CSM148 Homework 3

Due date: Friday, February 25 at 11:00 PM PST
Instructions: All work must be completed independently, and submitted independently.

Start each problem on a new page, and be sure to clearly label where each problem and subproblem begins. All problems must be submitted in order (all of P1 before P2, etc.).

No late homeworks will be accepted. This is not out of a desire to be harsh, but rather out of fairness to all students in this large course.

1 Multinomial Logistic Regression

In multinomial logistic regression with K classes, we fit $K-1$ independent binary logistic regression models, in which one outcome is chosen as a "pivot" and then the other $K-1$ outcomes are separately regressed against the pivot outcome. For example, if outcome K (the last outcome) is chosen as the pivot:

$$\ln\frac{P(Y_i=1)}{P(Y_i=K)}=\beta_1 X_i, \quad \cdots, \quad \ln\frac{P(Y_i=K-1)}{P(Y_i=K)}=\beta_1 X_i$$

.

(a) **(5 points)** Derive the probability for class K. *Hint: the probabilities should sum to 1.*

(b) **(5 points)** With a bit of algebraic manipulation, we can further normalize the probabilities to get the softmax function:

$$P(Y_i=k)=\frac{e^{X_i\beta_k}}{\sum_{j=1}^{K}e^{X_i\beta_j}}.$$

Derive the equations for the $K-1$ decision boundaries. *Hint: the prediction for X_i changes from $k\in[K]$ to $r\in[K]$ when $P(Y_i=r)>P(Y_i=k)$.*

2 Decision Trees

Consider the data given in the following table. Suppose you want to build a decision tree based on the given data using entropy gain to determine whether you enjoy going to stadium to see a match or not.

Match	Type	Price	Location	Level	OK
R1	Football	$	LA	Intermediate	0
R2	Tennis	$$	Burbank	Advanced	0
R3	Basketball	$$	Burbank	Beginner	0
R4	Basketball	$$$	Long Beach	Intermediate	0
R5	Basketball	$	LA	Intermediate	1
R6	Football	$$	Burbank	Beginner	1
R7	Tennis	$	Burbank	Beginner	1
R8	Basketball	$	Long Beach	Advanced	0
R9	Football	$$$	LA	Beginner	0
R10	Tennis	$$	Long Beach	Intermediate	1
R11	Basketball	$$	Long Beach	Advanced	1

(a) **(15 points)** Build a decision tree to decide whether you would go to see a particular match or not. Show at each level how you decided which attribute to expand next. Stop expanding a node if entropy gain is 0.

(b) **(2 points)** What is the training set error of your decision tree (i.e. the fraction of points in the training set that is misclassified)?

(c) **(3 points)** You are now given data from five more matches: To which one would you go?

Match	Type	Price	Location	Level
R12	Football	$	Burbank	Beginner
R13	Tennis	$$	Long Beach	Beginner
R14	Tennis	$	LA	Advanced
R15	Basketball	$	Long Beach	Intermediate
R16	Tennis	$	Burbank	Advanced

(d) **(3 points)** To verify your decision tree accuracy, you decide to go to stadium to watch them all. The results are:

Match	OK
R12	0
R13	1
R14	0
R15	1
R16	0

How good did your decision tree do? What is the test set error? What is the F1 score?

(e) **(10 points)** Based on the above test data, what is the best cost-effective Tree? *Hint: use the algorithm in slide 35 of lecture 12, and show the value of the objective discussed in the slide for all the pruned trees.*

3 Decision Tree (Short Answers)

(a) **(3 points)** On which one of the datasets in Fig. 1 decision trees tend to overfit and one which one they work well? Explain briefly WHY. Which model would you use instead to avoid overfitting?

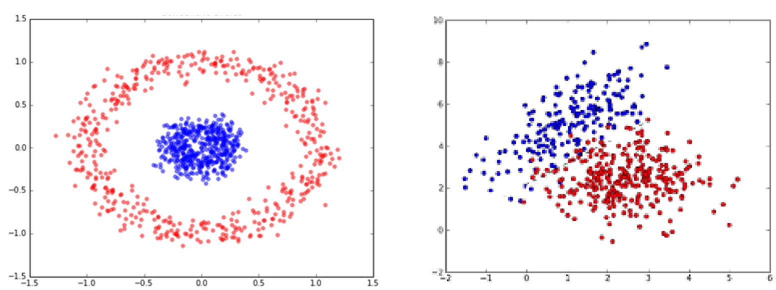


Figure 1: Datasets. Left: non-linearly separable, Right: not well-separable

(b) **(3 points)** Should we standardize our features when using a Decision Tree?

(c) **(3 points)** Are decision trees robust to outliers?

4 Random Forest

Suppose we are given a dataset generated by the following python code:

```
from sklearn.datasets import make_regression
X, y = make_regression(n_features=10, n_informative=5, random_state=10,
shuffle=False, n_samples=1000)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=10)
```

(a) **(2 points)** Fit a Random Forest with 100 trees, the maximum depth of a single tree is 5, and the number of predictors to randomly select at each split is 2. What is the R^2 for your trained Random Forest on the test dataset?

(b) **(5 points)** Suppose you are not satisfied with the Random Forest you trained in question (a), and you want to find a good Random Forest by tuning its hyper-parameters. Describe what will you do to achieve this goal and what is the best model you find (specify the hyper-parameters and R^2 for your best model)?

(c) **(2 points)** Give the variable importance for the Random Forest obtained in question (b).

(d) **(2 points)** Suppose you use bagging to fit a model based on the same hyper-parameters, i.e., the number of trees and the maximum depth of a single tree, in question (b). Give the variable importance for your bagging model and compare it with the results obtained in question (c).

5 Perceptron

You are provided with a perceptron that takes **2 inputs** and **outputs a 1 if the sum is greater than 0, else it returns a 0**. The **weights** for both inputs along with the constant are initially **set equal to 1**. The **Learning Rate** for the model is **equal to 2**.

For this model assume the perceptron updates after every data point, and that the weight of the coefficient is modified as well.

(a) **(10 points)** Given the following set of data, show how the model would update after each point. You only need to solve one complete epoch. *Hint: It may be easiest to represent inputs, outputs, and weights as a table.*

X1	X2	Y
2	-3	1
4	4	0
2	-3	0

(b) **(4 points)** The Perceptron Convergence Theorem proved that the learning model to find a solution would converge on a solution in finite time if a solution existed. Based on how your model has trained, has it converged on a solution? If not, would you expect it to if you continued training it for additional epoches?

6 Neural Networks

(a) **(4 points)** Assume you want to classify the following four points $(x_1, x_2) \in \mathbb{R}^2$:

X_1	0	0	1	1
X_2	0	1	0	1
class	+1	+1	+1	-1

In this question we will use activation units of the form

$$y = f_H(w_0 + \sum_{i=1}^{n} x_i w_i),$$

where f_H is the following threshold function

$$f_H(\alpha) = \begin{cases} -1, & \text{if } \alpha < 0 \\ +1, & \text{if } \alpha \geq 0 \end{cases}$$

Note that the numbers x_i are the inputs of the unit.

Show by specifying a correct set of parameters that the above dataset can be perfectly classified with a single activation unit.

(b) **(4 points)** Consider the following points:

X_1	0	0	1	1
X_2	0	1	0	1
class	+1	-1	-1	+1

Provide a simple argument explaining why a neural network with one unit (of the type in (a)) cannot perfectly classify these points. (You can choose to draw a picture if you prefer).

(c) **(5 points)** Explain if the dataset in part (b) be perfectly classified with more than one units (of the type in (a))?

7 SVM Decision Boundary

The following figure shows SVM decision boundaries resulting from using different kernels and/or different regularization. Circles and squares show two classes of training data. The solid circles and squares represent the support vectors. For each of the following options, match the corresponding decision boundary in the Figure and briefly explain WHY (in terms of both decision boundary and support vectors) you pick the figure for a given kernel. Note that there are 6 plots, but only 5 options, so one plot does not match any of the options.

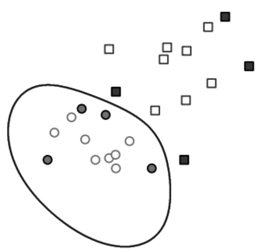
(a) **(2 points)** A soft-margin linear SVM with C = 0.1

(b) **(2 points)** A soft-margin linear SVM with C = 10

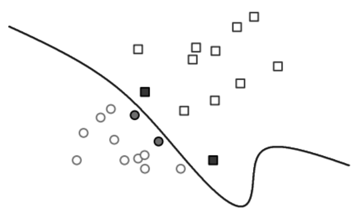
(c) **(2 points)** A hard-margin kernel SVM with $K(u, v) = u.v + (u.v)^2$

(d) **(2 points)** A hard-margin kernel SVM with $K(u, v) = \exp(-\frac{1}{4}\|u - v\|^2)$

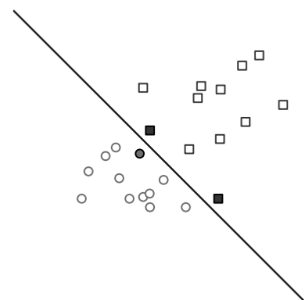
(e) **(2 points)** A hard-margin kernel SVM with $K(u, v) = \exp(-4\|u - v\|^2)$



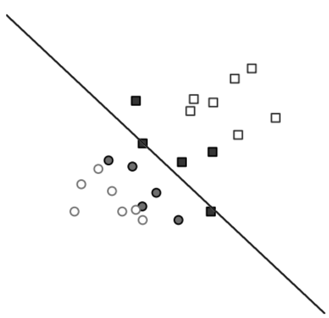
1



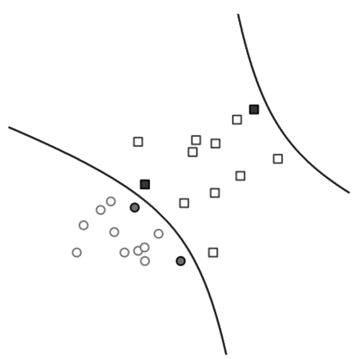
2



3



4



5

6