CIS 419/519: Applied Machine Learning **Fall 2020**

Homework 5

Handed Out: December 3, 2020 *Due: December 10, 2020 at 11:59pm*

- 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) **(6 points)** Multiple Choice Questions

(1) **(2 points)** You are training a binary classifier which predicts whether a test taken by a patient indicates if they have a rare disease (True) or not (False). Which one of the following performance measures would you like to optimize?

(a) Precision, because it is important not to have many false negative examples

(b) Recall, because it is important not to have many false positive examples

(c) Accuracy, because it is important to know how many correct predictions

(d) Recall, because it is important not to have many false negative examples

(2) **(2 points)** Suppose you have a feature space consisting of 50 boolean variables, $x_1, x_2, ..., x_{50}$. You are trying to learn the following function over these variables: $y = x_2 \lor x_{10} \lor x_{15} \lor x_{22} \lor x_{40} \lor x_{42} \lor x_{45}$. What is the minimum number of training examples you will have to see in order to learn this function?

(a) 45 (b) 7 (c) 8 (d) 50

(3) **(2 points)** Recall the mathematical representation of SVM is:

$$\textbf{min}\ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{j=1}^{m}\xi_i \tag{1}$$

$$\text{s.t } y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)} + \theta) \geq 1 - \xi_i, \xi_i \geq 0, \forall(\mathbf{x}^{(i)}, y^{(i)}) \in D \tag{2}$$

Select all of the statements that are true below.

(a) The cost parameter C in the SVM means the balance between simplicity and overfitting of the model

(b) SVM can apply an kernel tricks by extending Equation (2) to:

$$\text{s.t } y^{(i)}\phi(\mathbf{w}\cdot\mathbf{x}^{(i)}+\theta)\geq 1-\xi_i, \xi_i\geq 0, \forall(\mathbf{x}^{(i)}, y^{(i)})\in D$$

(c) SVM can't be used for Image Classification

(d) SVM is robust under noisy datasets with many outliers and overlapping

(b) **(5 points)** Adaboost Review

(1) **(2 points)** The performance of Adaboost depends on which of the following assumptions about the weighted error ϵ_i of the i^{th} hypothesis h_i?

(a) $\epsilon_i < \sigma$, $\forall\sigma > 0$

(b) $\epsilon_i < \frac{1}{2} - \gamma_i$, where $\gamma_i$is a positive constant

(c) $\epsilon_i < \frac{1}{2}$

(d) none of above

(2) **(3 points)** Assume that the i^{th} hypothesis h_i makes mistakes on the set of examples x_1, x_2, x_3, x_4 and h_{i+1} makes mistakes on exactly the same set of examples. What is the relationship between ϵ_i and ϵ_{i+1} $(<, >, =)$? Explain why.

(c) **(9 points)** Consider the hypothesis space consisting of all circles in the plane $\mathbb{R}^2$. Each hypothesis is an individual circle, and any point inside the circle is considered to have a positive label.

(1) **(2 points)** Show that this set of classifiers can shatter a set of three points (please submit the necessary illustrations for this in your latex document)

(2) **(5 points)** Can this set of classifiers shatter a set of 4 points? If not, explain.(Hint: try consider all different possible scenarios)

(3) **(2 points)** What is the VC dimension of this set of classifiers, and why?

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.

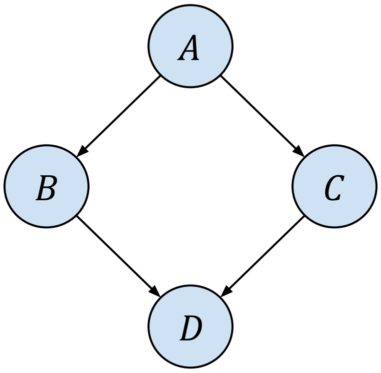


Figure 1: A Directed Graphical model over A, B, C, D

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.
$P(A, B, C, D,) =$

(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.

(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).

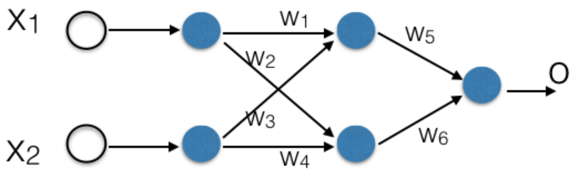
(f) **(4 points)** We are still using the Bayesian network depicted in Figure 1, with the parameters you computed in (d) above.

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).

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

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:

$$\text{(i)} \quad \frac{\partial Err}{\partial w_5} \qquad\qquad \text{(ii)} \quad \frac{\partial Err}{\partial w_1}$$

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:

$net_{13} = w_1x_1 + w_3x_2,$
$o_{13} = \sigma(net_{13} + b_{13}),$
$net_{24} = w_2x_1 + w_4x_2,$
$o_{24} = \sigma(net_{24} + b_{24}),$
$net_{56} = w_5o_{13} + w_6o_{24},$
$o = \sigma(net_{56} + b_{56}).$

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[20 Points]

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 $y = 1$ if ____________________________.

(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.

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

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

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

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

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

(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. $P(X_3 = 1) =$

Submission Instructions

We will be using Gradescope to turn in 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.”

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

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

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

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

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

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

5. for $x = (x_1, x_2, \dots x_n) \in R^n$, $L_2(x) = ||x||_2 = \sqrt{\sum_1^n x_i^2}$