Introduction to Machine Learning **Name (Print):** ____________________
Fall 2020
Exam 2
13/01/21
Time Limit: 90 Minutes

This exam contains 5 pages (including this cover page) and 4 problems. Check to see if any pages are missing. Enter all requested information on the top of this page, and put your initials on the top of every page, in case the pages become separated.

You may *not* use your books, notes, internet sources, or any other storage method besides your memory and biological brain on this exam.

You are required to show your work on each problem on this exam. The following rules apply:

- **Organize your work**, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.
- **Mysterious or unsupported answers will not receive full credit**. A correct answer, unsupported by calculations, explanation, or algebraic work will receive no credit; an incorrect answer supported by substantially correct calculations and explanations might still receive partial credit.
- If you need more space, use the back of the pages; clearly indicate when you have done this.

Do not write in the table to the right.

Problem	Points	Score
1	2	
2	3	
3	2	
4	3	
Total:	10	

1. (2 points) Answer the following questions:

   (a) (1 point) Describe two ways of forcing a bottleneck in an Autoencoder

   (b) (1 point) Ensemble learning is based on the aggregation of decisions of different classifiers. For ensemble learning to work classifiers must be diverse. What does diversity mean in this context? Name three different ways of achieving diversity (give a one-two lines description of each of them).

2. (3 points) You have been hired in the new trendy RamenTech company "Uzumaki RamenHiro". In our last project we need to design a machine learning technique displaying a non-linear decision boundary to achieve the required accuracy for the new service of predicting whether a new customer will order Miso or Shoyu based Ramen. A good prediction will definitely increase the service speed and stock management. This in turn will revert in a much better customer experience and removal of waste storage.

   To that end, the data science team in "Uzumaki RamenHiro" has appointed the most experienced data scientist in the country. This is you!

   Unfortunately, there are some constraints that must be obeyed,

   - We do not have direct access to the dataset but for a streaming API that allows to query and get the data. (This precludes the potential use of interaction variables).
   - Unfortunately, the engineering and legal departments are new and unexperienced and can only clear the use of logistic regression as base classifier.

   Your mission is to design a solution to this problem. Describe a solution detailing how to solve to the problem using just logistic regression classifiers. (The detail must be enough to help a ML engineer to code the solution, though it is not needed to code the solution or to strictly write the solution mathematically). To help in your quest you have in Figure 1 an example of the dataset.

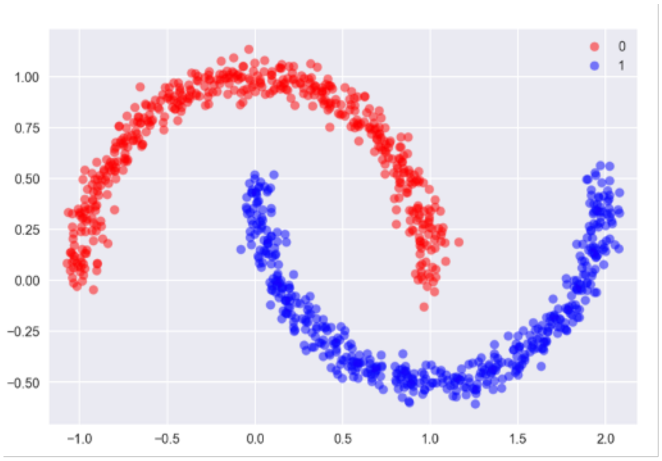

Figure 1: Example of the dataset of customer that prefer Shoyu Ramen (in red) and Miso Ramen (in blue).

   You need to explicitly specify:

   (a) (1 point) The hypothesis space

   (b) (1 point) The objective function

   (c) (1 point) The learning algorithm

3. (2 points) After the success in the design of the solution, the company has flourished. Due to this increment in popularity, the company wants to abide to strict ethical rules. The company is starting a discount plan project for customers fueled by machine learning. But, it is concerned with fairness when predicting for a customer's discount. To that respect it has hired a new team of data scientist that is using a well known problem formulation as model:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \|F\|_H + C\sum \xi_i && (1)\\ \text{subject to} \quad & F(x_i;\theta) \geq 1 - \xi_i \quad \{(x_i, y_i)|y_i = +1\} && (2)\\ & F(x_i;\theta) \leq -1 + \xi_i \quad \{(x_i, y_i)|y_i = -1\} && (3) \end{aligned}$$

where $D = \{(x_i, y_i)\}$ is the dataset, with y_i is a binary target variable that take values $+1$ if the customer can be awarded with a discount or -1 otherwise. x_i includes different features. One of the features is "Preference" and take values "Shoyu" or "Miso". Unfortunately, in practice, this model displays disparate treatment for shoyu and miso ramen lovers when competing for the advantage plan. For this reason, we want you to correct the model formulation.

In that respect we can consider the concept of *equal opportunity*. Equal opportunity is defined as displaying the same probability when conditioning with respect to the protected variable values. In our case, the protected variable is "Preference". In our binary problem this accounts for $P(y' = 1|\text{Preference} = \text{“Miso”}, y = 1) = P(y' = 1|\text{Preference} = \text{“Shoyu”}, y = 1)$ and $P(y' = -1|\text{Preference} = \text{“Miso”}, y = -1) = P(y' = -1|\text{Preference} = \text{“Shoyu”}, y = -1)$. We are going to adapt this definition by approximating this definition in a discriminative setting as follows: we want the classifier to produce the same outcome disregarding the value of variable "Preference".

In order to help in this respect, we will create a new dataset $\bar{D}$ that is exactly equal to the original dataset D except for the fact that we have changed the value of feature "Preference" to the opposite (if a smaple has "Preference = Shoyu" it is changed to "Preference = Miso" and the other way around).

(a) (1 point) **Rewrite** the equations in Eq.2 and Eq.3 to account for the data from both data sets D and $\bar{D}$ and **briefly describe** what would be the expected effect of this change in terms of accuracy and in terms of fairness. (If you need to adapt Eq.1 you may do it).

(b) (1 point) **Modify and write** the objective function in Eq.1 with the aim to explicitly fix the equal opportunity disparity. **Briefly describe** what would be the expected effect of this change in terms of accuracy and in terms of fairness.

4. (3 points) After solving the fairness problem we are faced we a new challenge. The CEO of ”Uzumaki RamenHiro” wants to design and deploy a new ML based product. Unfortunately, this product was not previously foreseen in our current stored data D. During the last weeks the data team has been labelling data, but it is not sufficient. Fortunately, we remember that there exists a technique called semi-supervised learning that may help in this situation. Semi-supervised learning considers the problem of classifying but it biases the solution so that the boundary passes through a low data density area. The technique that we are going to use is to automatically label all non-labelled data and then build a classifier on the new data set.

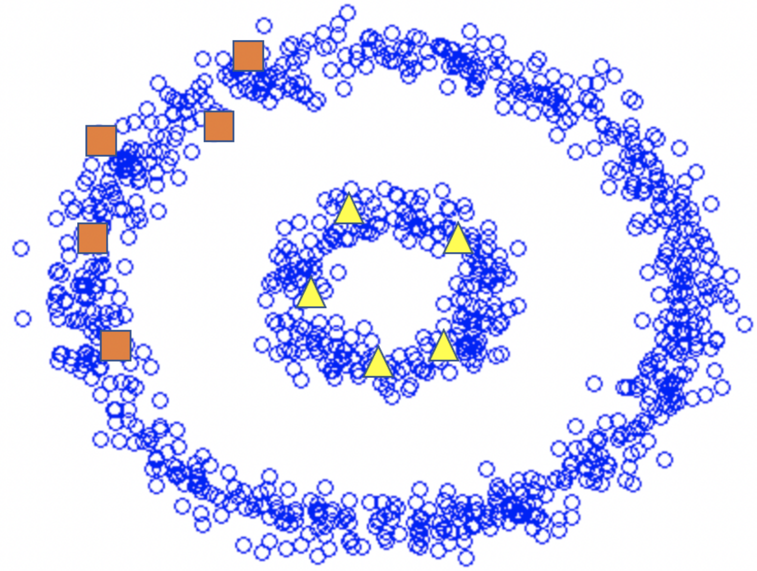

Figure 2: Example of the dataset. Blue circles correspond to unlabelled data. Orange squares are elements from class +1 and yellow triangles are samples from class -1.

To help in designing the solution, answer the following questions:

(a) (1 point) Use the former picture to **draw** the simplest decision boundary if we use only the labelled data.

(b) (1 point) **Describe** a method for automatically labelling all the points in the dataset (you can use the example in Figure 2 as an inspiration or to help in your explanation).

(c) (1 point) In order to solve the problem we want to use a kernel based method. **Define** a kernel that solves this problem as depicted in Fig.2 after relabelling.