CS534 — Final — Fall 2015

12:00pm - 1:50pm Thursday, Dec 11th

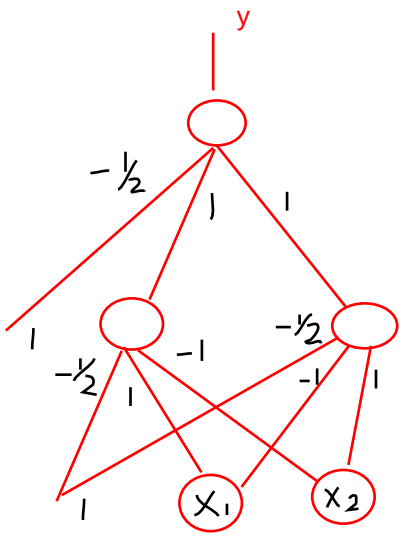
Name (Please print):

1. You have 110 minutes to finish the exam.
2. There are 7 pages in this exam (including cover page).
3. If you use the back of the page please indicate on the front of the page so we won't miss it.
4. This exam is **close book, close notes, but you are allowed one page cheat sheet (letter size, one sided)**.
5. Calculator is allowed but not necessary. No other electronics allowed.
6. Please give brief answers. Most of the questions can be answered in one or two sentences.

Problem	Max	score
1	9	
2	6	
3	15	
4	14	
5	14	
Total	58	

1. (9pts) Neural networks

   (a) (6pts) Consider the following binary function $y = (x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2)$, where $x_1, x_2 \in \{0, 1\}$. Provide a neural net to represent this function. You will use a threshold function as the activation function for the neurons. You will need to clearly specify the weights for your network.



   (b) (3pts) Why should we initialize the weights of a Neural Net to random small values?

Small random weights allow the sigmoid function to operate in the linear region and avoid saturation of the sigmoid function, which could lead to very small gradient.

2. (6 pts) Ensemble learning and bias variance decomposition

(a) (2pts) Applying Bagging to a base learner, how does it influence the bias and variance?

Bias: choose one
Increase Decrease No impact

Variance: choose one
Increase Decrease No impact

(b) (2pts) Applying Adaboost to a base learner, how does it influence the bias and variance?

Bias: choose one
Increase Decrease No impact

Variance: choose one
Increase Decrease No impact

(c) (2pts) Why is Adaboost sensitive to outliers?

The weights of the outliers get overly large, leading to overfitting to the outliers.

3. (15 pts) **Expectation Maximization and Maximum Likelihood Estimation.**
Students in Master Yoda's "Dark Side Studies" class come from two different backgrounds: Sith or Jedi.

- In the general population, we know $P(Sith) = P(Jedi) = 0.5$.
- The probability of passing the class for a Sith, $P(Pass|Sith)$, is known to be 0.8.
- The probability of passing for a Jedi, $P(Pass|Jedi)$, is unknown, denoted as θ.
- We observe a total of N_p students passing and N_f failing in this year's class.
- N_{jp} and N_{jf} denote the number of passed and failed Jedi students respectively, which are unknown.

(a) **Expectation:**(5 pts) Which of the following expressions compute the expected value of N_{jp} and N_{jf} given N_p, N_f and θ?

	N_{jp}	N_{jf}
(a)	$N_p \cdot \theta$	$N_f \cdot (1-\theta)$
(b)	$N \cdot 0.5 \cdot \theta$	$N \cdot 0.5 \cdot (1-\theta)$
(c)	$N_p \cdot \frac{0.5 \cdot \theta}{0.5 \cdot \theta + 0.5 \cdot 0.8}$	$N_f \cdot \frac{0.5 \cdot (1-\theta)}{0.5 \cdot (1-\theta) + 0.5 \cdot 0.2}$
(d)	$N_p \cdot \frac{0.5 \cdot \theta}{0.5 \cdot \theta + 0.5 \cdot 0.2}$	$N_f \cdot \frac{0.5 \cdot (1-\theta)}{0.5 \cdot (1-\theta) + 0.5 \cdot 0.8}$

(b) **Maximization**: (5 pts) Now given the expected value of N_{jp} and N_{jf}, which of the following expression computes the maximum likelihood estimate of θ?

(a) $\frac{N_{jp}+N_{jf}}{N}$

(b) $\frac{N_{jp}}{N_{jp}+N_{jf}}$

(c) $\frac{N_{jp}}{N_p}$

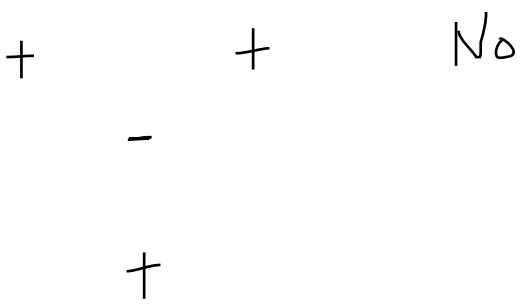
(d) $\frac{N_{jp}+N_{jf}}{N_p}$

(c) **Likelihood**: (5 pts) Please directly write out the likelihood of observing N_p and N_f as a function of θ. No need to solve for optimal θ.

$$\left(\frac{1}{2}\theta + \frac{1}{2} \cdot 0.8\right)^{N_p} \cdot \left(\frac{1}{2}(1-\theta) + \frac{1}{2} \cdot 0.2\right)^{N_f}$$

4. (12 pts) **Learning theory**

(a) (3pts) Consider the hypothesis space H of circles in a 2-d space, which classify points inside the circle to be positive and outside to be negative. Can H shatter the following four points?



The labels shown on the left cannot be correctly classified.

(b) (3pts) Instead of always labeling the interior of the circle positive, we now also have the choice of labeling the interior negative and the exterior positive. Can the four points be shattered by this new hypothesis space H' ?

Yes.

(c) (2pts) What can we state about the VC-dimension of H' based purely on the answer in (b). Choose an answer for each possibility:

If H' cannot shatter them:	$VC(H') \leq 4$	$VC(H') \geq 4$	$VC(H') = 4$	Nothing
If H' can shatter them:	$VC(H') \leq 4$	$VC(H') \geq 4$	$VC(H') = 4$	Nothing

(d) (3 pts) In class we showed that if a consistent learning algorithm for a finite hypothesis space H is provided with

$$m \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

randomly drawn training instances, then we can state a certain guarantee. What is that guarantee? Make sure to clearly indicate the roles of ϵ and δ.

With probability at least 1-\delta, consistent learn will output a hypothesis whose generalization error is less than \epsilon

(e) (3 pts) For a fixed hypothesis space H, below is a bound on the generalization error of the learned hypothesis h_L:

$$\epsilon(h_L) \leq \epsilon(h^*) + 2\sqrt{\frac{1}{2m}(\ln|H| + \ln\frac{1}{\delta})}$$

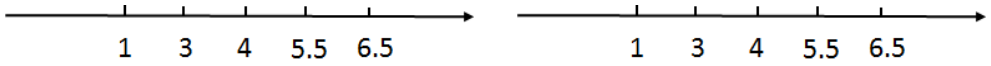
where h^* is the hypothesis in H that has the lowest generalization error. How would the two terms, and their combined sum change as we consider larger and more complex hypothesis space H? For each term, choose one.

$\epsilon(h^*)$:	Decrease	Increase	Both are possible
$2\sqrt{\frac{1}{2m}(\ln\|H\| + \ln\frac{1}{\delta})}$	Decrease	Increase	Both are possible
$\epsilon(h^*) + 2\sqrt{\frac{1}{2m}(\ln\|H\| + \ln\frac{1}{\delta})}$	Decrease	Increase	Both are possible

5. (14 pts) Clustering and Dimension reduction

(a) (6 pts) Apply single-link and complete-link HAC to the following five $1-d$ points and plot the resulting dendrograms.

$$\text{Sample} = (1, 3, 4, 5.5, 6.5)$$



(i) Single-link (ii)Complete-link

(b) (3pts) Explain why it is important to randomly restart Kmeans multiple times?

Different initialization leads to different local minimum. Randomly restart multiple times avoid bad local minimum.

(c) (5pts) Explain how ISOMAP estimates the geodesic distance between two points on the manifold?

It first constructs a graph where points are only connected to their close neighbors, and the edge weights are the euclidean distances. The geodesic distance between two points is then computed as the length of the shortest path between the two points on the graph.