a dataset resulted in the following models

$$\log\left(\frac{\Pr(Y=1|X_1,X_2)}{\Pr(Y=4|X_1,X_2)}\right) = X_1 - X_2 + 1$$
$$\log\left(\frac{\Pr(Y=2|X_1,X_2)}{\Pr(Y=4|X_1,X_2)}\right) = X_1 + X_2 - 1$$
$$\log\left(\frac{\Pr(Y=3|X_1,X_2)}{\Pr(Y=4|X_1,X_2)}\right) = X_1 - X_2 - 1$$

Assume that we modeled $p_4(X_1,X_2) = \Pr(Y=4|X_1,X_2) = \frac{1}{1+e^{\delta_1}+e^{\delta_2}+e^{\delta_3}}$, where $\delta_1, \delta_2, \delta_3$ are linear functions of X_1 and X_2.

(a) Determine $p_k(X_1,X_2) = \Pr(Y=k|X_1,X_2)$ for $k = 1,2,3,4$. Remember that p_k is a logistic function of X_1, X_2. Your answer must have numeric coefficients for X_1, X_2.

(b) To which class $(x_1,x_2) = (1,0)$ is classified?

20.

Consider the following logistic regression problem with two classes $Y = 1$ and $Y = 2$ and one feature:

$$\Pr(Y = 1|X = x) = \frac{e^{x-.5+\log .5}}{e^{x-.5+\log .5} + e^{2x-2+\log .5}}$$
$$\Pr(Y = 2|X = x) = \frac{e^{2x-2+\log .5}}{e^{x-.5+\log .5} + e^{2x-2+\log .5}}$$

Find an LDA classifier that yields the same decision boundary as the above logistic regression problem. Assume that $\pi_1 = \pi_2 = 0.5$. It is sufficient to find the mean of the feature in each class (i.e. μ_1 and μ_2) and the common variance of the feature in those two classes, $\sigma_1 = \sigma_2 = \sigma$.

21.

Consider the following text classification problem:

Document1: This movie is sad. It made me cry. (Negative class)
Document 2: Cry cry cry! What a sad sad movie. (Negative class)
Document 3: I love it! (Positive class)
Document 4: Love this movie. Love everything about it. (Positive class)

Dropping the stop words, the combined document (dictionary) for the above corpus has the following words: movie, sad, made, cry, love.

(a) Create TF (term frequency) features for each of the documents. Do NOT use IDF (Inverse Document Frequency).

(b) Apply 1-nearest neighbor classifier with Euclidean to this dataset to classify the document "Love this sad movie." Remember that ''this" is a stop word and should not be used to calculate the features for this test document.