University of Illinois Spring 2018

CS 446: Machine Learning
Midterm

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

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^*.**

Solution: Plot it out. From geometry, the line goes through (0,-1) and (0,1). $w_1 = 2$(1 points), $w_2 = -1$. (1 points)
Total Points: 2

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

Solution: Two observations. (1 points)

(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.**

Solution: Observe from the data behaves as x^2 (1 points). Therefore, square the features to get a better fit. We obtain the following program(1 points):

$$\operatorname*{argmin}_{w_1, w_2} \sum_{i}^{N} (y^{(i)} - w_1 \cdot (x^{(i)})^2 - w_2)^2$$

Total Points: 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.**

Solution: Derivation (1 points)

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} {x^{(1)}}^2 & 1 \\ {x^{(2)}}^2 & 1 \\ & \vdots \\ {x^{(N)}}^2 & 1 \end{bmatrix}$$

(1 points)

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}$$

(1 points)

Total Points: 3

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

Solution: Overfitting and poor generalization. (1 points)

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)

Solution:
(1 points)
Yes.

(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)

Solution:
(2 points)
Yes.

(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.

Solution:
(5 points)

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \sum_i \frac{-y^{(i)} \mathbf{x}^{(i)} \exp(-y^{(i)} \mathbf{w}^T \mathbf{x}^{(i)})}{1 + \exp(-y^{(i)} \mathbf{w}^T \mathbf{x}^{(i)})}$$

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

Solution:
(2 points)
Logistic/Binomial vs Gaussian probability model

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

Solution:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} ||\mathbf{w}||_2^2 + \sum_i^{|\mathcal{D}|} \alpha^{(i)} (1 - y^{(i)} \mathbf{w}^\intercal \mathbf{x}^{(i)} + b)$$

(1 point)
Total of 1 points

(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.

Solution: Take the gradient with respect to **w** and set it to 0. (1 point)

$$\sum_{i=1}^{N} \alpha^{(i)} - \frac{1}{2} || \sum_{i}^{N} \alpha^{(i)} y^{(i)} x^{(i)} ||_2^2$$

(1 point) s.t.

$$\alpha^{(i)} \geq 0$$

(1 point)

Total of 3 points

(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

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 $\mathbf{w}^*$.

Solution: Observe that x is one dimensional, then kernel can be written as $\phi^\intercal(x)\phi(z)$, where $\phi(x) = [x^2, 1]^\intercal$. (1 point)

Let $\mathbf{w} = [w_1, w_2]$.

From geometry b has to be the midpoint (1 point), $w_2 = 0$, $b* = 2$ and $w_1 = -c$ for some $c > 0$.

Plug in support vector example (2), $w_1 = -1$.

$\mathbf{w}^* = [-1, 0]$ (1 point) and $b^* = 2$ (1 point)

Total of 4 points

(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.*, $\mathbf{w}^* = \sum_{i=1}^{N} \alpha^{(i)}y^{(i)}\phi(\mathbf{x}^{(i)})$. Assume $\mathbf{w}^* = [-4 \quad 0]^T$ if you couldn't solve part (c).

Solution: Support vectors are example 2 and 3. (1 point)
From the dual formulation, we know that

$$\mathbf{w}^* = \sum_i^N \alpha^{(i)} y^{(i)} \mathbf{z}^{(i)},$$

where N is the size of the dataset and $\mathbf{z} = \phi(x)$.
Plug in the support vectors example (2) and (3) and solve.
$\alpha^{(1)} = 0$ as not a support vector and $\alpha^{(2)} = \frac{1}{2}$. (1 point)
Total of 2 points
If you couldn't solve part (c) using $w^* = [-4, 0]^T$ $\alpha^{(2)} = 2$. (1 point)

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?

Solution:
(2 points)

$$\hat{\xi}^{(i)} = \max_{\hat{y}} \{1 - w_{y^{(i)}}^\top \phi(x^{(i)}) + w_{\hat{y}}^\top \phi(x^{(i)})\}$$

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

Solution:
(1 point)

$$\min_w \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^n \max_{\hat{y}} \{1 - w_{y^{(i)}}^\top \phi(x^{(i)}) + w_{\hat{y}}^\top \phi(x^{(i)})\}$$

(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?

Solution:
(2 points) It encourages the difference between the score of true label $y^{(i)}$ and any class $\hat{y}$ on $x^{(i)}$ to be larger with a margin.

(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 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.

Solution:
(2 points) OVR classifier takes $9 \times 5000^2 = 2.25 \times 10^8$ nanoseconds or $10 \times 5000^2 = 2.5 \times 10^8$ nanoseconds, while OVO classifier takes $45 \times 1000^2 = 4.5 \times 10^7$ nanoseconds. So OVO classifier is faster.

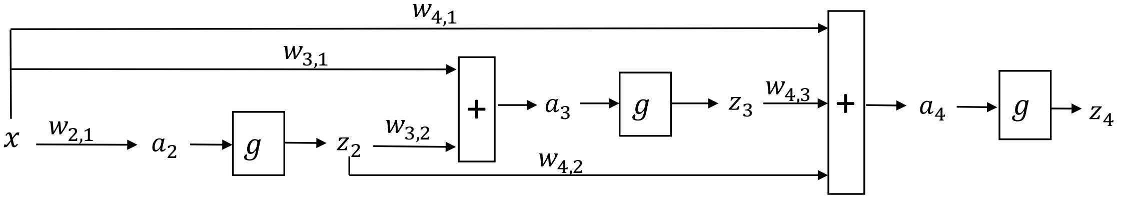
5. **[13 points]** Backpropagation

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 $\sigma(x)$.

Solution: $g'(x) = \sigma(x)(1 - \sigma(x))$

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

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

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

Solution: $\delta_3 = \delta_4 w_{4,3} g'(a_3)$

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

Solution: $\delta_3 w_{3,2} g'(a_2) + \delta_4 w_{4,2} g'(a_2)$

(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 $k \in \{j+1, \cdots, M\}$.

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

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.

Solution:

$$\max_b \left(\sum_{i=1}^{2} b_i(1) - b_i(0) \right) - b_{1,2}(1,1) + 2b_{1,2}(1,0) + b_{1,2}(0,1) - b_{1,2}(0,0)$$

$$\text{s.t.} \begin{cases} b_1(0), b_1(1), b_2(0), b_2(1) \in \{0, 1\} \\ b_{1,2}(0,0), b_{1,2}(1,0), b_{1,2}(0,1), b_{1,2}(1,1) \in \{0, 1\} \\ b_1(0) + b_1(1) = 1, b_2(0) + b_2(1) = 1 \\ b_{1,2}(0,0) + b_{1,2}(1,0) + b_{1,2}(0,1) + b_{1,2}(1,1) = 1 \\ b_1(0) = b_{1,2}(0,0) + b_{1,2}(0,1) \\ b_1(1) = b_{1,2}(1,0) + b_{1,2}(1,1) \\ b_2(0) = b_{1,2}(0,0) + b_{1,2}(1,0) \\ b_2(1) = b_{1,2}(0,1) + b_{1,2}(1,1) \end{cases}$$

(5 pts)

(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?

Solution: 24 domain constraints + 3 intra-region marginalization constraints + 8 inter-region marginalization constraints = 35 total constraints (1 pt)

(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?

Solution:
Yes, we can - the graph represented by the problem is a tree. (2 pts)

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

Solution: Some possible answers:
Linear programming relaxation of the ILP - is no longer NP Hard and we have good solvers, but still may be inefficient for larger problems.
Message Passing: Efficient due to analytically computable sub-problems, but it takes special care to find global optimum (2 pts) Graph Cut: Have fast solvers, but requires potentials to have specific properties to work.