Introduction to Machine Learning (Minor) (CS 419M)Midterm Exam
Computer Science and EngineeringMarch 1, 2020
Indian Institute of Technology Bombay

Instructions

1. This paper has three questions. Each question carries 10 marks. Therefore, the maximum marks is 30.

2. Write your answers on a paper, scan and submit them at the end of the exam.

3. Write your name, roll number and the subject number (CS 419M) on the top of each of your answer script.

4. There are multiple parts (sub-questions) in each question. Some sub-questions are objective and some are subjective.

5. There will be partial credits for subjective questions, if you have made substantial progress towards the answer. However there will be NO credit for rough work.

6. Please keep your answer sheets different from the rough work you have made. Do not attach the rough work with the answer sheet. You should ONLY upload the answer sheets.

1. Assume that we are given a set of features $\{(\boldsymbol{x}_i, y_i) \mid i \in \{1, 2, ..., N\}\}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, $y \in \{-1, +1\}$. We wish to train a function $h : \mathbb{R}^d \to \mathbb{R}$, so that $\text{Sign}(h(\boldsymbol{x})) = y$. To that aim, we seek to solve the following:

$$\underset{h \in \mathcal{H}}{\text{minimize}} \sum_{i=1}^{N} \mathbb{I}[\text{Sign}(h(\boldsymbol{x}_i)) \neq y_i] \tag{1}$$

where $\mathbb{I}[A] = 1$ if A is true and 0 if A is false. Moreover, $\mathcal{H}$ is the set of all functions that map from $\mathbb{R}^d$ to $\mathbb{R}$.

1.a This problem is hard to solve in general. That is why, we resort to several approximations. In the following, mark and explain which ones are good approximator of $\mathbb{I}[\text{Sign}(h(\boldsymbol{x}_i)) \neq y_i]$ in Eq. 1.

(i) $\max\{0, 1 - y_i \cdot h(\boldsymbol{x}_i)\}$ ______ (Yes/No)

(ii) $\min\{0, 1 - y_i \cdot h(\boldsymbol{x}_i)\}$ ______ (Yes/No)

(iii) $\dfrac{\exp(-y_i \cdot h(\boldsymbol{x}_i))}{1 + \exp(-y_i \cdot h(\boldsymbol{x}_i))}$ ______ (Yes/No)

(iv) $\dfrac{1}{1 + \exp(-y_i \cdot h(\boldsymbol{x}_i))}$ ______ (Yes/No)

Explanation: ______________________________

1.a	/1+1+1+1+2	

1.b Suppose we restrict $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$, *i.e.*, $h(\boldsymbol{x})$ is a linear function. Then write the approximation of the optimization problem defined in Eq. 1 in terms of any (correct) one approximation in the previous question. Specifically, fill up the gaps

$$\underset{\_\_\_\_\_\_}{\text{minimize}} \sum_{i=1}^{N} \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

1.b	/2	

1.c Note that, in the loss function proposed in Eq. 1 is called a pointwise loss since we aim to minimize the misclassification loss per-instance level (i.e. the sum $\sum_{i=1}^{N}$ is taken over all instances).

In contrast to this, in document retrieval in web search applications, one rather tries to ensure that if we sort all N instances in the decreasing order of $h(\boldsymbol{x}_i)$, then the instances with $y_i = +1$ will appear on the top of the ranked list. The loss which encodes this condition is called ranking loss. There are several ranking losses in literature. One of them considers pairwise ranking loss, which takes all pairs of instances (i, j) so that $y_i = +1, y_j = -1$ and tries to ensure that the instances i (i.e., with $y_i = +1$) enjoy higher $h(\boldsymbol{x}_i)$ than the instances j (i.e., with $y_j = -1$). More specifically, we seek to minimize the number of pairs which violate that condition, *i.e.*:

$$\underset{h \in \mathcal{H}}{\text{minimize}} \sum_{j: y_j = -1} \sum_{i: y_i = +1} \mathbb{I}[h(\boldsymbol{x}_i) \leq h(\boldsymbol{x}_j)]. \tag{2}$$

Provide one approximation of the above objective in the similar line of Question 1.a:

$$\underset{h \in \mathcal{H}}{\text{minimize}} \sum_{j: y_j = -1} \sum_{i: y_i = +1} \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. \tag{3}$$

1.c	/2	

2. This question is about linear regression problem. Given $\mathcal{D} = \{(\boldsymbol{x}_i, y_i), i \in 1, ..., N\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, and $y_i \in \mathbb{R}$, we wish to estimate a linear function $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ which will approximate y.

2.a Generally, we aim to minimize the following loss with respect to $\boldsymbol{w} = [w_1, ..., w_d]$.

$$L(\boldsymbol{w}) = \sum_{i=1}^{|\mathcal{D}|} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \sum_{j=1}^{d} w_j^2 \tag{4}$$

Assume that $\mathcal{D}$ has only one element, *i.e.*, $\mathcal{D} = (\boldsymbol{x}, y)$. Then, prove that the minimum solution $L(\boldsymbol{w}^*)$ is given by:

$$L(\boldsymbol{w}^*) = y^2 \left[1 - \boldsymbol{x}^\top (\lambda \mathbb{I} + \boldsymbol{x}\boldsymbol{x}^\top)^{-1} \boldsymbol{x}\right] \tag{5}$$

2.a	/2	

2.b Use the following equality [1]

$$\left(A + uv^\mathsf{T}\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\mathsf{T}A^{-1}}{1 + v^\mathsf{T}A^{-1}u},$$

to (i) simplify Eq. 5 and then using that simplification find the values of (ii) $\lim_{\lambda \to 0} L(\boldsymbol{w}^*)$ and $\lim_{\lambda \to \infty} L(\boldsymbol{w}^*)$.

2.b	/2+1	

2.c Discuss the probabilistic interpretation of the objective in Eq. 4.

2.c	/2	

2.d we wish to assign different weightage α_i to the samples. To that aim, we seek to minimize the following regularized loss function (Here, $\mathcal{D}$ contains more than one samples):

$$\min_{\boldsymbol{w}} \sum_{i=1}^{|\mathcal{D}|} \alpha_i (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \sum_{i=1}^{d} w_j^2$$

Derive the expression for the optimal solution of $\boldsymbol{w}^*$.

2.d	/3	

[1] Source: Wikipedia, this is called Sherman-Morrison lemma

3. This problem is about support vector machine. Suppose $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \mid i = 1, ..., N\}$ with $y_i \in \{+1, -1\}$ and $\boldsymbol{x}_i \in \mathbb{R}^2$, *i.e.*, the features are two dimensional vectors and $||\boldsymbol{a}|| = \sqrt{a_1^2 + a_2^2 + ... + a_d^2}$ for any d element vector. For instances which are linearly separable, we first note three equivalent SVM formulations:

Formulation-1:

$$\min_{\boldsymbol{w}, b, \{\xi_i\}} ||\boldsymbol{w}||^2 + C \sum_{i=1}^{|\mathcal{D}|} \xi_i$$

$$\text{such that, } y_i \cdot (\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0$$

$$i = 1, .., N.$$

Formulation-2:

$$\min_{\boldsymbol{w}, b} ||\boldsymbol{w}||^2 + C \sum_{i=1}^{|\mathcal{D}|} \max(0, 1 - y_i \cdot (\boldsymbol{w}^\top \boldsymbol{x}_i + b))$$

Formulation-3:

$$\min_{0 \leq \alpha_i \leq C} \sum_{i,j=1}^{|\mathcal{D}|} \frac{1}{2} \alpha_i \alpha_j y_i y_j \cdot (\boldsymbol{x}_i^\top \boldsymbol{x}_j) - \sum_{i=1}^{|\mathcal{D}|} \alpha_i$$

$$\text{such that, } \sum_{i=1}^{|\mathcal{D}|} y_i \alpha_i = 0$$

3.a Assume $b = 0$, then show that the optimal solution $\boldsymbol{w}^*$ is given by, $||\boldsymbol{w}^*|| \leq \sqrt{C|\mathcal{D}|}$.

3.a /3

3.b Assume that $y_i = 1$ if $||\boldsymbol{x}_i - [1, -1]|| < 0.5$ and $y_i = -1$ if $||\boldsymbol{x}_i - [1, 1]|| < 0.5$. Sketch a scatter plot of $(\boldsymbol{x}_i, y_i)$ and show qualitatively, how the SVM solution would look like?

3.b /3

3.c Assume that $y_i = 1$ if $||\boldsymbol{x}_i|| < 0.5$ and $y_i = -1$ if $1 < ||\boldsymbol{x}_i|| < 2$. Sketch a scatter plot of $(\boldsymbol{x}_i, y_i)$. Then explain that such a situation cannot work with the above linear SVM models. How can we modify one of these formulations to classify these points?

3.c /4

Total: 30