CS 412: Introduction to Machine Learning
Fall 2016
Final Exam

Name: _____

User ID: _____

Instructions:

1. Write your name and user ID above. Do not begin the exam (look at other pages) until told to do so.
2. There should be 10 pages. Count the pages (without looking at the questions).
3. Do not discuss the exam with students who have not taken the exam!

Some useful formulas:

- $P(\mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{x}, \mathbf{y})$ (marginalization)
- $P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{y})}$ (conditioning)
- $P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}$ (Bayes theorem)
- $\mathbb{E}_{x\sim P}[f(X)] = \sum_x P(x)f(x)$ (Expectation)
- $X \sim \text{Normal}(\mu, \sigma) \implies P(X = x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $X \sim \text{Multinoulli}(\theta) \implies P(\mathbf{x}) = \prod_{i=1}^{K} \theta_i^{x_i}$
- AdaBoost weights: $\alpha_m = \frac{1}{2} \ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$

	Points
Q1	/20
Q2	/25
Q3	/25
Q4	/30
Total	

Q1. True-False questions (5 questions, 20 points total)

Circle correct answer. Give one sentence explanation or small picture.

Q1.1: (4 points) A classification method with low bias and more variance should always be preferred over one with more bias and lower variance.

True / False. Explanation:

Q1.2: (4 points) A support vector machine with polynomial kernel $K(x, y) = (x \cdot y + 1)^d$ will always provide at least as good of accuracy on training data as a linear support vector machine trained to heavily favor hinge-loss minimization over margin width (i.e., little/no regularization).

True / False. Explanation:

Q1.3: (4 points) The expectation-maximization algorithm is guaranteed to converge to a global optima.

True / False. Explanation:

Q1.4: (4 points) The weak classifier weights α_i in AdaBoost for binary classification are always non-negative in practice.

True / False. Explanation:

Q1.5: (4 points) Deep neural network architectures cannot be overfit to training data in practice.

True / False. Explanation:

Q2. Short Answer (4 questions, 25 points total)

Q2.1: (4 points) What are the main differences between frequentist parameter estimation (maximum likelihood) and Bayesian parameter estimation?

Q2.2: (4 points) What are the advantages and disadvantages of discriminative learning methods (logistic regression, support vector machines) versus generative learning methods (Naïve bayes)?

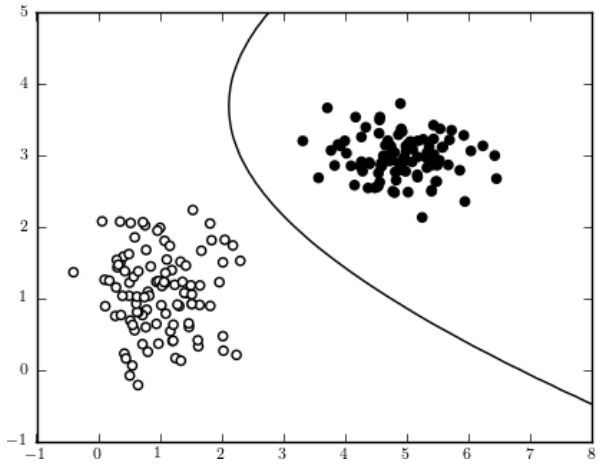
Advantages:

Disadvantages:

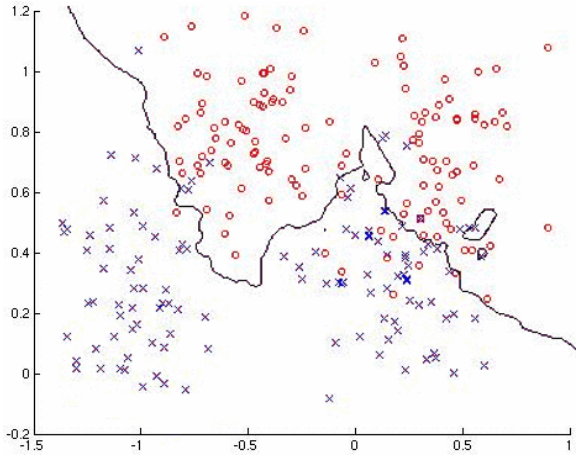
Q2.3: (5 points) What is the reason for using decision tree bagging or random forests instead of a single decision tree? What advantage(s) do random forests provide versus decision tree bagging?

Q2.4: (12 points) Consider the folowing predictors: decision tree, k-nearest neighbor, Gaussian naïve Bayes, logistic regression that is linear in the inputs, and linear support vector machines. You will need to explain which predictor(s) could produce a particular decision boundary and the parameters of the predictor that would do so (*how*).

I.



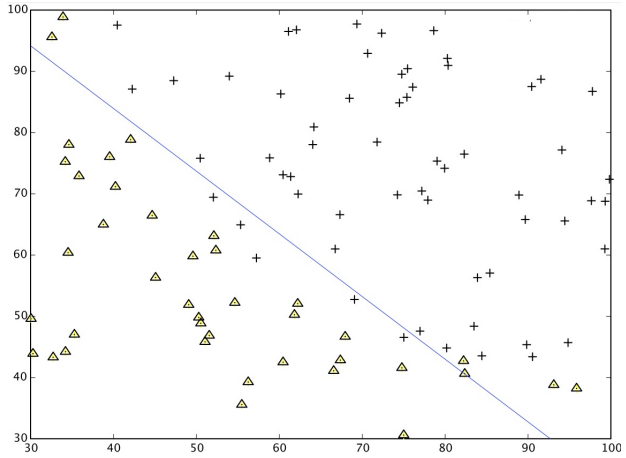
II.



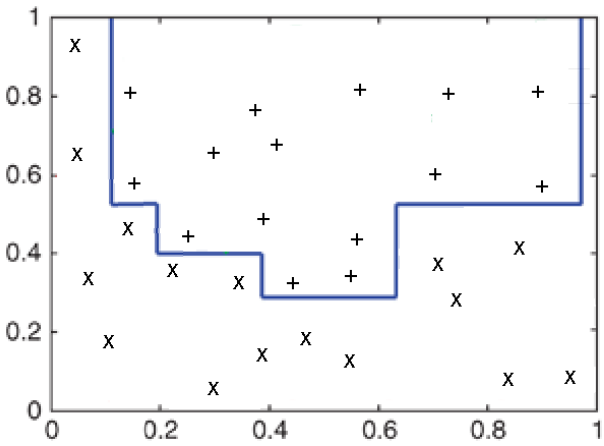
(a) **(3 points)** What predictor(s) produces the decision boundary from figure I and how?

(b) **(3 points)** What predictor(s) produces the decision boundary from figure II and how?

III.



IV.

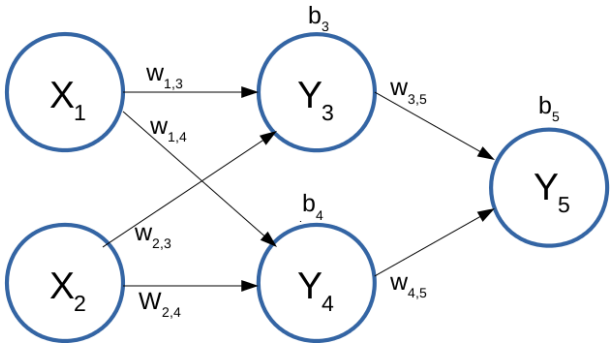


(c) **(3 points)** What predictor(s) produces the decision boundary from figure III and how?

(d) **(3 points)** What predictor(s) produces the decision boundary from figure IV and how?

Q3. Neural Networks (25 points total)

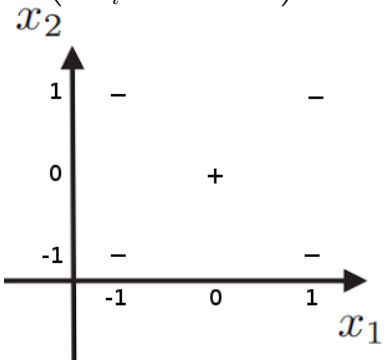
Consider the following neural network with weights w for the edges and offset b for the nodes:



Assume that the outputs of nodes Y_3, Y_4, and Y_5 are based on rectified linear functions:

$$Y_3 = \max\left(0, \sum_i x_i w_{i,3} + b_3\right) \text{ or } Y_5 = \max\left(0, \sum_i y_i w_{i,5} + b_5\right).$$

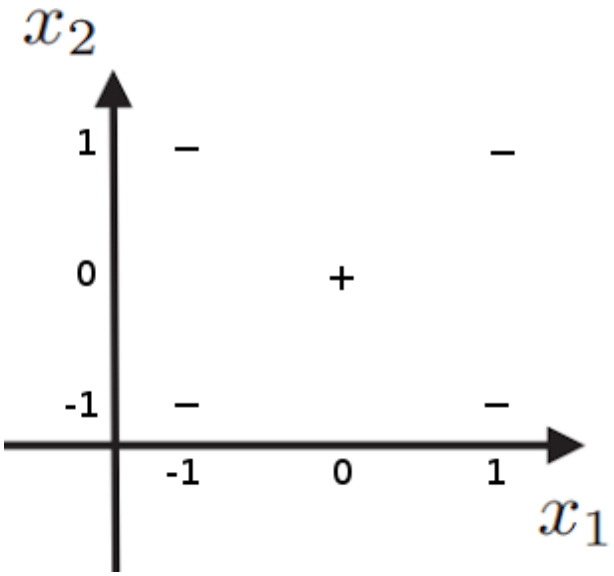
Q3.1: **(15 points)**
Consider the dataset on the right with negative $(-)$ and positive $(+)$ class labels. Assuming all neural network weights are -1, 0, or 1, choose a set of weights that makes $Y_5 = 0$ for each negative example and $Y_5 > 0$ for the positive example.



(Circle one choice for each parameter)

$w_{1,3} = -1$	$w_{1,3} = 0$	$w_{1,3} = 1$
$w_{1,4} = -1$	$w_{1,4} = 0$	$w_{1,4} = 1$
$w_{2,3} = -1$	$w_{2,3} = 0$	$w_{2,3} = 1$
$w_{2,4} = -1$	$w_{2,4} = 0$	$w_{2,4} = 1$
$w_{3,5} = -1$	$w_{3,5} = 0$	$w_{3,5} = 1$
$w_{4,5} = -1$	$w_{4,5} = 0$	$w_{4,5} = 1$
$b_3 = -1$	$b_3 = 0$	$b_3 = 1$
$b_4 = -1$	$b_4 = 0$	$b_4 = 1$
$b_5 = -1$	$b_5 = 0$	$b_5 = 1$

Q3.2: **(10 points)** Draw the decision boundaries (i.e., boundaries where Y_5 transitions from 0 to > 0) for your weights specified in Q3.1:

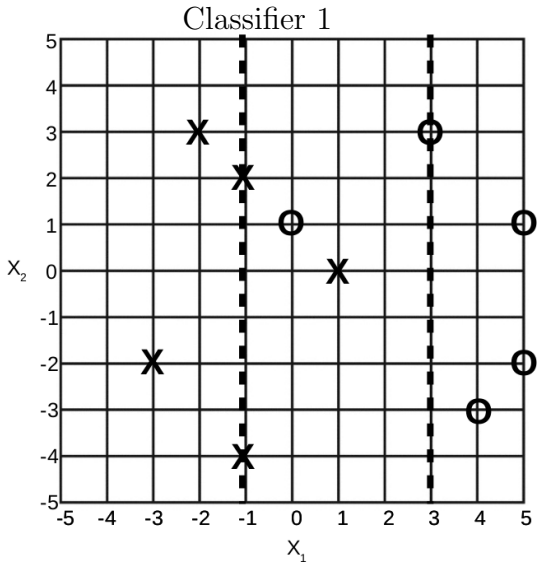


Q4. Support Vector Machines (30 points)

Q4.1 Linear SVM Consider the linear support vector machine **with the formulation we saw in class**, with M being the distance from decision boundary to margin boundary and C being a "slack budget" for the classifier:

$$\max_{M,\beta_0,\beta_1,\beta_2,\xi} M$$
$$\text{such that: } y_i(\beta_2 x_{i,2} + \beta_1 x_{i,1} + \beta_0) \geq M(1-\xi_i), \text{ and } \xi_i \geq 0, \forall_i \in \{1,...,n\};$$
$$\beta_1^2 + \beta_2^2 = 1; \text{ and } \sum \xi_i \leq C;$$

on the following dataset (with 'X' class defined as positive, $y_i = 1$ and 'O' class defined as negative, $y_i = -1$):
For the support vector classifer with margin boundaries (dashed lines) shown to the right:



(a) **(5 points)** Which datapoints are support vectors? Please circle them on the figure.

(b) **(5 points)** What is the value of M for this classifier?

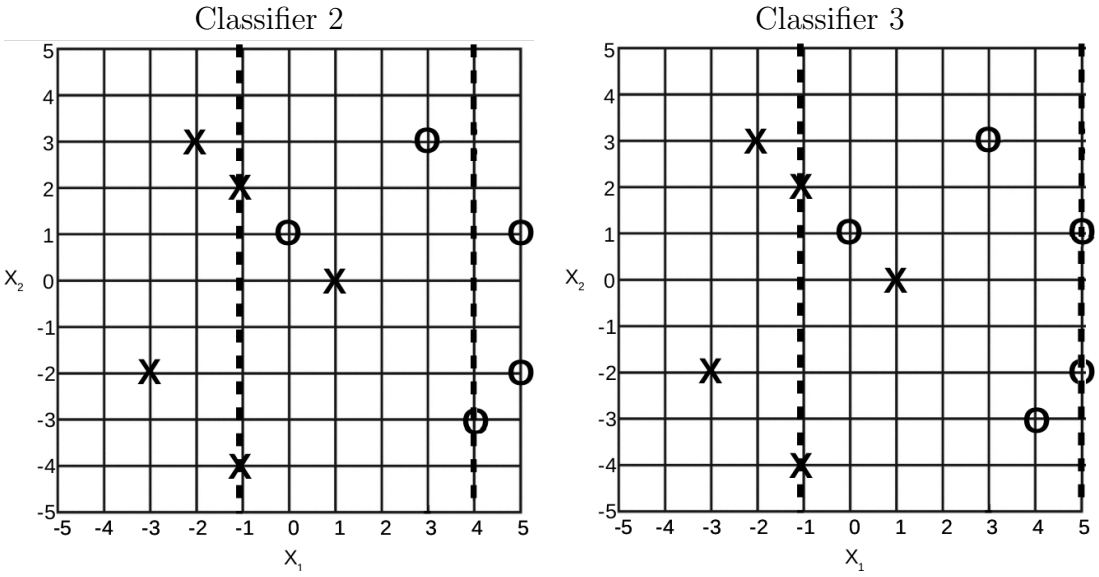
(c) **(6 points)** What are the values of β_2, β_1, and β_0 for this classifer?

$\beta_0 =$

$\beta_1 =$

$\beta_2 =$

(d) **(6 points)** What is the sum of slacks, $\sum_i \xi_i$, for this classifier?



(d) **(8 points)** Consider the classifer when the margin boundary at $x_1 = 3$ is instead moved to $x_1 = 4$ (Classifier 2) or moved to $x_1 = 5$ (Classifier 3). When will the optimization above prefer Classifier 2 or Classifier 3 (in terms slack budget parameters C) instead of Classifier 1? (Hint: what is the margin/sum of slacks for each?)

Extra space