CS446: Machine Learning **Fall 2011**

Mid-term Exam Solution

October 25^{th}, 2011

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam. Note that there is an appendix with possibly useful formulae and computational shortcuts at the end.
- This exam booklet contains **five** problems, out of which you are expected to answer **four** problems of your choice.
- The exam ends at 10:45 AM. You have 75 minutes to earn a total of 100 points. You can earn 25 additional (bonus) points if you successfully attempt all five problems.
- If you choose to attempt all five problems, the four problems with the highest points will be considered for your mid-term score and the lowest will be considered as bonus.
- 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.**

The solutions are highlighted in red.

Good Luck!

Name (NetID): (1 Point)

Problem 1 (25 points):	
Problem 2 (25 points):	
Problem 3 (25 points):	
Problem 4 (25 points):	
Problem 5 (24 points):	
Total (100 points):	
Bonus (25 points):	

Problem 1: Decision Trees [25 points]

You are following the result of past eight matches Illini football team played, and analyze the performance based on a few features you felt were important.

#	Opponent	QtrBack	Fouls	Result
1	Weak	Strong	No	Win
2	Strong	Strong	Many	Loss
3	Strong	Weak	Many	Loss
4	Weak	Weak	Many	Loss
5	Strong	Weak	No	Win
6	Weak	Weak	Few	Win
7	Strong	Weak	Few	Loss
8	Strong	Strong	Few	Win

(a) **[4 points]** What is the entropy of the data set?

Solution:
Number of `Win` = Number of `Loss` = 4. So,

$$Entropy = -\sum_{i=1}^{2} \frac{4}{8} \log\left(\frac{4}{8}\right) = 1$$

(b) **[6 points]** What is the information gain if you split the dataset based on the attribute `Fouls`?

Solution:

$$\begin{aligned} Gain(S, \texttt{Fouls}) &= 1 - \frac{3}{8}\left[-\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)\right] \\ &\approx 1 - \frac{3}{8}\left[\frac{1}{3}\times\frac{3}{2} + \frac{2}{3}\times\frac{1}{2}\right] \\ &= 1 - \frac{3}{8}\left[\frac{1}{2} + \frac{1}{3}\right] = \frac{11}{16} \end{aligned}$$

(c) **[4 points]** We tell you that $Gain(S, \texttt{Opponent}) = Gain(S, \texttt{QtrBack}) = 0.05$. Based on your answer in (b) and this infromation, which attribute will you choose as the root node for the decision tree? Circle the appropriate option below.

i. `Opponent`
ii. `QtrBack`
iii. Either `Opponent` or `QtrBack`
iv. `Fouls`

Solution: (iv) `Fouls`, since $Gain(S, \texttt{Fouls})$ is the highest.

(d) **[6 points]** Build out a decision tree that it is *consistent* with the given data set.

Solution:

Either one of these two trees is acceptable.

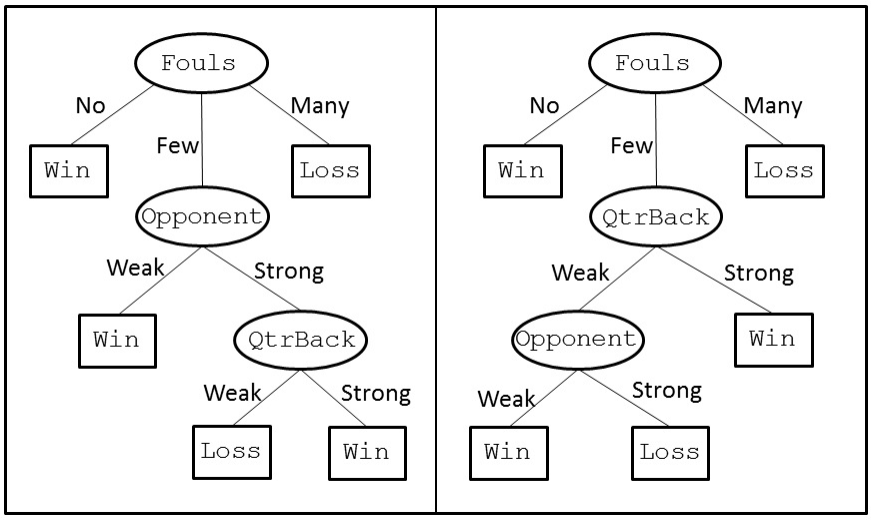


Figure 1: Two possible answers

(e) **[5 points]** Based on the decision tree you learned, what is the accuracy of your prediction on the following matches:

#	Opponent	QtrBack	Fouls	Result	Your Prediction
1	Strong	Strong	No	Win	Win
2	Weak	Strong	Few	Win	Win
3	Strong	Weak	missing	Win	Loss

Accuracy = 66.67%

Note: The key issue is how to handle the case when Fouls attribute is missing. One of the options is to consider all three values, weighing them based on how frequently they appear in the training set, and then take the majority vote. Counting this way, we get 2 votes for Win when Fouls is No, 3 votes for Loss when Fouls is Many, and another 3 votes for Loss when Fouls is Few. So, by majority voting, the prediction is Loss.

Problem 2: VC dimension [25 points]

Consider a concept space **H** of axis-parallel origin-centered embedded rectangles (see Figure 2). Formally, a concept $h \in \mathbf{H}$ is defined by four non-negative real parameters a,b,c,$d \in \mathbb{R}^+$, such that $a < b$ and $c < d$. An example $(x, y) \in \mathbb{R}^2$ is labeled `positive` if and only if (x, y) is in the rectangle $-b < x < b$ and $-d < y < d$, but not in the rectangle $-a < x < a$ and $-c < y < c$.

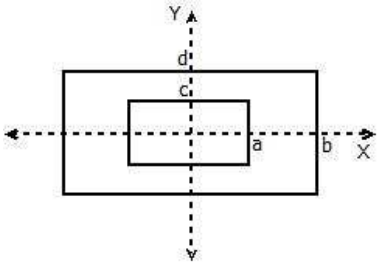
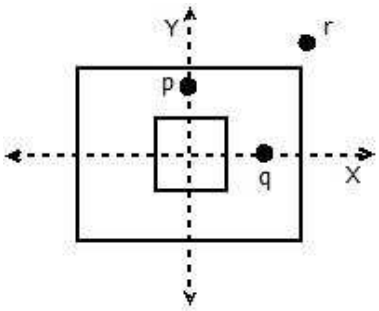


Figure 2: Sample concept in **H**

We will check if the following configuration of points can be shattered by concepts in **H**. Fill in the blanks below the figures.



(a) **[5 points]** Three points

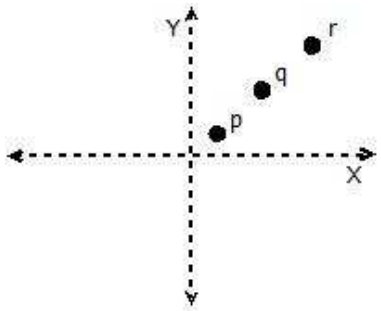
i. Need to check $\underset{\{\text{a number}\}}{2^3 = 8}$ cases.

ii. The points can be shattered : $\underset{\{\text{Yes} \mid \text{No}\}}{\text{Yes}}$

iii. If points can be shattered, give an example labeling you tried:
$\mathbf{p} = \underset{\{+ \mid -\}}{+}$ $\mathbf{q} = \underset{\{+ \mid -\}}{+}$ $\mathbf{r} = \underset{\{+ \mid -\}}{-}$
Draw a concept above that satisfies this labeling.

iv. If points cannot be shattered, justify.

v. Hence, VC(**H**) $\underset{\{\geq \mid < \mid =\}}{\geq}$ 3.



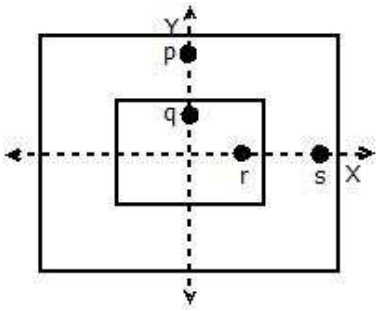
(b) **[3 points]** Three points (alternate case)

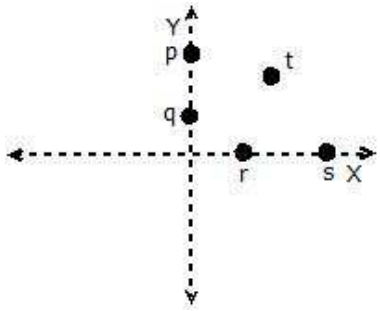
i. Plot three points **p**, **q**, **r** above such that they are not shattered by concepts in **H**.

ii. Show a labeling that justifies this claim:
$\mathbf{p} = \underset{\{+ \mid -\}}{+}$ $\mathbf{q} = \underset{\{+ \mid -\}}{-}$ $\mathbf{r} = \underset{\{+ \mid -\}}{+}$

iii. Based on the analysis so far on three points,

VC(**H**) $\underset{\{\geq \mid < \mid =\}}{\geq}$ 3.





(c) [**5 points**] Four points

i. Need to check $\underset{\{\text{a number}\}}{2^4 = 16}$ cases.

ii. The points can be shattered : $\underset{\{\text{Yes} \mid \text{No}\}}{\text{Yes}}$

iii. If points can be shattered, give an example labeling you tried:
$\mathbf{p} = \underset{\{+ \mid -\}}{+}$ $\mathbf{q} = \underset{\{+ \mid -\}}{-}$ $\mathbf{r} = \underset{\{+ \mid -\}}{-}$ $\mathbf{s} = \underset{\{+ \mid -\}}{+}$
Draw a concept above that satisfies this labeling.

iv. If points cannot be shattered, justify if you can or cannot shatter any configuration of 4 points.

v. Hence, VC($\mathbf{H}$) $\underset{\{\geq \mid < \mid =\}}{\geq}$ 4.

(d) [**10 points**] Five points

i. Need to check $\underset{\{\text{a number}\}}{2^5 = 32}$ cases.

ii. The points can be shattered : $\underset{\{\text{Yes} \mid \text{No}\}}{\text{No}}$

iii. If points can be shattered, give an example labeling you tried:
$\mathbf{p} = \underset{\{+ \mid -\}}{}$ $\mathbf{q} = \underset{\{+ \mid -\}}{}$ $\mathbf{r} = \underset{\{+ \mid -\}}{}$ $\mathbf{s} = \underset{\{+ \mid -\}}{}$ $\mathbf{t} = \underset{\{+ \mid -\}}{}$
Draw a concept above that satisfies this labeling.

iv. If points cannot be shattered, justify if you can or cannot shatter any configuration of 5 points. (You have more space below and on the next page, if you want.)
See proof below.

v. Hence, VC($\mathbf{H}$) $\underset{\{\geq \mid < \mid =\}}{<}$ 5.

We will prove that $\mathbf{H}$ cannot shatter any set of five points (i.e. VC($\mathbf{H}$) < 5).
Consider any 5 points $(x, y) \in \mathbb{R}^2$, and find the points with $\min |x|$, $\min |y|$, $\max |x|$, $\max |y|$. This way, we have selected at most 4 distinct points. If there are ties, just pick one. (It is okay if a single point satisfies more than one of the four boundary conditions above). Let these points be labeled `positive`. Any concept in $\mathbf{H}$ must have the embedded rectangle covering these four points in its interior. Now consider the remaining point. There exists at least one such point (x, y), such that $x_{\min} \leq x \leq x_{\max}$ and $y_{\min} \leq y \leq y_{\max}$. But this would mean (x, y) is also in the interior of the embedded rectangle. So, this point can not be labeled `negative` by any concept in $\mathbf{H}$. Since this is true for any five points, we proved that no concept in $\mathbf{H}$ can shatter five points.
Hence, VC($\mathbf{H}$) < 5.

(e) [**2 points**] Based on your answers in (a) to (d) above, VC($\mathbf{H}$) = 4
Why?

Solution:
We showed in parts (c) and (d) that VC($\mathbf{H}$) ≥ 4 and VC($\mathbf{H}$) < 5.
So, VC($\mathbf{H}$) $= 4$.

This page is intentionally left blank. You may continue your answer here.

Problem 3: Perceptrons and Kernels [25 points]

You are given a collection of m labeled documents $\{(\mathbf{d}_1, \ell_1), \ldots, (\mathbf{d}_m, \ell_m)\}$. The documents consist of words taken from a vocabulary of n words $\{w_1, \ldots, w_n\}$. Each document $\mathbf{d}$ is represented as a binary vector of words, $\mathbf{d} \in \{0, 1\}^n$, where the j^{th} component in $\mathbf{d}$ indicates whether the word w_j appears in document $\mathbf{d}$ ($\mathbf{d}[j] = 1$) or not ($\mathbf{d}[j] = 0$).

Each document is labeled as `relevant` ($\ell = 1$) or `irrelevant` ($\ell = -1$).

You have been asked to learn a Perceptron to classify a document as `relevant` or `irrelevant`. You want to include, in addition to the word features (whether or not w_j appears in a document), also features that indicate if a pair of words $\langle w_i, w_j \rangle$ appear anywhere in the document.

(a) **[4 points]** What is the size of the expanded feature space (as a function of the vocabulary size n)?

Solution:
The size of the expanded space is $n + \binom{n}{2}$

(b) **[5 points]** Assume a document $\mathbf{d}$ has only 4 words:

$$\mathbf{d} : \{w_5 = \texttt{machine}, w_6 = \texttt{learning}, w_{10} = \texttt{is}, w_{20} = \texttt{cool}\}$$

We denote by $\phi(\mathbf{d})$ the representation of $\mathbf{d}$ in the expanded space. How many of the features in $\phi(\mathbf{d})$ will have the value `1`?

Solution:
Number of features that will be `1` in $\phi(\mathbf{d})$ is $4 + \binom{4}{2} = 10$.

Write down the set of features that will have value `1` in $\phi(\mathbf{d})$.

Solution:
Let us define features f_i, $\forall i \in [1, n]$, each corresponding to one word w_i, such that $f_i = I[w_i \text{ is in the document}]$, i.e. f_i is 1 if w_i is in the document, and $f_i = 0$ otherwise. The features in $\phi(\mathbf{d})$ that will have value `1` are:

f_5, f_6, f_{10}, f_{20}, $f_5 \wedge f_6$, $f_5 \wedge f_{10}$, $f_5 \wedge f_{20}$, $f_6 \wedge f_{10}$, $f_6 \wedge f_{20}$, and $f_{10} \wedge f_{20}$.

(c) **[4 points]** Given two documents $\mathbf{d}$ and $\mathbf{e}$:

$$\mathbf{d} : \{w_5 = \texttt{machine}, w_6 = \texttt{learning}, w_{10} = \texttt{is}, w_{20} = \texttt{cool}\}$$

$$\mathbf{e} : \{w_6 = \texttt{learning}, w_9 = \texttt{theory}, w_{10} = \texttt{is}, w_{33} = \texttt{hard}\}$$

compute the dot product $\phi(\mathbf{d}) \cdot \phi(\mathbf{e})$

Solution:
Since there are two words common in $\mathbf{d}$ and $\mathbf{e}$, the number of features active in both $\phi(\mathbf{d})$ and $\phi(\mathbf{e})$ are $2 + \binom{2}{2} = 3$.

(d) **[4 points]** It is given that document $\mathbf{d}$ is `relevant` ($\ell_{\mathbf{d}} = 1$) and document $\mathbf{e}$ is `irrelevant` ($\ell_{\mathbf{e}} = -1$). You start learning the Perceptron in the ϕ space with weight vector $\mathbf{w} = \mathbf{0}$, and your Perceptron makes mistakes on both $\mathbf{d}$ and $\mathbf{e}$. What will $\mathbf{w}$ be after you have seen these two documents? Let learning rate $R = 1$.

Solution:

$$\begin{aligned} \text{Initially,} \quad & \mathbf{w} \leftarrow \mathbf{0} \\ \text{After document } \mathbf{d}, \quad & \mathbf{w} \leftarrow \mathbf{0} + \ell_{\mathbf{d}}\phi(\mathbf{d}) = \phi(\mathbf{d}) \\ \text{After document } \mathbf{e}, \quad & \mathbf{w} \leftarrow \phi(\mathbf{d}) + \ell_{\mathbf{e}}\phi(\mathbf{e}) = \phi(\mathbf{d}) - \phi(\mathbf{e}) \end{aligned}$$

Hence the final value of $\mathbf{w}$ after the two documents is $\phi(\mathbf{d}) - \phi(\mathbf{e})$.

(e) **[8 points]** We want to learn a Perceptron that behaves as if it operates in the ϕ space, but does not require that all the documents are expanded out. Design a kernel $K(\mathbf{d}_1, \mathbf{d}_2)$ that computes the value of the dot product $\phi(\mathbf{d}_1) \cdot \phi(\mathbf{d}_2)$ directly in the original space without expanding to the ϕ space.

Solution:
Many kernels are possible. One option is to use

$$K(\mathbf{d}_1, \mathbf{d}_2) = common(\mathbf{d}_1, \mathbf{d}_2) + \binom{common(\mathbf{d}_1, \mathbf{d}_2)}{2}$$

where $common(\mathbf{d}_1, \mathbf{d}_2)$ is the number of words that appear in both documents $\mathbf{d}_1$ and $\mathbf{d}_2$.

We can also use a polynomial kernel, if we also assume a bias (static) feature (corresponding to an empty word $\emptyset$) and features for pair of same words $\langle w_i, w_i \rangle$. See lecture slides for further details.

Problem 4: Boosting [25 points]

Consider that multiple rounds of AdaBoost are being run over m labeled training examples $\{(x_1, y_1), \ldots, (x_m, y_m)\}$. Let h_t be the hypothesis learned by the weak learner during the t^{th} round of AdaBoost. Let D_t be the distribution over which h_t was learned, and let $D_t(i)$ be the weight assigned to the i^{th} example in that distribution, D_t. ϵ_t is the error of h_t with respect to D_t, and AdaBoost uses ϵ_t to create the distribution D_{t+1} for the next round.

(a) **[5 points]** Write down the expression for ϵ_t in terms of $D_t(i)$, $h_t(x_i)$, y_i, and m.
Solution:
ϵ_t is given as:

$$\epsilon_t = \sum_{i=1}^{m} D_t(i) I[h_t(x_i) \neq y_i]$$

where $I[\cdot]$ is the indicator function that evaluates to 1 if its argument is true and 0 otherwise.

(b) **[4 points]** Assume that the initial distribution D_1 is uniform and that your first hypothesis h_1 labels 3 training examples incorrectly. What is the value of ϵ_1?
Solution: $\frac{3}{m}$

(c) **[8 points]** Recall that the normalization factor in round t, Z_t, is defined as

$$Z_t = \sum_{i=1}^{m} D_t(i) 2^{-\alpha_t y_i h_t(x_i)} \quad \text{and} \quad \alpha_t = \frac{1}{2} \log_2 \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$$

where ϵ_t is the error in round t. Prove that

$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

Solution:
First note that

$$2^{-\alpha_t} = \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \qquad \text{and} \qquad 2^{\alpha_t} = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$$

Using these, we have

$$\begin{aligned}
Z_t &= \sum_{i=1}^{m} D_t(i) 2^{-\alpha_t y_i h_t(x_i)} \\
&= \sum_{\substack{i=1 \\ h_t(x_i)=y_i}}^{m} D_t(i) 2^{-\alpha_t} + \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{m} D_t(i) 2^{\alpha_t} \\
&= 2^{-\alpha_t} \sum_{\substack{i=1 \\ h_t(x_i)=y_i}}^{m} D_t(i) + 2^{\alpha_t} \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{m} D_t(i) \\
&= \frac{\sqrt{\epsilon_t}}{\sqrt{1-\epsilon_t}}(1-\epsilon_t) + \frac{\sqrt{1-\epsilon_t}}{\sqrt{\epsilon_t}}(\epsilon_t) = 2\sqrt{\epsilon_t(1-\epsilon_t)}
\end{aligned}$$

(d) **[8 points]** Recall that D_{t+1} is given as

$$D_{t+1}(i) = \begin{cases} \frac{1}{Z_t}\, D_t(i)\, 2^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ \\ \frac{1}{Z_t}\, D_t(i)\, 2^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$

where Z_t is the normalization factor. Using (c), prove that the error of h_t with respect to D_{t+1}, i.e.

$$\sum_{i=1}^{m} D_{t+1}(i) I[h_t(x_i) \neq y_i] = \frac{1}{2}$$

where $I[\cdot]$ is the indicator function that evaluates to 1 if its argument is true and 0 otherwise.

Solution:

$$\begin{aligned} \sum_{i=1}^{m} D_{t+1}(i) I[h_t(x_i) \neq y_i] &= \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{m} D_{t+1}(i) \\ &= \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{m} \frac{1}{Z_t}\, D_t(i)\, 2^{\alpha_t} \\ &= \frac{1}{Z_t}\, 2^{\alpha_t} \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{m} D_t(i) \\ &= \frac{1}{2\sqrt{\epsilon_t}\sqrt{1-\epsilon_t}} \times \frac{\sqrt{1-\epsilon_t}}{\sqrt{\epsilon_t}} \times \epsilon_t = \frac{1}{2} \end{aligned}$$

Problem 5: Short Questions [24 points]

Please give a *brief* answer to each of the following questions.

(a) **[6 points]** Assume that all examples are in an n-dimensional Boolean space, $\mathbf{x} \in \{0,1\}^n$. An example is labeled `positive` if at least ℓ out of n variables ($1 \leq \ell \leq n$) are `1`, otherwise it is labeled `negative`. Write down a linear threshold function that is consistent with this labeling.

Solution:
The linear threshold function can be written as $sgn(\mathbf{w} \cdot \mathbf{x} - \theta)$, where $\mathbf{w} = [1, 1, \ldots, 1]$ and $\theta = \ell$.

(b) **[6 points]** You are given a data set of examples in an n-dimensional space, $\mathbf{x} \in \mathbb{R}^n$. The examples are labeled `positive` or `negative`, and the data set is consistent with labeling based on some linear threshold function $sgn(\mathbf{w} \cdot \mathbf{x} + \theta)$. Show that in an $(n+1)$-dimensional space, the data set is consistent with another linear threshold function that passes **through the origin** and is of **unit length**. Give the new representation for the instances ($\mathbf{x}'$) and weight vector ($\mathbf{w}'$).

Solution:
Writing $\mathbf{x}'$ and $\mathbf{w}'$ as column vectors, we have

$$\mathbf{x}' = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{w}' = \frac{1}{\sqrt{||\mathbf{w}||^2 + \theta^2}} \begin{bmatrix} \mathbf{w} \\ \theta \end{bmatrix}$$

(c) **[6 points]** Fill in the blanks with options given below:

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

A concept class $\mathbf{C}$ defined over the instance space $\mathbf{X}$ (with instances of length n) is ***PAC learnable*** by learner $\mathbf{L}$ using a hypothesis space $\mathbf{H}$ if

for all $f \in$ $\underset{\{\mathbf{C} \mid \mathbf{H}\}}{\mathbf{C}}$

for all distributions $\mathbf{D}$ over $\mathbf{X}$, and 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 $\underset{\{\text{at least} \mid \text{at most} \mid \text{equal to}\}}{\text{at least}}$ $\underset{\{\text{one of (a) to (f)}\}}{1-\delta}$

a hypothesis $g \in$ $\underset{\{\mathbf{C} \mid \mathbf{H}\}}{\mathbf{H}}$

with error $(\text{Error}_{\mathbf{D}} = \Pr_{\mathbf{D}}[f(x) \neq g(x)])$ $\underset{\{\text{at least} \mid \text{at most} \mid \text{equal to}\}}{\text{at most}}$ $\underset{\{\text{one of (a) to (f)}\}}{\epsilon}$

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

n, $1/\delta$, $1/\epsilon$, and size($\mathbf{C}$) .

{four of (a) to (j)}

(d) **[6 points]** Write down the update rule for Stochastic Gradient Descent.

Let $\mathbf{w} \in \mathbb{R}^n$ be the weight vector, $\mathbf{x}^{(i)}$ be the i^{th} sample from the training data, and $L(\mathbf{w}, \mathbf{x})$ be a loss function (assume that it is continuous and differentiable). We observe example $\mathbf{x}^{(i)}$ and predict using the current hypothesis $\mathbf{w}$. We then update the weight vector. Express the new weight vector as a function of $\mathbf{w}$, $L(\mathbf{w}, \mathbf{x}^{(i)})$, and the learning rate R.

$$\mathbf{w} \leftarrow \mathbf{w} + R \cdot \left[-\nabla L(\mathbf{w}, \mathbf{x}^{(i)})\right]$$

where $\nabla L(\mathbf{w}, \mathbf{x}^{(i)})$ is partial differential of $L(\mathbf{w}, \mathbf{x}^{(i)})$ with respect to $\mathbf{w}$.

Some formulae you may need

- $P(A, B) = P(A|B)P(B)$
- $Entropy(S) = -p^+ \log(p^+) - p^- \log(p^-) = -\sum_{i=1}^{k} p_i \log(p_i)$, where k is number of values
- $Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$
- $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$
- $\log_2(3) \approx \frac{3}{2}$