MIDTERM: CS 6375
INSTRUCTOR: VIBHAV GOGATE
October, 27 2014

The exam is closed book. You are allowed a one-page cheat sheet. Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, use an additional sheet (available from the instructor) and staple it to your exam.

- **NAME** ______________________________

- **UTD-ID if known** ______________________________

Section Number	**Points**
1	
2	
3	
4	
5	
6	
Total	

1 SHORT ANSWERS [11 points]

1. (3 points) For linearly separable data, can a small slack penalty hurt the training accuracy when using a linear SVM (no kernel)? If so, explain how. If not, why not?

2. (3 points) Let us assume that the data (y) was generated from the following distribution:

$$\Pr(y) = \frac{\theta^y e^{-5\theta}}{y!}$$

   Let us assume that you are given n data points $y_1, \ldots, y_n$ independently drawn from this distribution. Write down the expression for the maximum likelihood estimate for θ.

3. (2 points) Gaussian Naive Bayes is a linear classifier. True or False. Explain.

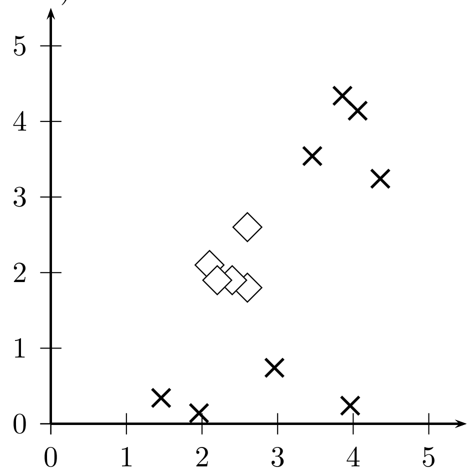
Leave-one-out cross validation is a special case of k-fold cross validation in which k equals the number of examples n. Thus, at each iteration i, where $1 \leq i \leq n$, we use the i-th example for testing and the remaining $n-1$ examples for training. The leave-one-out cross validation error equals the average error over the n iterations.

4. (3 points) The Leave-one-out cross validation error of 1-nearest neighbor classifier is always zero (assume that you are using the Euclidean distance as your distance measure). True or False. Explain your answer.

2 Classification [9 points]

For each of the following datasets, write alongside each classifier whether it will have zero training error on the dataset. Also, explain why in one sentence or by drawing a decision boundary. No credit if the explanation is incorrect.

1. (3 points)

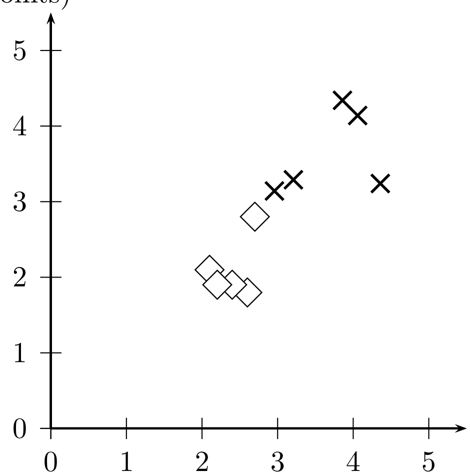


Logistic Regression:

3-nearest neighbors:

SVMs with a quadratic kernel:

2. (3 points)

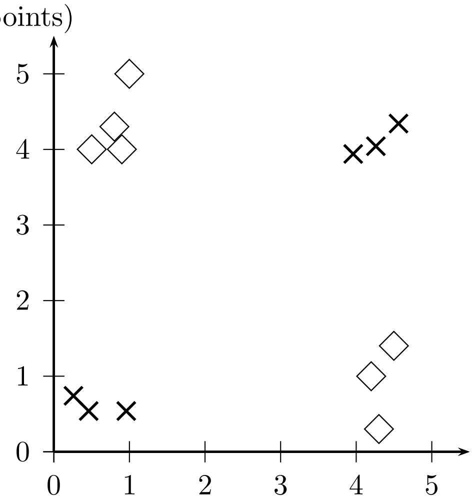


Logistic Regression:

3-nearest neighbors:

SVMs with a quadratic kernel:

3. (3 points)



Logistic Regression:

3-nearest neighbors:

SVMs with a quadratic kernel:

3 Linear Regression [8 points]

Consider a linear regression problem $y = w_1 x + w_0$, with a training set having m examples $(x_1, y_1), \ldots, (x_m, y_m)$. Suppose that we wish to minimize the mean 3^{rd} degree error (loss function) given by:

$$Loss = \frac{1}{m}\sum_{i=1}^{m}(y_i - w_1 x_i - w_0)^5$$

1. (4 points) Calculate the gradient with respect to the parameters w_1 and w_0. Hint: $\frac{dx^k}{dx} = kx^{k-1}$.

2. (4 points) Write down pseudo-code for online gradient descent for this problem.

4 Short Questions (12 points)

1. (4 points) Consider a d-bit parity problem in which the input is a d-dimensional Boolean vector $\mathbf{x} = (x_1, \ldots, x_d)$ and the output is 1 or *true* if the number of 1's in $\mathbf{x}$ is odd and 0 or *false* otherwise. Is the problem linearly separable for $d > 1$? Explain.

2. (4 points) Estimate the number of nodes that a decision tree will require to model the d-bit parity function perfectly. Assume that your training data contains all possible, 2^d combinations and each split is on a value of a SINGLE Boolean variable. (**Hint**: Construct trees for $d = 3$ and $d = 4$ and compare the number of nodes in them to the maximum number of nodes that a decision tree can have).

3. (4 points) Consider a non-uniform prior which assigns positive probability mass to each possible hypothesis. As the number of data points grows to infinity, the MLE estimate of a parameter approaches the MAP estimate to arbitrary precision. True of False. Explain.

5 Support Vector Machines [10 points]

Consider the training data given below (Y is the class variable)

X	Y
-2	1
-1	-1
1	-1
2	1

1. (3 points) Assume that you are using a linear SVM. Let $\alpha_1, \alpha_2, \alpha_3$ and α_4 be the lagrangian mutlipliers for the four data points. Write the precise expression for the lagrangian dual optimization problem that needs to be solved in order to compute the values of $\alpha_1, \ldots, \alpha_4$ for the dataset given above.

2. (2 points) Do you think, you will get zero training error on this dataset if you use linear SVMs (Yes or No)? Explain your answer.

The dataset is replicated here for convenience.

X	Y
-2	1
-1	-1
1	-1
2	1

3. (3 points) Now assume that you are using a quadratic kernel: $(1+x_i^T x_j)^2$. Again, let $\alpha_1, \alpha_2, \alpha_3$ and α_4 be the lagrangian mutlipliers for the four data points. Write the precise expression for the lagrangian dual optimization problem that needs to be solved in order to compute the values of $\alpha_1, \ldots, \alpha_4$ for the dataset and the quadratic kernel given above.

4. (2 points) Do you think, you will get zero training error on this dataset if you use the quadratic kernel (Yes or No)? Explain your answer.

6 Neural networks [25 points]

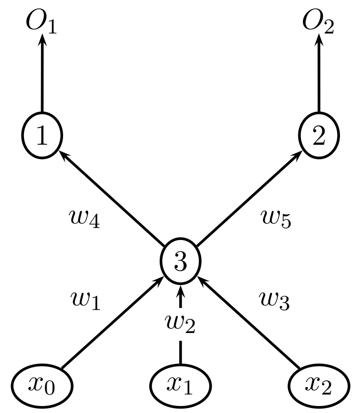
1. (2 points) Is it always possible to express a neural network made up of only linear units without a hidden layer? Namely, given a neural network with several layers made up of only linear units, can you find a perceptron that is equivalent to the neural network? Explain your answer.

2. (8 points) Draw a neural network that represents the function $f(x, y)$ defined below:

x	y	$f(x,y)$
0	0	10
0	1	-5
1	0	-5
1	1	10

Note that to get full credit, you have to write down the precise numeric weights (e.g., -1, -0.5, $+1$, etc.) as well as the precise units that you will use at each node (e.g., sigmoid, linear, simple threshold, tanh etc.).

Consider the neural network given below:



Assume that all internal nodes compute the sigmoid function. Write an explicit expression that shows how back propagation (applied to minimize the least squares error function) changes the values of w_1, w_2, w_3, w_4 and w_5 when the algorithm is given the example $x_1 = 0$, $x_2 = 1$, with the desired response $y_1 = 0$ and $y_2 = 1$ ($x_0 = 1$ is the bias term). Assume that the learning rate is α and that the current values of the weights are: $w_1 = 3$, $w_2 = 2$, $w_3 = 2$, $w_4 = 3$ and $w_5 = 2$. Let O_1 and O_2 be the output of the output units 1 (which models y_1) and 2 (which models y_2) respectively . Let O_3 be the output of the hidden unit 3.

3. (5 points) Forward propagation. Write equations for O_1, O_2 and O_3 in terms of the given weights and example.

4. (5 points) Backward propagation. Write equations for δ_1, δ_2 and δ_3 in terms of the given weights and example where δ_1, δ_2 and δ_3 are the values propagated backwards by the units denoted by 1 and 2 and 3 respectively in the neural network.

5. (5 points) Give an explicit expression for the new (updated) weights w_1, w_2, w_3, w_4 and w_5 after backward propagation.