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 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. $Y \in \{-1, 1\}$:

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

   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 $Y = 1$?

[1] https://www.gutenberg.org/files/1404/1404-h/1404-h.htm

3. Although in practice, we usually model conditional distributions of features in each class as Gauusians for Bayesian Discriminant Analysis, one must notice that Bayesian Discriminant Analysis is doable using any type of distribution. Assume that we wish to perform Discriminant Analysis with only one positive predictor $x \geq 0$ (i.e. $p = 1$), but instead of normal class pdfs, we have Lognormal class pdfs whose formula is: (40 pts)

$$f_k(x) = \frac{1}{x\sqrt{2\pi}\sigma_k} \exp\left(\frac{-(\ln x - \mu_k)^2}{2\sigma_k^2}\right), x \geq 0$$

We know that the posterior class probability for class k is:

$$\Pr(Y = k|X = x) = p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

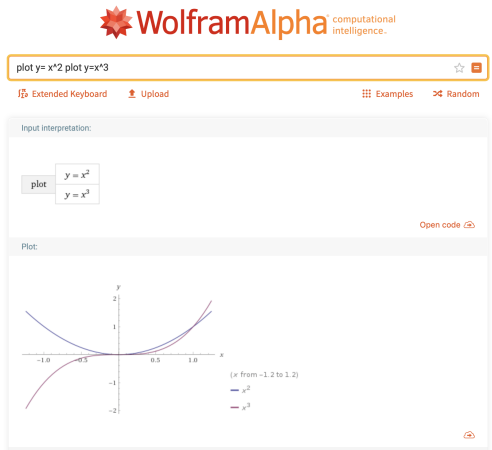
where π_l's are prior class probabilities.

(a) Determine the simplest form for the discriminant score $\delta_k(x)$ if we assume that $\sigma_1 = \sigma_2 = \ldots = \sigma_K$; that is, there is a shared parameter σ_k across all K classes. Are the discriminant scores linear functions of x in this case? Hint: follow the same procedure as the one we used for Gaussian pdfs.

(b) For binay classification ($K = 2$), when classes are balanced ($\pi_1 = \pi_2$), show that the decision boundary between the two classes is the threshold

$$x^* = \sqrt{\exp(\mu_1)}\sqrt{\exp(\mu_2)}$$

(c) Plot the pdfs and the decision boundary threshold when $K = 2, \pi_1 = \pi_2$, and $\mu_1 = 1, \mu_2 = 5$, and $\sigma_1 = \sigma_2 = 1$ over the range $x \in [0, 70]$. The easiest way is to use https://www.wolframalpha.com. A sample snippet is included below.



4. The following data set was collected to classify people who evade taxes: (40 pts)

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*.

(a) Estimate prior class probabilities, π_1 and π_2.

(b) For continuous feature(s), assume conditional Gaussianity and estimate class conditional pdfs $f_k(x)$.

(c) For each discrete feature X, assume that the number of instances in class k for which $X = x_j$ is n_{jk} and the number of instances in class k is n_k. Estimate the probability mass $f_k(x_j) = \Pr(X = x_j|Y = k)$ as n_{jk}/n_k for each discrete feature. Is this a valid estimate of the pmf? (a valid pmf sums to one!)

(d) There is an issue with using the estimate you calculated in 4c. Explain why the Laplace correction $(n_{jk} + 1)/(n_k + l)$, where l is the number of levels X can assume,[2] solves the problem with the estimate given in 4c. Is this a valid estimate of the pmf?

(e) Using the priors, pdfs, and pmfs you calculated in (4a), (4b) and (4d) respectively, determine how the algorithm classifies test data points (Yes, Single, 110K) and (No, Divorced, 93K).[3]

5. The test confusion matrix for a medical data set on HIV is shown below

	Actually Neg	Actually Pos
Predicted Neg	8826	30
Predicted Pos	23	456

Calculate Precision, Recall, True Negative Rate, Negative Predictive Value, and F1 score for the test data. (15 pts)

[2]For example, if $X \in \{apple, orange, pear, peach, blueberry\}$, then $l = 5$.

[3]Ask yourself: does this resemble KNN?