University of Illinois Spring 2018

CS 446: Machine Learning
Midterm

Tuesday, March 13, 2018, 6:00pm to 7:30pm Central Time

Name: (in **BLOCK CAPITALS**) NAMAN SHUKLA

NetID: namans2

Signature: Naman.

Instructions

This exam is closed book except that one 8.5" ×11" sheet of notes is permitted: both sides may be used. Calculators, laptop computers, PDAs, iPods, cellphones, e-mail pagers, headphones, etc. are not allowed. Remove everything beyond a pen, the 8.5" ×11" sheet of notes, and your ID from the desk.

The exam consists of 6 problems worth a total of 59 points. The problems are not weighted equally, so it is best for you to pace yourself accordingly. Write your answers in the spaces provided, and reduce common fractions and expressions to lowest terms, but DO NOT convert them to decimal fractions (for example, write $\frac{3}{4}$ instead of $\frac{24}{32}$ or 0.75).

SHOW YOUR WORK; WRITE YOUR ANSWERS IN PROVIDED BOXES. Answers without appropriate justification will receive very little credit. Answers outside the provided boxes will receive very little credit. If you need extra space, use the back of the previous page and clearly mark what problem the solution corresponds to. Ambiguous answers will receive very little credit.

Grading

1. 9 points	9
2. 10 points	8
3. 10 points	4
4. 7 points	5
5. 13 points	13
6. 10 points	9
Total (59 points)	48

Scanned by CamScanner

1. **[9 points]** Regression

(a) Consider the following dataset $\mathcal{D}$ in the one-dimensional space.

i	$x^{(i)}$	$y^{(i)}$
1	0	-1
2	1	2
3	1	0

Table 1: Data for $\mathcal{D}$

For a set of observations $\{(y^{(i)}, x^{(i)})\}$, where $\{(y^{(i)}, x^{(i)})\} \in \mathbb{R}$ and $i \in \{1, 2, \ldots, |\mathcal{D}|\}$, we optimize the following program.

$$\underset{w_1, w_2}{\operatorname{argmin}} \sum_{(y^{(i)}, x^{(i)}) \in \mathcal{D}} (y^{(i)} - w_1 \cdot x^{(i)} - w_2)^2 \tag{1}$$

Find the optimal w_1^*, w_2^* given the aforementioned dataset $\mathcal{D}$ and justify your answer. **Compute the scalars w_1^* and w_2^*.**

Your answer:

$w_1^* = 2$

$w_2^* = -1$

2

by setting $\nabla f_{w_1} = 0 \rightarrow w_2 = -1$

$\nabla f_{w_2} = 0 \rightarrow w_1 = 2$

$(-1 - w_2)^2 \rightarrow \cancel{(0)}$

$(2 - w_1 - w_2)^2 \rightarrow \cancel{(3 - w_1)^2}$

$(-w_1 - w_2)^2 \rightarrow \cancel{(1 - w_1)^2}$

(b) What is the minimum number of observations that are required to obtain a unique solution for the program in Eq. (1)?

Your answer: 2

1

$(-1 - w_2)^2 + (2 - w_1 - w_2)^2 + (-w_1 - w_2)^2$

∇_{w_2}: $-2(-1 - w_2) - 2(2 - w_1 - w_2) - (w_1 - w_2) = 0$

∇_{w_1}: $+ (2 - w_1 - w_2) + (w_1 - w_2) = 0$ $\quad \boxed{w_1 = -1}$

$(3 - w_1) + (1 - w_1) = 0$

$\boxed{\frac{1}{2} = w_1}$

Scanned by CamScanner

(c) Consider another dataset $\mathcal{D}_1$, where $x^{(i)}, y^{(i)} \in \mathbb{R}$

i	$x^{(i)}$	$y^{(i)}$
1	0	0
2	1	1
3	2	4
4	3	9
5	4	16

Table 2: Data for $\mathcal{D}_1$

Clearly $\mathcal{D}_1$ can not be fit exactly with a linear model. In class, we discussed a simple approach of building a nonlinear model while still using our linear regression tools. How would you use the linear regression tools to obtain a nonlinear model which better fits $\mathcal{D}_1$, *i.e.*, what feature transform would you use? Provide your reasons and write down the resulting program that you would optimize using a notation which follows Eq. (1), *i.e.*, make all the trainable parameters explicit. **Do NOT plug the datapoints from $\mathcal{D}_1$ into your program and solve for its parameters. Just provide the program.**

Your answer: As y varies with square of x. We should use $\phi(x) = x^2$ with loss function as

$$\frac{1}{2}(y - w^T\phi(x))^2$$

2

(d) Write down a program equivalent to the one derived in part (c) using matrix-vector notation. Carefully define the matrices and vectors which you use, their dimensions and their entries. Show how you fill the matrices and vectors with the data. Derive the closed form solution for this program using the symbols which you introduced. **Do NOT compute the solution numerically.**

$x^T(wx-y)=0$

Your answer:

$$\phi(x) = \begin{bmatrix} 0 & 1 \\ 1^2 & 1 \\ 2^2 & 1 \\ 3^2 & 1 \\ 4^2 & 1 \end{bmatrix}$$

$w = [w \; b]$

$y = [0 \; 1 \; 4 \; 9 \; 16]$

then $w\phi^T(x) =$

$$[w \; b]\begin{bmatrix} 0 & 1^2 & 2^2 & 3^2 & 4^2 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

OLS sol $\Rightarrow (\phi(x)^T\phi(x))^{-1}\phi^T(x)\, y$

3

(e) Briefly describe the problem(s) that we will encounter if we were to fit a very high degree polynomial to the dataset $\mathcal{D}_1$?

Your answer:

we might overfit the data

✓

1

Scanned by CamScanner

2. **[10 points]** Binary Classifiers

(a) Assume $y \in \{-1, 1\}$. Consider the following program for linear regression:

$$\min_{\mathbf{w}} f(\mathbf{w}) := \frac{1}{2} \sum_i \left(1 - y^{(i)} \mathbf{w}^T \mathbf{x}^{(i)}\right)^2$$

Is the objective function $f(\mathbf{w})$ convex in $\mathbf{w}$ assuming everything else given and fixed? (Yes or No)

Your answer: Yes.

1 ✓

(b) Assume $y \in \{-1, 1\}$. Consider the following program for logistic regression:

$$\min_{\mathbf{w}} f(\mathbf{w}) := \sum_i \log\left(1 + \exp(-y^{(i)} \mathbf{w}^T \mathbf{x}^{(i)})\right)$$

Is the objective function $f(\mathbf{w})$ convex in $\mathbf{w}$ assuming everything else given and fixed? (Yes or No)

Your answer: Yes.

2 ✓

(c) We want to use gradient descent to address the above **logistic regression** program. What is the gradient $\nabla_{\mathbf{w}} f(\mathbf{w})$? Use the symbols and notation which was used in the cost function.

Your answer:

$$\nabla_w f(w) = \sum \frac{(-y^{(i)} w^T x^{(i)}) \exp(-y^{(i)} w^T x^{(i)}) . (\cancel{y^{(i)} x^{(i)}})}{1+\exp(-y^{(i)} w^T x^{(i)})}$$

4 ✗ ✓

(d) What is the probability model assumed for logistic regression and linear regression? Give names rather than equations.

Your answer: ~~Logit function~~ sigmoid function.

$$P\left(\frac{y_i = 1}{x_i}\right) = \frac{1}{1+\exp(-w^T x_i)}$$

1 ✓

Scanned by CamScanner

3. **[10 points]** Support Vector Machine

(a) Recall, a hard-margin support vector machine in the primal form optimizes the following program

$$\min_{\mathbf{w},b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{s.t.} \;\; y^{(i)}(\mathbf{w}^\intercal \mathbf{x}^{(i)} + b) \geq 1 \;\; ,\forall(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D} \tag{2}$$

What is the Lagrangian, $L(\mathbf{w}, \alpha)$, of the constrained optimization problem in Eq. (2)?

Your answer:

$$L(\underset{b}{w,\alpha}) = \frac{1}{8}\|w\|^2 + \sum_{i\in D} \alpha^{(i)}(1 - y^{(i)}(w^T x^{(i)} + b)) \qquad \alpha \geq 0$$

1

(b) Consider the Lagrangian

$$L(\mathbf{w}, \alpha) := \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \alpha^{(i)}(1 - y^{(i)}\mathbf{w}^\intercal \mathbf{x}^{(i)}) \tag{3}$$

where $\alpha^{(i)}$ are elements of α. *Note: This Lagrangian is not the same as the solution in the previous part.*

Derive the dual program for the Lagrangian given in Eq. (3). Provide all its constraints if any.

Your answer: putting $\nabla_w L(w,\alpha) = 0 = w + \sum_i \alpha(0 - y^{(i)}x^{(i)})$

eliminating "w" we get

Dual :

$$\inf L(w,\alpha) = \begin{cases} \sum \alpha_i - \frac{1}{2}\|\sum \alpha_i y_i x_i\|^2 & \alpha \geq 0 \\ -\infty \end{cases}$$

3

4

Scanned by CamScanner

(c) Recall that a kernel SVM optimizes the following program

$$\max_{\alpha} \sum_{i=1}^{|\mathcal{D}|} \alpha^{(i)} - \frac{1}{2} \sum_{i,j=1}^{|\mathcal{D}|} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \kappa(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \tag{4}$$

s.t. $\alpha^{(i)} \geq 0$ and $\sum_{i}^{|\mathcal{D}|} \alpha^{(i)} y^{(i)} = 0$

We have chosen the kernel to be

$$\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\intercal \mathbf{z})^2 + 1$$

Consider the following dataset $\mathcal{D}_2$ in the one-dimensional space; $x^{(i)}, y^{(i)} \in \mathbb{R}$.

i	$x^{(i)}$	$y^{(i)}$
1	$\frac{1}{2}$	+1
2	-1	+1
3	$\sqrt{3}$	-1
4	4	-1

$$K = \begin{bmatrix} \frac{1}{16} & \frac{1}{4} & \frac{3}{4} & 4 \\ \frac{1}{4} & 1 & 3 & 16 \\ \frac{3}{4} & 3 & 9 & 16\times4 \\ 4 & 16 & 16\times4 & 16^2 \end{bmatrix} \text{tanh()}$$

What are the optimal primal parameters, $\mathbf{w}^*$, b^* when optimizing the program in Eq. (4) on the dataset $\mathcal{D}_2$. Note: b is NOT included in the margin or the features (treat it explicitly).

Hint: First, construct a feature vector $\phi(\mathbf{x}) \in \mathbb{R}^2$ such that $\kappa(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\intercal \phi(\mathbf{z})$ for the given one dimensional dataset. Then use this feature vector to transform the data $\mathcal{D}_2$ into feature space and plot the result. Read of the bias term b and the optimal weight vector $\mathbf{w}^*$.

Your answer:

$K(x^{(i)}, x^{(j)}) = \left[\cancel{\frac{5}{4}} \ \cancel{\frac{1}{2}} \ 1+\frac{\sqrt{3}}{2} \ \cancel{2} \ \cancel{3/17} \right.$ $\left[\frac{1}{2} \ -1 \ \sqrt{3} \ 4\right]^T \begin{bmatrix} \frac{1}{2} \\ -1 \\ \sqrt{3} \\ 4 \end{bmatrix}^T$

$W = [-4, 0]$

0

(d) (Continuing from previous part) Which of the points in $\mathcal{D}_2$ are support vectors? What are $\alpha^{(1)}$ and $\alpha^{(2)}$?

Hint: To find $\alpha^{(2)}$ make use of the relationship between the primal solution and the dual variables, *i.e.*, $\mathbf{w}^* = \sum_{i=1}^{N} \alpha^{(i)} \phi(\mathbf{x}^{(i)})$. Assume $\mathbf{w}^* = [-4 \quad 0]^T$ if you couldn't solve part (c).

Your answer:

vectors i : 1, 4.

0

0

Scanned by CamScanner

4. **[7 points]** Multiclass Classification

Consider the objective function of a multiclass SVM given by

$$\min_{w,\xi^{(i)}\geq 0} \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \xi^{(i)}$$

$$\text{s.t.} \quad w_{y^{(i)}}^{\top}\phi(x^{(i)}) - w_{\hat{y}}^{\top}\phi(x^{(i)}) \geq 1 - \xi^{(i)} \quad \forall i \in \{1,\ldots,n\}; \hat{y} \in \{0,\ldots,K-1\}$$

where $w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{K-1} \end{bmatrix}$.

(a) What's the optimal value of $\xi^{(i)}$, given $\phi(x^{(i)}), y^{(i)}$, and w?

Your answer:

$\xi^{(i)} = \max_{\hat{y}} \left(1 - (w^T_{y^{(i)}} \phi(x^i) - w^T_{\hat{y}} \phi(x^i))\right)$ 2

(b) Rewrite the objective function in unconstrained form, using the optimal value of $\xi^{(i)}$.

Your answer: $\min_{w} \frac{1}{2}\|w\|_2^2 + \Sigma \max (1 + w^T_{\hat{y}} \phi(x^i)) - w^T_{y^{(i)}} \phi(x^i)$ 1

(c) Briefly explain using English language the reason for using $w_{y^{(i)}}^{\top}\phi(x^{(i)}) - w_{\hat{y}}^{\top}\phi(x^{(i)}) \geq 1 - \xi^{(i)}$ in a multiclass SVM formulation, *i.e.*, what does this constraint encourage?

Your answer: The above equation is giving margin of 1 units (could be any $L(y,\hat{y})$) with softness with $\xi^{(i)}$. where as LHS implies that when $\hat{y}$ is correct it gets minimum penalty (enforces $\hat{y}$ to be equal to ground truth) 2

5

Scanned by CamScanner

(d) Suppose we want to train a set of one-vs-rest classifiers and a set of one-vs-one classifiers on a dataset of 5,000 samples and 10 classes, each class having 500 samples. Suppose the running time of the training underlying binary classifier we use is n^2 in nanoseconds, where n is the size of the training dataset. Which one is faster, training of the one-vs-rest classifiers or training of the one-vs-one classifiers? Explain your reason.

Your answer: (1 vs Many) or (one vs rest is faster) D

as in one vs rest total comparisons are only 10

but in 1-vs-1 there are 9 X 5 = 45 comparisons.

1v1

$\frac{10\times 9}{2} \times (500)^2$

$500 \times \frac{\cancel{10}^{5}\times 9}{\cancel{2}} \times (5000)^2$

45

1vm.

$500 \times 10 \times (500)^2$

$500 \times 10 \times (5000)^2$

0

Scanned by CamScanner

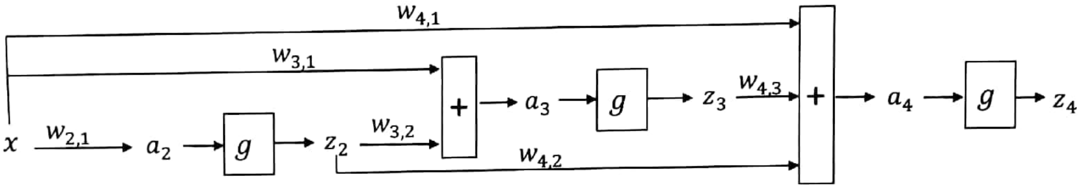
5. **[13 points]** Backpropagation

Consider the neural network given in the figure below. The network has a scalar input variable $x \in \mathbb{R}$ and a scalar target $t \in \mathbb{R}$ and is defined as follows:

$$z_j = \begin{cases} x, & \text{if } j = 1 \\ g(a_j) & \text{if } j \in \{2,3,4\} \text{ with } a_j = \sum_{i=1}^{j-1} w_{j,i} z_i \end{cases} \tag{5}$$

Suppose that the network is trained to minimize the L2 loss per sample, *i.e.*, $E = \frac{1}{2}(z_4 - t)^2$. The error gradient can be written as:

$$\frac{\partial E}{\partial w_{j,i}} = \delta_j z_i \tag{6}$$



(a) [2 pts] For $g(x) = \sigma(x) = \frac{1}{1+e^{-x}}$, compute the derivative $g'(x)$ of $g(x)$ as a function of $\sigma(x)$.

Your answer:

$g'(x) = \sigma(x)\,(1 - \sigma(x))$

2

$E = \frac{1}{2}(z_4 - t)^2$

$\frac{\partial E}{\partial z_4} = 2(z_4 - t)$

$(z_4 - t)\, g'(a_4)$

(b) [2 pts] Compute δ_4 as a function of z_4, t and $g'(a_4)$.

Your answer:

$\delta_4 = g'(a_4)\,(z_4 - t)$

2

(c) [2 pts] Compute δ_3 as a function of δ_4, $w_{4,3}$ and $g'(a_3)$.

Your answer:

$a_4 = w_{41}\, x + w_{42}\, z_2 + \underline{\underline{w_{43}\, z_3}}$

$\delta_3 = \delta_4 \cdot w_{43} \cdot g'(a_3)$

2

Scanned by CamScanner

(d) [3 pts] Compute δ_2 as a function of δ_3, δ_4, $w_{3,2}$, $w_{4,2}$ and $g'(a_2)$.

Your answer:

$$\delta_2 = g'(a_2)\,(w_{32}\,\delta_3 + w_{42}\,\delta_4)$$

3

(e) [4 pts] Write down a recursive formula for computing δ_j for $j \in \{2, \cdots, M-1\}$, as a function of δ_k, $w_{k,j}$ and $g'(a_j)$ for $k \in \{j+1, \cdots, M\}$.

Your answer:

$$\delta_j = g'(a_j) \sum_{k=j+1}^{M} w_{kj}\,\delta_k$$

4

$$\delta_1 = g'(a_1)\,(w_{41}\,\delta_4 + w_{31}\,\delta_3 + w_{21}\,\delta_2)$$

$$\delta_2 = g'(a_2)\,(w_{32}\,\delta_3 + w_{42}\,\delta_2)$$

$$\delta_3 = g'(a_3)\,(w_{43}\,\delta_4)$$

$$\delta_j = g'(a_j) \sum_{j=i+1}^{M} w_{ij}\,\delta_i$$

Scanned by CamScanner

6. **[10 points]** Inference in Discrete Markov Random Fields

(a) Inference in Markov random fields amounts to finding the highest scoring configuration for a set of variables. Suppose we have two variables $x_1 \in \{0,1\}$ and $x_2 \in \{0,1\}$ and their local evidence functions $\theta_1(x_1)$ and $\theta_2(x_2)$ as well as pairwise function $\theta_{1,2}(x_1,x_2)$. Using this setup, inference solves $\arg\max_{x_1,x_2} \theta_1(x_1) + \theta_2(x_2) + \theta_{1,2}(x_1,x_2)$. Using

$$\theta_1(x_1) = \begin{cases} -1 & \text{if } x_1 = 0 \\ 1 & \text{otherwise} \end{cases} \qquad \theta_2(x_2) = \begin{cases} -1 & \text{if } x_2 = 0 \\ 1 & \text{otherwise} \end{cases}$$

$$\theta_{1,2}(x_1,x_2) = \begin{cases} 2 & \text{if } x_1 = 1 \ \& \ x_2 = 0 \\ 1 & \text{if } x_1 = 0 \ \& \ x_2 = 1 \\ -1 & \text{otherwise} \end{cases}$$

what is the integer linear programming (ILP) formulation of the inference task? Make cost function and constraints explicit for the given problem, i.e., do not use a general formulation.

Your answer: 4.5

$$\begin{bmatrix} b_1(0) \\ b_1(1) \\ b_2(0) \\ b_2(1) \\ b_{12}(0,0) \\ b_{12}(1,0) \end{bmatrix}^T \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 2 \end{bmatrix}$$

$b_{12}(0,1)$ 1

$b_{12}(1,1)$ -1

cost func : $\max_{b}$ $-b_1(0) + b_1(1) - b_2(0) + b_2(1) - b_{12}(0,0) + 2\, b_{12}(1,0) + b_1(0,1) - b_{12}(1,1)$

constraints : (1) $b_1(x), b_2(x), b_{12}(x) \in \{0,1\}$ wher $x \in \{0,1\}$.

(2) $b_1(0) + b_1(1) = 1$
$b_2(0) + b_2(1) = 1$
$b_{12}(00) + b_{12}(01) + b_{12}(10) = 1$
$+ b_{12}(1,1)$

(3) $b_{12}(0,0) + b_{12}(0,1) = b_1(0)$
$b_{12}(1,0) + b_{12}(1,1) = b_1(1)$
$b_{12}(1,0) + b_{12}(0,0) = b_2(0)$
$b_{12}(1,1) + b_{12}(0,1) \quad b_2(1)$

(b) If the two variables instead took on values $x_1, x_2 \in \{0,1,2,3\}$, how many constraints would the integer linear program have?

Your answer: Total = 11 (equality) + 5 (inequality) ← one per assignment per variable

$\frac{1}{2}$

(3) marginal : $2 \times 4 = 8$ ✓

(2) Local : 3 ✓

(1) $b_1(x) \in \{0,1\}$
$b_2(x) \in \{0,1\}$
$b_{12}(x) \in \{0,1\}$
$x \in \{0,1\}$

(c) Let's say we wanted to use a different method to solve this inference problem. Can we use a dynamic programming method? Why or why not?

Your answer: Yes, it is a tree structure.

2

7

Scanned by CamScanner

(d) Name two other inference methods that may be more efficient than ILP, and name one advantage and one disadvantage for each.

Your answer: (1) Linear program : Advantage : Faster
Disadvantage : do not give integer solutions.

2

(2) Dynamic program : Advantage : Computationally inexpensive
Disadvantage : only works for trees

2

Scanned by CamScanner