Assignment 1

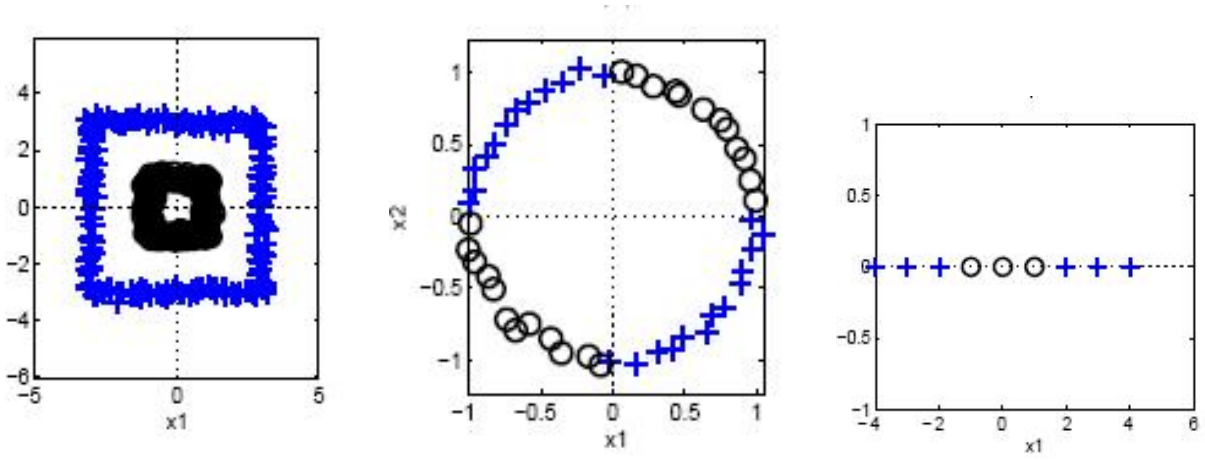
Due: 11.59pm Monday 10^{th} May 2021

Rules

1. This is a group assignment. (There are approximately 3 people per group and by now you should know your assigned group.)
2. While `R` is the default package / programming language for this course you are free to use `R` or `Python` for the programming components of this assignment.
3. Within each group **I strongly encourage each person to attempt each question by his / herself first** before discussing it with other members of the group.
4. Students should **not** consult students in other groups when working on their assignments.
5. Late assignments will **not** be accepted and all assignments must be submitted through the Hub with one assignment submission per group. Your submission should include a PDF report with your answers to each question as well as any relevant code. Make sure your PDF clearly identifies each member of the group by CID and name.

---

1. **Basis Functions and Linear Separation (15 marks)**
   In each of the following three figures there are two classes denoted by the colors blue and black. In each case, the classes are not linearly separable. For each case suggest basis functions which would render the classes linearly separable.



(a) (b) (c)

Solution: There are many possible answers to this questions. Here are some possibilities:

(a) The single basis function $f(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$ will separate the classes. Likewise the two basis functions $f_1(x_1, x_2) = |x_1|$ and $f_2(x_1, x_2) = |x_2|$ will also separate them.

(b) $f(x_1, x_2) = x_1 x_2$

(c) $f(x_1) = |x_1|$

Note: This question was adopted from the late Ben Taskar's CIS520 Machine Learning course at the University of Pennsylvania.

---

2. **Tahoe Healthcare Systems (50 marks)**
Read the Tahoe Healthcare Systems case carefully and then answer the following questions. (The data for the case can be found in the *Tahoe_Healthcare_Data.csv* file. This file and the case study itself are both posted on the Hub.) All of your answers below should be computed **taking the data-set as representative** of what will happen in a given year if nothing is done to reduce the readmissions rate.

(a) Suppose we consider the status quo and don't implement CareTracker. How much will this cost Tahoe due to the loss in Medicare reimbursements? You can assume that (as stated in the case) the loss would be \$8,000 per re-admitted patient. (Hint: Your answer should be \$7,984,000 but you should explain where this comes from!) **(5 marks)**

Solution: In the data-set there were 998 readmitted AMI patients and with each one costing \$8,000, this would amount to a loss of $998 \times 8,000 = \$7,984,000$.

(b) Suppose CareTracker was implemented for all AMI patients. What would be the net change in cost (over the status quo) from doing this? Should Tahoe implement CareTracker for all AMI patients? **(5 marks)**

Solution: CareTracker costs \$1,200 per person and with 4,382 patients in the data-set this comes to $1200 \times 4382 = \$5,258,400$. The savings will come from the 40% reduction in readmissions which results in a savings of $.4 \times 998 \times 8,000 = \$3,193,600$. The net savings will therefore be $3,193,600 - 5,258,400 = -\$2,064,800$. There is no net benefit then to introducing CareTracker to all patients and so Tahoe should not do this.

(c) Despite this seemingly bad news it may be possible to apply CareTracker to the subset of the AMI patient population who are most at risk (of readmission) and save money in the process. To investigate this, Tahoe want to build a classification model to predict those AMI patients who will be admitted and those who won't. Before doing so, however, they want to estimate an upper bound on the possible savings they could make using such a classification model. Specifically, suppose Tahoe had perfect foresight regarding what patients will need to be readmitted. What savings could they make in this event? **(10 marks)**

Solution: If Tahoe had perfect foresight they would apply CareTracker only to the patient population that would be readmitted. There are 998 such patients in the database and so the savings from this perfect foresight would be (why?)

$$998 \times (8,000 \times 40\% - 1,200) = \$1,996,000.$$

This number is important because it gives an upper bound on the savings achievable from a perfect classification algorithm. We will see below to what extent some of these idealistic savings are attainable.

(d) One simple idea for a classification algorithm is to classify based on the *severity.score* variable since (and you can check this) the mean value of *severity.score* is considerably higher for readmitted patients (*readmit30* = 1) than it is for patients who were not readmitted (*readmit30* = 0). Consider then the simple classifier which predicts that a patient will be readmitted if the patient's severity score, S, satisfies $S > S^*$, for some fixed threshold S^*.

Write a piece of code that computes the cost savings (over the status quo) in the data-set for values of S^* increasing from 25 to 100 in increments of 1. Create a graph of these estimated cost savings as a function of S^*. What is the best value (in terms of cost savings) for the threshold S^*? (*Hint:* As a check on your work, you should obtain a cost savings of \$111,200 over the status quo when you take $S^* = 50.5$.) **(10 marks)**

Solution: See the *R Notebook* that accompanies these solutions to see the code and plot. You should have found a value of $S^* = 41$ was the best threshold with a corresponding savings of \$136,800. We now know (or at least estimate) that CareTracker can be profitable if applied to the appropriate subpopulation of the AMI patient population!

(e) Classification algorithms based on *severity.score* alone are very crude, however, and Tahoe now expects that they can do better using other more sophisticated algorithms. In particular, they want to use logistic regression to fit a model using the entire data-set to estimate the probability of readmission. Write a piece of code to fit a logistic regression model to the entire data-set. The goal is to predict *readmit30* as a function of the other covariates. The *glm* function in `R` can do this for you.) **(10 marks)**

Solution: See the *R Notebook* that accompanies these solutions to see the code.

(f) Based on the fitted logistic regression model they plan (analogously to what we did earlier with the *severity.score*-based classifier) to construct a series of classifiers based on the predicted probability of re-admission. In particular, consider the classifier which predicts a patient will be readmitted if the patient's estimated probability (p) of readmission (according to the fitted logistic regression model) satisfies $p > p^*$, for some fixed threshold p^*.

Write a piece of code that computes the cost savings (over the status quo) in the data-set for values of p^* increasing from .1 to .9 in increments of .01. Create a graph of these estimated cost savings as a function of p^*. What is the best value (in terms of cost

savings) for the threshold p^*? (*Hint:* As a check on your work, you should obtain a cost savings of $\approx \$320,000$ over the status quo when you take $p^* = .6$.) What is the best value (in terms of cost savings) for the threshold p^*? **(10 marks)**

Solution: See the *R Notebook* that accompanies these solutions to see the code. The best value of p^* is 0.4 with corresponding savings over the status quo of $495,200.

Remarks

(i) You certainly don't have to do this but if your plotting abilities in `R` (or `Python`) are good, then it's pretty straightforward to adapt your code to create the ROC curves for both the *severity.score* and logistic regression classifiers. If you plot both curves on the same graph you will see that the logistic ROC curve is clearly better than the *severity.score* ROC curve in that it has a larger "area under the curve". (Remember the best possible situation is that your ROC curve has an "area under the curve" equal to 1.)

(ii) In practice, we could (and should) have created a training and validation set from the .csv file so as to estimate the out-of-sample cost savings and the best threshold (p^*) on the validation set. Alternatively, we could have used cross-validation to do this but ... one step at a time.

(iii) Speaking of cross-validation, there is an `R` package called **glmnet** which allows you to fit a logistic regression model with a Lasso regularization penalty. (In fact we use **glmnet** to do Lasso and ridge regression!) We could also have used this package and chosen the appropriate level of regularization using cross-validation. Indeed one could have used cross-validation to choose both the amount of regularization, i.e. size of the Lasso penalty, as well as the threshold p^* to optimise the cost-savings out-of-sample. But again ... one step at a time.

---

3. **Logistic Regression (20 marks)**
Suppose we are faced with a classification problem where the goal is to predict the outcome $Y \in \{0, 1\}$ given a feature vector $\mathbf{X} \in \mathbb{R}^m$.

(a) Which of the following is *false* regarding logistic regression? **(5 marks)**

(i) It fits a model of the form $p(Y \mid \mathbf{X})$ where $p(Y \mid \mathbf{X})$ denotes the conditional probability distribution of $\mathbf{Y}$ given X.

(ii) It is fit via maximum likelihood estimation.

(iii) It fits a model of the form $p(\mathbf{X}, Y)$ where $p(\mathbf{X}, Y)$ denotes the joint probability distribution of $(\mathbf{X}, Y)$.

(iv) It always produces a linear classifier.

Solution: (iii)

(b) Explain your answer from part (a). **(15 marks)**

Solution: Logistic regression assumes

$$\mathrm{P}(Y = 1 \mid \mathbf{X}) = \frac{\exp\left(\mathbf{w}^\top \mathbf{X}\right)}{1 + \exp\left(\mathbf{w}^\top \mathbf{X}\right)}$$

where $\mathbf{w} \in \mathbb{R}^m$ is the vector of parameters to be estimated. (i) is therefore true and in fact it is (iii) that is false.

It's also the case that logistic regression is fit via maximum likelihood estimation and so (ii) is also true. Finally to see that logistic regression produces a linear classifier, we note that boundary between classes $Y = 0$ and $Y = 1$ is the boundary defined by $\mathrm{P}(Y = 1 \mid \mathbf{X}) = \mathrm{P}(Y = 0 \mid \mathbf{X}) = 1 - \mathrm{P}(Y = 1 \mid \mathbf{X})$. That boundary is therefore determined by

$$\begin{aligned} \frac{\exp\left(\mathbf{w}^\top \mathbf{X}\right)}{1 + \exp\left(\mathbf{w}^\top \mathbf{X}\right)} &= 1 - \frac{\exp\left(\mathbf{w}^\top \mathbf{X}\right)}{1 + \exp\left(\mathbf{w}^\top \mathbf{X}\right)} \\ &= \frac{1}{1 + \exp\left(\mathbf{w}^\top \mathbf{X}\right)} \end{aligned}$$

which after simplifying yields $\mathbf{w}^\top \mathbf{X} = 0$ which is linear. So (iv) is also true.
