Name:

USC ID:

Notes:

- Write your name and ID number in the solution you submit.
- No books, cell phones or other notes are permitted. Only two letter size cheat sheets (back and front) and a calculator are allowed.
- Problems are not sorted in terms of difficulty. Please avoid guess work and long and irrelevant answers.
- Show all your work and your final answer. Simplify your answer as much as you can.
- Open your exam only when you are instructed to do so.

Problem	Score	Earned
1	20	
2	20	
3	20	
4	20	
5	20	
6	20	
Total	120	

1. Assume that we have a Ridge regression problem with only one predictor, and the true model is linear *without an intercept*, i.e. $Y = \beta_1 X + \epsilon$. Assume that we have n samples, $(x_i, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and we want to find the $\mathcal{L}_2$ regularized least squares estimate $\hat{\beta}_1$ from the data.

   (a) Formulate the objective function in terms of a candidate $\hat{\beta}_1$ and x_i's and y_i's, which are known. Assume that the regularization parameter is λ

   (b) Find $\hat{\beta}_1$ in terms of λ and the data.

2. The data the specialists have collected about malignant and benign tumors in a specific organ in the body is presented below. There are three attributes of each tumor that the cancer specialists can determine without an operation: 1) is the tumor large? 2) is the tumor hard? 3) is the tumor symmetrical? These attributes are abbreviated L, H and S respectively and their truth value is given in the table. For the collected data, the specialists also know whether the tumor turned out to be malignant or not. This fact is in the last column labeled with an M.

L	H	S	M
T	F	F	T
F	F	F	T
F	T	T	T
F	F	T	F
T	F	T	F

   (a) Train a decision tree on the data using cross entropy. The stopping criterion is zeo entropy, i.e. when all data points in a region are in the same class.

   (b) Using the decision tree you trained, predict the labels in the following test data:

L	H	S	M
F	F	F	?
T	T	T	?
F	F	T	?

3. Consider the following training data:

Index	X_1	X_2	Y
1	1	1	1
2	2	2	1
3	2	0	1
4	0	0	-1
5	1	0	-1
6	0	1	-1

(a) Carefully sketch these six training points. Are the classes linearly separable?

(b) Construct the weight vector of the maximum margin hyperplane by inspection and identify the support vectors..

(c) For each of the support vectors answer the following question: if you remove the support vector, does the size of the optimal margin remain the same, increase, or decrease?

4. Consider a MLP with one input, two layers, one neuron in each layer, and one output. The activation function of the first layer is $f^{(1)}(n) = 2n$ and the activation function of the second layer is $f^{(2)}(n) = 3n$. The initial weights of the first and the second layers are respectively $w^{(1)} = 0$ and $w^{(2)} = 0$. Assume that we present the data point with $x = 1$ and $y = 2$ to the network. Perform one step of the Stochastic Gradient Descent algorithm by using the backpropagation algorithm, assuming the learning rate $\alpha = 0.5$. Use the MSE objective function, i.e. $J = (y - a^{(2)})^2$. This means that you should calculate the updated weights

5. Assume that you have a labeled dataset. Explain how you can use only K-means clustering to build a classification model for this dataset. What can go wrong?

6. Choose either T (True) or F (False):

   (a) Complexity of random forests significantly changes when the number of trees B increases, and they become more prone to overfitting. T F

   (b) Gaussian Kernels are always the kernels of choice for implicit expansion of feature space in SVMs. T F

   (c) Instead of majority polling, one can use SVMs in each segment of the feature space in a decision tree to predict labels. T F

   (d) Semi-supervised learning cannot be used for regression problems. T F

   (e) A Multilayer Perceptron with one hidden layer of sigmoids and an output layer with linear activation functions can learn a linear regression function with any precision and this does NOT defy the no free lunch theorem. T F

   (f) The $\mathscr{L}_1$ regularizer can easily exclude correlated features when combined with logistic regression. T F

   (g) To convert a multiclass classification problem with 10,000 classes to multiple binary classification problems, it makes more sense to choose the One-Versus-All (OVA) method over the One-Versus-One (OVO) method. T F

   (h) For a multi-label classification problem with 100 binary labels, the label power set method is the best method of converting the problem into a multi-class problem. T F

   (i) In a binary classification problem, if we are interested in class conditional probabilities, we can not use Support Vector Classifiers. T F

   (j) The results of K-means and Hierarchical Clustering can be different for the same data set, because K-Means does not assume that clusters are nested. T F

Scratch paper

Name:

USC ID:

Scratch paper

Name:

USC ID: