CS446: Machine Learning **Fall 2014**

Mid-term Exam Solutions

October 23^{rd}, 2014

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam booklet contains **four** problems. You need to solve all problems to get 100%.
- The exam ends at 1:45 PM. You have 75 minutes to earn a total of 100 points.
- Answer each question in the space provided. If you need more room, write on the reverse side of the paper and indicate that you have done so.
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**

Good Luck!

Name (NetID): (1 Point)

Short Questions		/24
Decision Trees		/25
Online Learning		/25
Kernels		/25
Total		/100

Short Questions [24 points]

(a) [8 points] Consider the hypothesis space $\mathbf{H}$ defined by all n-dimensional hyperplanes that pass through the origin. That is, $h \in \mathbf{H}$ is defined by $\mathbf{w} \in \mathcal{R}^n$ and an example $\mathbf{x} \in \mathcal{R}^n$ is labeled positive if and only if $\mathbf{w}^T\mathbf{x} \geq 0$.
Prove that the VC dimension of $\mathbf{H}$ is at least n.

Note: We do not ask you to compute the VC dimension of $\mathbf{H}$ *exactly*; you only need to show that it is *at least* n. Write formally what you need to show; then provide an explanation for why this is true.

Solution:
We need to show that there exists a set of n points that can be shattered by $\mathbf{H}$. To show that, we consider $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathcal{R}^n$, such that the i-th component of $\mathbf{x}_i$ is 1 and the remaining components are 0 (i.e., $\mathbf{x}_i = [0, \ldots, 0, \underbrace{1}_{\text{the i-th element}}, 0, \ldots, 0]^T$).

For any label assignment $\{y_1, y_2, \ldots, y_n\}$, where y_i is the assignment to x_i in a specific dichotomy, we can find a hypothesis defined as: $\mathbf{w} = [y_1, y_2, \ldots, y_n]^T \in \mathbf{H}$, such that $y_i\mathbf{w}^T\mathbf{x}_i = 1 > 0$ for all i. Check the dot product and convince yourself that this $\mathbf{w}$ makes all the examples labeled 1 positive and all the examples labeled 0 negative. Therefore, these n points can be shattered by $\mathbf{H}$, and $VC(\mathbf{H}) \geq n$.

(b) [4 points] Consider the hypothesis space of all l-of-m-of-n Boolean functions $H_{l,m,n}$. As you already know from the homework, a function $h \in H$ is a Boolean function on the n-dimensional Boolean cube $\{0,1\}^n$ and there is a set of m of the n attributes such that an example $x \in \{0,1\}^n$ is positive if and only if at least l of these m attributes are active in the example. l, m, and n define the function class $H_{l,m,n}$.

Show that l-of-m-of-n functions are linearly separable functions.
(Hint: Find a weight vector $\mathbf{w}$ and a bias θ such that $sgn(\mathbf{w}^T\mathbf{x} + \theta)$ will make exactly the same predictions as a given l-of-m-of-n function.)

Without loss of generality, we assume the first m attributes in the given l-of-m-of-n function from $H_{l,m,n}$ are relevant and the remaining attributes are irrelevant. Let $\theta = -l + 0.5$ and $\mathbf{w} = [\underbrace{1, 1, \ldots, 1}_{m}, \underbrace{0, 0, \ldots, 0}_{n-m}]^T$; then $sgn(\mathbf{w}^T\mathbf{x} + \theta)$ will make the same prediction as the given l-of-m-of-n function. This suggests that l-of-m-of-n functions are linearly separable functions.

(c) [8 points] Given l, m, n, show that the VC dimension of the hypothesis class $H_{l,m,n}$ of l-of-m-of-n functions is upper bounded by K, where $K = O(m \log(n))$.

We proved in class that the VC dimension of a finite hypothesis class H is no more than $log(H)$. The reason is that the number of dichotomies supported by this class, 2^{VC}, must be small than $|H|$. The size of $H_{l,m,n}$ is $C(n,m) \sim n^m$. Therefore, $VC(H_{l,m,n})$ is upper bounded by $log(C(n,m)) = O(m \log(n))$.

(d) [4 points] In the following we provide three statements; two about PAC learning and one about Boosting. In each statement we left a few blank fields. Fill in the blanks by choosing, for each empty field, one of the options given below. Note that under each line defining a blank we provided a small set of options for you to choose from.

(a) δ	(b) ϵ	(c) $1/\delta$	(d) $1/\epsilon$	(e) $1-\delta$	(f) $1-\epsilon$
(g) m	(h) n	(i) $n\epsilon/\delta$		(j) size($\mathbf{H}$)	
(k) number of examples		(l) instance size		(m) computation time	
(n) linear		(o) polynomial		(p) exponential	
(q) $\frac{1}{2}-\gamma$	(r) $\frac{1}{2}+\gamma$	(s) $1-\gamma$			

(1) A concept class $\mathbf{C}$ defined over the instance space $\mathbf{X}$ (with instances of length n) is *strongly* PAC learnable by learner $\mathbf{L}$ using a hypothesis space $\mathbf{H}$ if for all concepts $f \in \mathbf{C}$, for all distributions $\mathbf{D}$ on $\mathbf{X}$, and for all fixed $\delta, \epsilon \in [0,1]$, given a sample of m examples sampled independently according to the distribution $\mathbf{D}$, the learner $\mathbf{L}$ produces with a probability

at least {at least | at most | equal to} $1-\delta$ {one of (a) to (f)} a hypothesis $g \in \mathbf{H}$ with error

($\text{Error}_{\mathbf{D}} = \Pr_{\mathbf{D}}[f(x) \neq g(x)]$) at most {at least | at most | equal to} ϵ {one of (a) to (f)}

where the number of examples {one of (k) to (m)} is polynomial {one of (n) to (p)} in

n, $1/\delta$, $1/\epsilon$, and size($\mathbf{H}$) {four of (a) to (j)}.

(2) A concept class $\mathbf{C}$ is *weakly* PAC learnable, if the above statement holds only for values of ϵ such that $\epsilon \geq$ $\frac{1}{2}-\gamma$ {one of (q) to (s)} for some $\gamma > 0$.

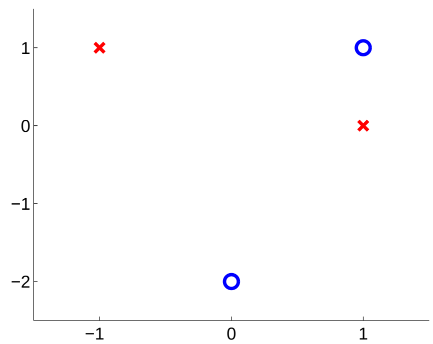
(3) The Boosting theory suggests that:
weak learnability implies {implies | does not imply | may or may not imply} strong learnability.

Kernels [25 points]

Consider the following 4 training examples in $\Re^2$,

i	$\vec{z_i}$	y_i
z_1	$(1,1)$	1
z_2	$(1,0)$	-1
z_3	$(-1,1)$	-1
z_4	$(0,-2)$	1

where $\vec{z_i}$ is the feature vector for example z_i and y_i is the corresponding label.



(a) Determine if the four examples depicted above are linearly separable in $\Re^2$.

no
{yes | no}

In the rest of the problem we will consider the question of learning a linear separator consistent with these four points in a new space. We will use the kernel perceptron algorithm. Recall that the hypothesis used by kernel perceptron is defined as the following function of example x:

$$f(\vec{x}) = \sum_{z_i \in M} y_i K(\vec{x}, \vec{z_i})$$

(b) What is M in the above equation? Be precise when defining the elements in M.

A collection of examples on which mistakes were made during training (with repetitions).

Using the definition of $f(\vec{x})$ given earlier, the prediction rule of our algorithm is given by:

$$Th_0(f(\vec{x})) = \begin{cases} 1, & \text{if } f(\vec{x}) \geq 0, \\ -1, & \text{if } f(\vec{x}) < 0 \end{cases}$$

(c) Complete the kernel perceptron algorithm below.

```
KernelPerceptron
  M ← ∅
  for each example (x⃗_i, y_i)

      if  ____y_i ≠ Th_0(f(x⃗_i))____

          M ← ____M ∪ {x_i}____
```

(d) We want to make the four examples given above linearly separable in $\Re^3$. In order to do it we define a mapping from $\vec{z} = (z_1, z_2)$ to a new space $t(z) = (z_1^2, \sqrt{2}z_1z_2, z_2^2)$.
Write down the kernel function $K(\vec{x}, \vec{z})$ represented by this mapping in terms of of $\vec{x}$ and $\vec{z}$.
$K(\vec{x}, \vec{z}) = (\vec{x}^T\vec{z})^2$

(e) Show that your definition above is indeed a kernel.
Hint: use the mapping $t(z)$ defined above.
$K(\vec{x}, \vec{z}) = (\vec{x}^T\vec{z})^2 = [(x_1, x_2)^T(z_1, z_2)]^2 = (z_1^2, \sqrt{2}z_1z_2, z_2^2) \cdot (x_1^2, \sqrt{2}x_1x_2, x_2^2)$

(f) Assume that when training the kernel perceptron algorithm, your hypothesis made three mistakes. One on z_1 and two on z_2. Write down the resulting weight vector $\vec{w} \in \Re^3$ that defines a hyperplane through the origin and is equivalent to the dual representation of the hypothesis learnt by running kernel perceptron algorithm above.

$$\begin{aligned}
\vec{w} &= \sum_{i \in M} y_i t(\vec{x_i}) \\
&= y_1 \times t(\vec{x_1}) + y_2 \times t(\vec{x_2}) + y_2 \times t(\vec{x_2}) \\
&= (1, \sqrt{2}, 1) - (1, 0, 0) - (1, 0, 0) = (-1, \sqrt{2}, 1)
\end{aligned}$$

On-Line Learning [25 points]

In this question, we will deal with a few on-line learning algorithms.

Let $D = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\}$, be a sequence of examples, where the j-th example $\mathbf{x}^{(j)}$ is associated with the label $y^{(j)} \in \{-1, +1\}$.

We wish to learn a weight vector $\mathbf{w}$ and a threshold θ so that an example $\mathbf{x}$ is positive if and only if:

$$\mathbf{w} \cdot \mathbf{x} + \theta \geq 0,$$

where $\mathbf{w} \in \mathbb{R}^n$, $\theta \in \mathbb{R}$.

(a) Write down the Hinge Loss function

$Loss(D, \mathbf{w}, \theta) =$ $\sum_{i=1}^{m} \max(0, 1 - y^{(i)}(\mathbf{w} \cdot \mathbf{x^{(i)}} + \theta))$

(b) Use the loss function you provided above to derive an update rule for the Stochastic Gradient Descent Algorithm.

Let R be the learning rate.

if $y^{(i)}(\mathbf{w} \cdot \mathbf{x^{(i)}} + \theta) < 1$

$\mathbf{w} \leftarrow \mathbf{w} + Ry^{(i)}\mathbf{x}^{(i)}$

$\theta \leftarrow \theta + Ry^{(i)}$

Otherwise: no update.

(c) What is the name of the algorithm that uses this update rule?

Perceptron (with margin)

(d) Write down the update rule of the Winnow Algorithm (Choose your favorite version of the algorithm).

Let α be the learning rate.

Set $\theta = -n$ and initialize $\mathbf{w}$ to be a vector with all ones.

For every $(\mathbf{x^{(i)}}, y^{(i)}) \in D$

if $y^{(i)}(\mathbf{w} \cdot \mathbf{x^{(i)}} + \theta) < 0$

$$w_j \leftarrow \begin{cases} w_j * \alpha, & x_j^{(i)} = 1 \text{ and } y^{(i)} = 1 \\ w_j / \alpha, & x_j^{(i)} = 1 \text{ and } y^{(i)} = -1 \\ w_j & \text{Otherwise.} \end{cases}.$$

(e) Assume now that we have a dataset labeled by a 3-DNF function. Is Winnow a PAC learning algorithms for the class of 3-DNF functions?

yes
___
{yes | no}

Explain:

3-DNF functions are not linearly separable functions over the original feature space. Therefore we cannot directly learn a consistent hypothesis using Winnow in the original feature space. However, if we consider a new feature space, where each feature is a conjunction containing up to 3 literals, then the dataset is linearly separable in the new feature space (why?), and Winnow becomes a consistent learner. Assume we have n literals total. The size of new feature space is $\binom{n}{3} \cdot 2^3 + \binom{n}{2} \cdot 2^2 + \binom{n}{1} \cdot 2^1 = O(n^3)$ which is polynomial. Also, the size of the hypothesis space is the size of the space of conjunctions over $O(n^3)$ terms and the log of this size is polynomial in n. Therefore, Winnow is a PAC learning algorithm for the class of 3-DNF functions.

Decision Trees [25 points]

Here is a collection S of different interesting animals. The goal is to use the available attributes to decide if an animal is a *Mammal (+)* or a *Bird (-)*. Note that the "note" column is only for notational convenience and as information for the readers; it has no bearing on the problem.

All computations in this problem are simple and only require the use of fractions; to see that, you should use the following identities and approximations in any computations.

$log(a \cdot b) = log(a) + log(b)$
$log(a/b) = log(a) - log(b)$
$log_2(3) \approx 3/2$
$log_2(5) \approx 11/5$

	Attributes						
	x_1	x_2	x_3	x_4	x_5	y	
#	Quacks	Lays Eggs	Flat-Bill	Biped	Aquatic	**Label**	(Note)
1	1	0	0	1	0	Mammal (+)	Human
2	1	1	1	1	1	Bird (-)	Duck
3	0	1	0	1	1	Bird (-)	Coot
4	0	1	1	0	1	Mammal (+)	Platypus
5	0	1	0	0	0	Mammal (+)	Echidna
6	0	0	0	0	0	Mammal (+)	Cow
7	0	1	0	1	0	Bird (-)	Emu
8	1	1	1	1	1	Bird (-)	Goose

(a) Calculate the entropy of the label y.

$H(y) = -(1/2)log_2(1/2) - (1/2)log_2(1/2) = 1$

(b) Calculate the *entropy* of the "Lays Eggs" attribute.
$H(x_2) = -(1/4)log_2(1/4) - (3/4)log_2(3/4) = 7/8$

(c) Compute $Gain(S, \text{Biped})$.
$H(y) - (5/8)H(y \mid x_4 = 1) - (3/8)H(y \mid x_4 = 0) = 5/8$

(d) Use the following information gain information

Gain(S, Quack) $\approx 1/20$
Gain(S, Lays Eggs) $\approx 1/3$
Gain(S, Flat-Bill) $\approx 1/20$
Gain(S, Aquatic) $\approx 1/8$
Gain(S, Biped) $\approx$???

along with the value you computed above, to choose the top node of the decision tree; continue to construct the minimal decision tree you can find that is consistent with the data. (There is no need to compute additional Information Gains).

```
If (biped):
    if !(lays eggs):
        Mammal
    else:
        Bird
else:
    Mammal
```

(e) Express the function Is-Mammal as a simple Boolean function over the features $\{x_1, x_2, ..., x_5\}$. That is, write down a simple Boolean function that is True on an example in the Table if and only if the example is a Mammal.

$\neg x_4 \vee \neg x_2$

(f) Express the Boolean function from (e) as a linear threshold function over the features $\{x_1, x_2, ..., x_5\}$. The answer in (d) is a disjunction. Therefore, it can be represented as

$(1 - x_2) + (1 - x_4) > 0 \Rightarrow -x_2 - x_4 + 2 > 0$