CIS519: Applied Machine Learning	**Fall 2019**
Homework 5	
Handed Out: December 2, 2019	*Due: December 9, 2019 at 11:59pm*

Version 1

- Feel free to talk to other members of the class in doing the homework. I am more concerned that you learn how to solve the problem than that you demonstrate that you solved it entirely on your own. You should, however, **write down your solution yourself**. Please include at the top of your document the list of people you consulted with in the course of working on the homework.
- While we encourage discussion within and outside the class, cheating and copying code is strictly not allowed. Copied code will result in the entire assignment being discarded at the very least.
- Please use Piazza if you have questions about the homework. Also, please come to the TAs recitations and to the office hours.
- Handwritten solutions are not allowed. All solutions must be typeset in Latex. Consult the class' website if you need guidance on using Latex. If you don't have a lot of experience with Latex (or even if you do), we recommend using Overleaf (`https://www.overleaf.com`) to write your solutions. You will submit your solutions as a single pdf file (in addition to the package with your code; see instructions in the body of the assignment).
- The homework is due at 11:59 PM on the due date. We will be using Gradescope for collecting the homework assignments. You should have been automatically added to Gradescope. If not, please ask a TA for assistance. Please do **not** hand in a hard copy of your write-up. Post on Piazza and contact the TAs if you are having technical difficulties in submitting the assignment.

1 Short Questions [20 Points]

(a) **(4 points)** Consider the following SVM formulation with the squared hinge-loss

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i [\max(0, 1 - y_i \mathbf{w}^\intercal \mathbf{x}_i)]^2$$

(1) What is i in the expression above. (Choose **one** of the following)
   (a) index of features of examples
   (b) index of examples in training data
   (c) index of examples which are support vectors
   (d) index of examples in test data

(2) One of the following optimization problems is equivalent to the one stated above. Which one? (Choose **one** of the following)
   (a)

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ s.t. \forall i \quad & y_i \mathbf{w}^\intercal \mathbf{x}_i \geq -\xi_i \\ \forall i \quad & \xi_i \geq 0 \end{aligned}$$

(b)

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i^2 \\ s.t. \forall i \quad & y_i \mathbf{w}^\intercal \mathbf{x}_i \geq 0 \\ \forall i \quad & \xi_i \geq 0 \end{aligned}$$

(c)

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ s.t. \forall i \quad & y_i \mathbf{w}^\intercal \mathbf{x}_i \geq 1 - \xi_i \\ \forall i \quad & \xi_i \geq 0 \end{aligned}$$

(d)

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i^2 \\ s.t. \forall i \quad & y_i \mathbf{w}^\intercal \mathbf{x}_i \geq 1 - \xi_i \\ \forall i \quad & \xi_i \geq 0 \end{aligned}$$

(b) **(9 points)** Consider the instance space $X = \{1, 2, \ldots N\}$ of natural numbers. An *arithmetic progression* over X is a subset of X of the form $X \cap \{a + bi : i = 0, 1, 2, \ldots\}$ where a and b are natural numbers. For instance, for $N = 40$ the set $\{9; 16; 23; 30; 37\}$ is an arithmetic progression over $\{1, 2, \ldots N\}$ with $a = 9, b = 7$.

Let C be the concept class consisting of *all arithmetic progressions over* $\{1, 2, \ldots N\}$. That is, each function $c_{a,b} \in C$ takes as input a natural number and says 'yes' if this natural number is in the $\{a, b\}$ arithmetic progression, 'no' otherwise.

A neural network proponent argues that it is necessary to use a deep neural networks to learn functions in C. You decide to study how expressive the class C is in order to determine if this argument is correct.

(1) **(2 points)** Determine the order of magnitude of the VC dimension of C (Circle one of the options below).

(A) $O(\log N)$ (B) $O(N)$ (C) $O(N^2)$ (D) ∞

(2) **(5 points)** Give a brief justification for the answer above.

(3) **(2 points)** Use your answer to (a) to settle the argument with the deep network proponent. (Circle on of the options below)

(A) Deep networks are needed (B) Simpler Classifiers are sufficient

(c) **(7 points)** We showed in class that the training error of the hypothesis h generated by the AdaBoost algorithm is bounded by

$$e^{-2\gamma^2 T},$$

where T is the number of rounds and γ is the advantage the weak learner has over chance.

Assume that you have a weak learner with an advantage of at least 1/4. (That is, the error of the weak learner is below 1/4.) You run AdaBoost on a dataset of $m = e^{14} (\approx 10^6)$ examples, with the following tweak in the algorithm:

> Instead of having the algorithm loop from $t = 1, 2, \ldots, T$, you run the loop with the following condition: If the Adaboost classifier h misclassifies at least one example, do another iteration of the loop.

That is, you run AdaBoost until your hypothesis is consistent with your m examples.

Question: Will your algorithm run forever, or can you guarantee that it will halt after some number of iterations of the loop?

(1) Choose one of the following options (**2 points**):

(A) Run forever (B) Halt after some number of iterations

(2) Justify (**5 points**): If you think it may run forever, explain why. If you think it will halt after some number of iterations of the loop, give the best *numeric* bound you can on the maximum # of iterations of the loop that may be executed. (You don't need a calculator here).

2 Graphical Models [20 Points]

We consider a probability distribution over 4 Boolean variables, A, B, C, D.

(a) **(2 points)** What is the largest number of independent parameters needed to define a probability distribution over these 4 variables? (Circle one of the options below).

(A) 3 (B) 4 (C) 15 (D) 16

In the rest of this problem we will consider the following Bayesian Network representation of the distribution over the 4 variables.

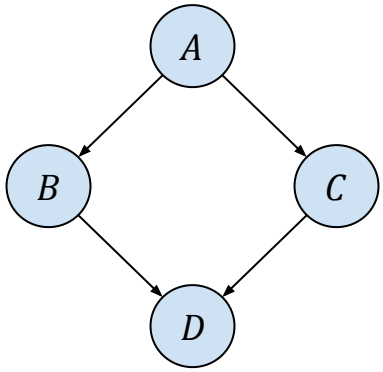


Figure 1: A Directed Graphical model over A, B, C, D

In all computations below you can leave the answers as fractions.

(b) **(2 points)** Write down the joint probability distribution given the graphical model depicted in Figure 1.

$P(A, B, C, D,) = \ldots$

(c) **(4 points)** What are the parameters you need to estimate in order to completely define this probability distribution? Provide as answers the indices of these parameters from the table below.

Answer:

(1) $P[A = 1]$	(2) $P[B = 1]$	(3) $P[C = 1]$
(4) $P[D = 1]$	(5) $P[A = 1\|B = i], i \in \{0, 1\}$	(6) $P[A = 1\|C = i], i \in \{0, 1\}$
(7) $P[B = 1\|A = i], i \in \{0, 1\}$	(8) $P[B = 1\|D = i], i \in \{0, 1\}$	(9) $P[C = 1\|A = i], i \in \{0, 1\}$
(10) $P[B = i\|A = 1], i \in \{0, 1\}$	(11) $P[B = i\|D = 1], i \in \{0, 1\}$	(12) $P[C = i\|A = 1], i \in \{0, 1\}$
(13) $P[C = 1\|D = i], i \in \{0, 1\}$	(14) $P[D = 1\|B = i, C = j], i, j \in \{0, 1\}$	(15) $P[D = 1\|C = i], i \in \{0, 1\}$

(d) **(4 points)** You are given the following sample S of 10 data points:

Example	1	2	3	4	5	6	7	8	9	10
A	1	1	0	1	1	1	0	0	0	1
B	0	0	1	1	0	0	0	1	1	1
C	0	1	0	1	0	1	1	0	1	1
D	0	0	0	1	1	1	0	1	1	0

Use the given data to estimate the most likely parameters that are needed to define the model (those you have chosen in part (3) above).

(e) **(4 points)** Compute the likelihood of the following set of independent data points given the Bayesian network depicted in Figure 1, with the parameters you computed in (d) above. (You don't need a calculator; use fractions; derive a numerical answer).

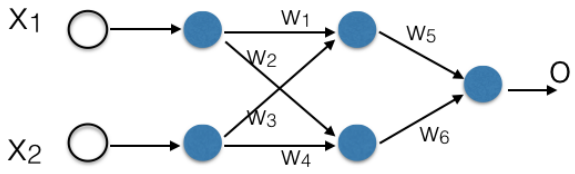
Example	A	B	C	D
1	1	0	1	0
2	0	1	0	1

(f) **(4 points)** We are still using the Bayesian network depicted in Figure 1, with the parameters you computed in (d) above.

If we know that $A = 1$, what is the probability of $D = 1$? (Write the expressions needed and keep your computation in fractions; derive a numerical answer).

3 Neural Networks [20 Points]

Consider a simple neural network shown in the diagram below.



It takes as input a 2-dimensional input vector of real numbers x $= [x_1, x_2]$, and computes the output o using the function $NN(x_1, x_2)$ as follows:

$$o = NN(x_1, x_2)$$

where

$$NN(x_1, x_2) = \sigma\left(w_5\sigma(w_1x_1 + w_3x_2 + b_{13}) + w_6\sigma(w_2x_1 + w_4x_2 + b_{24}) + b_{56}\right)$$

where $w_i \in \mathbb{R}$, $b_{ij} \in \mathbb{R}$, and $\sigma(x) = \frac{1}{1+e^{-x}}$. We want to learn the parameters of this neural network for a prediction task.

(a) **(2 points)** What is the range of values the output o can have ? (Circle one of the following options).

(A) $\mathbb{R}$ (B) $[0, -\infty)$ (C) $(0, 1)$ (D) $(-1, 1)$

(b) **(4 points)** We were informed that **the cross entropy loss** is a good loss function for our prediction task. The cross entropy loss function has the following form:

$$L(y, \hat{y}) = -y \ln \hat{y} - (1 - y) \ln(1 - \hat{y})$$

where y is the desired output, and $\hat{y}$ is the output of our predictor. First confirm that this is indeed a loss function, by evaluating it for different values of y and $\hat{y}$. Select from the following options **all** that are correct.

(a) $\lim_{\hat{y}\to 0^+} L(y=0,\hat{y}) = 0$ (b) $\lim_{\hat{y}\to 0^+} L(y=0,\hat{y}) = \infty$

(c) $\lim_{\hat{y}\to 0^+} L(y=1,\hat{y}) = 0$ (d) $\lim_{\hat{y}\to 1} L(y=0,\hat{y}) = \infty$

(c) **(2 points)** We are given m labeled examples $\{(\mathrm{x}^{(1)}, y^{(1)}), (\mathrm{x}^{(2)}, y^{(2)}), \ldots, (\mathrm{x}^{(m)}, y^{(m)})\}$, where $\mathrm{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}]$. We want to learn the parameters of the neural network by minimizing the cross entropy loss error over these m examples. We denote this error by Err. Which of the following options is a correct expression for Err ?

(a) $Err = \frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - NN(x_1^{(i)}, x_2^{(i)})\right)^2$

(b) $Err = \frac{1}{m}\sum_{i=1}^{m} -y^{(i)} \ln\left(NN(x_1^{(i)}, x_2^{(i)})\right) - (1-y^{(i)}) \ln\left(1 - NN(x_1^{(i)}, x_2^{(i)})\right)$

(c) $Err = \frac{1}{m}\sum_{i=1}^{m}\left(NN(x_1^{(i)}, x_2^{(i)}) - y^{(i)}\right)^2$

(d) $Err = \frac{1}{m}\sum_{i=1}^{m} -NN(x_1^{(i)}, x_2^{(i)}) \ln y^{(i)} - \left(1 - NN(x_1^{(i)}, x_2^{(i)})\right) \ln(1-y^{(i)})$

(d) **(7 points)** We will use gradient descent to minimize Err. We will only focus on two parameters of the neural network : w_1 and w_5. Compute the following partial derivatives:

(i) $\dfrac{\partial Err}{\partial w_5}$ (ii) $\dfrac{\partial Err}{\partial w_1}$

Hints: You might want to use the chain rule. For example, for (ii):

$$\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial net_{56}} \cdot \frac{\partial net_{56}}{\partial o_{13}} \cdot \frac{\partial o_{13}}{\partial net_{13}} \cdot \frac{\partial net_{13}}{\partial w_1}$$

where E is the error on a single example and the notation used is:

$net_{13} = w_1 x_1 + w_3 x_2,$
$o_{13} = \sigma(net_{13} + b_{13}),$
$net_{24} = w_2 x_1 + w_4 x_2,$
$o_{24} = \sigma(net_{24} + b_{24}),$
$net_{56} = w_5 o_{13} + w_6 o_{24},$
$o = \sigma(net_{56} + b_{56}).$

In both (i) and (ii) you have to compute all intermediate derivatives that are needed, and simplify the expressions as much as possible; you may want to consult the formulas in the Appendix on the final page. Eventually, write your solution in terms of the notation above, x_is and w_is.

(e) **(3 points)** Write down the update rules for w_1 and w_5, using the derivatives computed in the last part.

(f) **(2 points)** Based on the value of o predicted by the neural network, we have to say either 'YES' or 'NO'. We will say 'YES' if $o > 0.5$, else we will say 'NO'. Which of the following is true about this decision function ?

(a) It is a linear function of the inputs (b) It is a non-linear function of the inputs

4 Naive Bayes

We have four random variables: a label, $Y \in \{0, 1\}$, and features $X_i \in \{0, 1\}, i \in \{1, 2, 3\}$. Give m observations $\{(y, x_1, x_2, x_3)_1^m\}$ we would like to learn to predict the value of the label y on an instance (x_1, x_2, x_3). We will do this using a Naive Bayes classifier.

(a) **[3 points]** Which of the following is the Naive Bayes assumption? (Circle one of the options below.)

(a) $\Pr(x_i, x_j) = Pr(x_i)Pr(x_j) \quad \forall i, j = 1, 2, 3$
(b) $\Pr(x_i, x_j|y) = Pr(x_i|y)P(x_j|y) \quad \forall i, j = 1, 2, 3$
(c) $\Pr(y|x_i, x_j) = Pr(y|x_i)Pr(y|x_j) \quad \forall i, j = 1, 2, 3$
(d) $\Pr(x_i|x_j) = Pr(x_j|x_i) \quad \forall i, j = 1, 2, 3$

(b) **[3 points]** Which of the following equalities is an outcome of these assumptions? (Circle one and only one.)

(a) $\Pr(y, x_1, x_2, x_3) = \Pr(y) \Pr(x_1|y) \Pr(x_2|x_1) \Pr(x_3|x_2)$
(b) $\Pr(y, x_1, x_2, x_3) = \Pr(y) \Pr(x_1|y) \Pr(x_2|y) \Pr(x_3|y)$
(c) $\Pr(y, x_1, x_2, x_3) = \Pr(y) \Pr(y|x_1) \Pr(y|x_2) \Pr(y|x_3)$
(d) $\Pr(y, x_1, x_2, x_3) = \Pr(y|x_1, x_2, x_3) \Pr(x_1) \Pr(x_2) \Pr(x_3)$

(c) **[3 points]** Circle **all** (and only) the parameters from the table below that you will need to use in order to *completely* define the model. You may assume that $i \in \{1, 2, 3\}$ for all entries in the table.

(1) $\alpha_i = \Pr(X_i = 1)$	(6) $\beta = \Pr(Y = 1)$
(2) $\gamma_i = \Pr(X_i = 0)$	(7) $p_i = \Pr(X_i = 1 \mid Y = 1)$
(3) $s_i = \Pr(Y = 0 \mid X_i = 1)$	(8) $q_i = \Pr(Y = 1 \mid X_i = 1)$
(4) $t_i = \Pr(X_i = 1 \mid Y = 0)$	(9) $u_i = \Pr(Y = 1 \mid X_i = 0)$
(5) $v_i = \Pr(Y = 0 \mid X_i = 0)$	

(d) **[5 points]** Write an **algebraic** expression for the naive Bayes classifier in terms of the model parameters chosen in (c).
(Please note: by "algebraic" we mean, **no** use of words in the condition).
Predict $y = 1$ iff _____.

(e) **[6 points]** Use the data in table 1 below and the naive Bayes model defined above on (Y, X_1, X_2, X_3) to compute the following probabilities. Use Laplace Smoothing in all your parameter estimations.

Table 1: Data for this problem.

#	y	x_1	x_2	x_3
1	0	1	0	1
2	0	1	1	0
3	1	1	0	0
4	1	1	0	0
5	1	0	1	0
6	1	0	0	0
7	1	1	1	0
8	1	0	0	0

(1) Estimate directly from the data (with Laplace Smoothing):

i. $P(X_1 = 1|Y = 0) =$

ii. $P(Y = 1|X_2 = 1) =$

iii. $P(X_3 = 1|Y = 1) =$

(2) Estimate the following probabilities using the naive Bayes model. Use the parameters identified in part (c), estimated with Laplace smoothing. Provide your work.

i. $P(X_3 = 1, Y = 1) =$

ii. $P(X_3 = 1) =$

5 Tree-Dependent Distributions

A tree dependent distribution is a probability distribution over n variables, $\{x_1, \ldots, x_n\}$ that can be represented as a tree built over n nodes corresponding to the variables. If there is a directed edge from variable x_i to variable x_j, then x_i is said to be the parent of x_j. Each directed edge $\langle x_i, x_j \rangle$ has a weight that indicates the conditional probability $\Pr(x_j|x_i)$. In addition, we also have probability $\Pr(x_r)$ associated with the root node x_r. While computing joint probabilities over tree-dependent distributions, we assume that a node is independent of all its non-descendants given its parent. For instance, in our example above, x_j is independent of all its non-descendants given x_i.

To learn a tree-dependent distribution, we need to learn three things: the structure of the tree, the conditional probabilities on the edges of the tree, and the probabilities on the nodes. Assume that you have an algorithm to learn an *undirected* tree T with all required probabilities. To clarify, for all *undirected* edges $\langle x_i, x_j \rangle$, we have learned both probabilities, $\Pr(x_i|x_j)$ and $\Pr(x_j|x_i)$. (There exists such an algorithm and we will be covering that in class.) The only aspect missing is the directionality of edges to convert this undirected tree to a directed one.

However, it is okay to not learn the directionality of the edges explicitly. In this problem, you will show that choosing any arbitrary node as the root and directing all edges away from it is sufficient, and that two directed trees obtained this way from the same underlying undirected tree T are equivalent.

1. **[7 points]** State exactly what is meant by the statement: "*The two directed trees obtained from T are equivalent.*"
2. **[13 points]** Show that no matter which node in T is chosen as the root for the "direction" stage, the resulting directed trees are all equivalent (based on your definition above).

Appendix

1. **Losses and Derivatives**
   (a) sigmoid(x) $= \sigma(x) = \dfrac{1}{1 + exp^{-x}}$
   (b) $\dfrac{\partial}{\partial x}\sigma(x) = \sigma(x)\Big(1 - \sigma(x)\Big)$
   (c) ReLU(x) $= max(0, x)$
   (d) $\dfrac{\partial}{\partial x} ReLU(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$
   (e) $\tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$

(f) $\frac{\partial}{\partial x} tanh(x) = 1 - tanh^2(x)$

(g) Zero-One loss(y, y^*) $= \begin{cases} 1, & \text{if } y \neq y^* \\ 0, & \text{if } y = y^* \end{cases}$

(h) Hinge loss(w, x, b, y*) $= \begin{cases} 1 - y^*(w^T x + b), & \text{if } y^*(w^T x + b) < 1 \\ 0, & \text{otherwise} \end{cases}$

(i) $\frac{\partial}{\partial w}$ Hinge loss(w, x, b, y*) $= \begin{cases} -y^*(x), & \text{if } y^*(w^T x + b) < 1 \\ 0, & \text{otherwise} \end{cases}$

(j) Squared loss(w, x, y^*) $= \frac{1}{2}(w^T x - y^*)^2$

(k) $\frac{\partial}{\partial w}$ Squared loss$(w, x, y^*) = x(w^T x - y^*)$

2. **Other derivatives**
   (a) $\frac{d}{dx} \ln(x) = \frac{1}{x}$
   (b) $\frac{d}{dx} x^2 = 2x$
   (c) $\frac{d}{dx} f(g(x)) = \frac{d}{dg} f(g(x)) \frac{d}{dx} g(x)$
3. Logarithm rules: Use the following log rules and approximations for computation purposes.
   (a) $\log(a \cdot b) = \log(a) + \log(b)$
   (b) $\log(\frac{a}{b}) = \log(a) - \log(b)$
   (c) $\log_2(1) = 0$
   (d) $\log_2(2) = 1$
   (e) $\log_2(4) = 2$
   (f) $\log_2(3/4) \approx 3/2 - 2 = -1/2$;
   (g) $\log_2(3) \approx \frac{3}{2} \approx 1.5$
4. sgn(x) $= \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{if } x < 0 \end{cases}$
5. for $x = (x_1, x_2, \ldots x_n) \in R^n$, $L_2(x) = ||x||_2 = \sqrt{\sum_1^n x_i^2}$

Submission Instructions

We will be using Gradescope to turn in both the Python code and writeup pdfs. You should have been automatically added to Gradescope. If you do not have access, please ask the TA

staff on Piazza.

For this homework assignment, there is one Gradescope assignment where you should upload your writeup as a PDF: "Homework 5."