Introduction to Machine Learning (Minor) (CS 419M)Endsem Exam
Computer Science and EngineeringMay 3, 2020
Indian Institute of Technology Bombay

-

Instructions

1. This paper has four questions. Each of first three questions carries 15 marks. The last question has 10 marks. Therefore, the maximum marks is 55.

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. A is a user in Facebook. B, C, D are her neighbors. If A makes one post, it will be visible by B, C and D. Then each of these neighbors may (or may not) mark one "like" in the post. Assume that the probability that a user $u \in \{B, C, D\}$ will like a post of a topic `Topic` is $p_{u,\texttt{Topic}}$. We call these probabilities as preference probabilities. Assume that `Topic` $\in \{\texttt{sports}, \texttt{cinema}, \texttt{politics}\}$ and the probability that A will make a post with topic `Topic` is given by $q_{\texttt{Topic}}(t)$ at time $t \in \{1, ..T\}$.

1.a What is the expected number of likes A will receive for T random posts.

1.a	/3	

1.b Suppose A does not care about the number of likes she receives from C and D. Hence, she wants to *maximize* the likes she receives from B. Therefore, she posts T messages with suitable topics in order to maximize the total number of likes. Suppose A knows the preference probabilities. If $p_{B,\texttt{sports}} = 0.6$, $p_{B,\texttt{cinema}} = 0.5$ and $p_{B,\texttt{politics}} = 0.56$, then fill up the gaps:

$$q_{\texttt{sports}}(t) = \ldots$$

$$q_{\texttt{cinema}}(t) = \ldots$$

$$q_{\texttt{politics}}(t) = \ldots$$

Provide a clear explanation of the above choice.

1.b	/3+2	

1.c Consider the previous setup. However, here, we assume that A does not know the preference probabilities. However, still she want to maximize the expected number of likes he receives from B. More specifically, she wants to maximize

$$R(T) = \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}[B \text{ likes the post made at time } t]]$$

where $\mathbb{I}[x] = 1$ if x is true and 0 otherwise. Please suggest an algorithm which will maximize $R(T)$.

1.c	/4	

1.d Finally, suppose A wants to *maximize* the likes she receives from B, C and D. Therefore, she post T messages with suitable topics in order to maximize the total number of likes from B, C and D. Suppose A knows the preference probabilities. For any general preference probabilities, what should be the values of $q_{\bullet}(t)$:

$$q_{\texttt{sports}}(t) = \ldots$$

$$q_{\texttt{cinema}}(t) = \ldots$$

$$q_{\texttt{politics}}(t) = \ldots$$

1.d	/3	

2. This question consists of different parts from different topics taught in the class.

2.a We are given $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ and we wish to cluster these points in K clusters. Write the objective for K-means clustering and the algorithm for optimizing this objective.

2.a	/2+4	

2.b Describe the stochastic gradient descent and minibatch gradient descent algorithms. What are the individual advantages and disadvantages.

2.b	/3+3	

2.c Consider the neural networks.

$$\begin{aligned}\text{Linear}_1(\boldsymbol{x}) &= W_1^\top \boldsymbol{x} \\ \text{Linear}_2(\boldsymbol{x}) &= W_2^\top \boldsymbol{x} \\ \text{ReLU}(\cdot) &= \max[0, \cdot] \qquad (1)\end{aligned}$$

Then we predict the response $y \in \mathbb{R}$ from the features $\boldsymbol{x} \in \mathbb{R}^d$ as follows:
$y = \text{Linear}_2(\text{ReLU}(\text{Linear}_1(\text{Linear}_1(\boldsymbol{x}))))$. Find the total number of trainable parameters.

2.c	/3	

3. We say that a learning algorithm is stable if the output of the learning algorithm does not change by a large amount if we add a single data. Specifically, say we have $\mathcal{S} = \{z_1, z_2, ..., z_n\}$ and $\mathcal{S}' = \{z_1, z_2, ...z_i', .., z_n\}$ where, $z = (\boldsymbol{x}, y)$, i.e., $\mathcal{S}$ and $\mathcal{S}'$ differ by only i-th element. Then we say the learned parameter
$\boldsymbol{w}^*(\mathcal{S}) = \operatorname{argmin}_{\boldsymbol{w}} L(\boldsymbol{w}, \mathcal{S}) = \sum_{z_i \in \mathcal{S}} \ell(\boldsymbol{w}, z_i)$
is stable if $||\boldsymbol{w}^*(\mathcal{S}) - \boldsymbol{w}^*(\mathcal{S}')||$ is small. Smaller the value it takes, more stable is the loss. Consider two loss functions one for unregularized SVM and the other for regularized SVM.

$$L(\boldsymbol{w}, \mathcal{S}) = \sum_{i \in \mathcal{S}} \max[0, 1 - y_i \boldsymbol{w}^\top \boldsymbol{x}_i]$$
$$L_\lambda(\boldsymbol{w}, \mathcal{S}) = L(\boldsymbol{w}, \mathcal{S}) + \lambda |\mathcal{S}| ||\boldsymbol{w}||^2 \tag{2}$$

Mark the correct choices:

3.a If $\boldsymbol{w}^*(\mathcal{S}) = \operatorname{argmin}_{\boldsymbol{w}} L(\boldsymbol{w}, \mathcal{S})$, then it is always the case that $||\boldsymbol{w}^*(\mathcal{S}) - \boldsymbol{w}^*(\mathcal{S}')|| \le C/|\mathcal{S}|$ where C is independent of $|\mathcal{S}|$.
____________ (True/False)

If $\boldsymbol{w}^*(\mathcal{S}) = \operatorname{argmin}_{\boldsymbol{w}} L_\lambda(\boldsymbol{w}, \mathcal{S})$, then it is always the case that $||\boldsymbol{w}^*(\mathcal{S}) - \boldsymbol{w}^*(\mathcal{S}')|| \le C/|\mathcal{S}|$ where C is independent of $|\mathcal{S}|$.
____________ (True/False)

3.a	/3	

3.b Explain the above choice

3.b	/3	

3.c In the second case, i.e. when $\boldsymbol{w}^*(\mathcal{S}) = \operatorname{argmin}_{\boldsymbol{w}} L_\lambda(\boldsymbol{w}, \mathcal{S})$, how does $||\boldsymbol{w}^*(\mathcal{S}) - \boldsymbol{w}^*(\mathcal{S}')||$ depend on λ?

3.c	/3	

3.d Suppose, there are two convex loss functions $L^{(k)}(\boldsymbol{w}, \mathcal{S}) = \sum_{z_i \in \mathcal{S}} \ell^{(k)}(\boldsymbol{w}, z_i)$, $k = 1, 2$, so that $\left\| \frac{d\ell^{(1)}(\boldsymbol{w}, z)}{d\boldsymbol{w}} \right\|_2 < \left\| \frac{d\ell^{(2)}(\boldsymbol{w}, z)}{d\boldsymbol{w}} \right\|_2$ for all $\boldsymbol{w}$ and z. Compare the stability of $L^{(1)}$ and $L^{(2)}$ with explanation.

3.d	/3	

3.e Prove that
$$\min_{\boldsymbol{w}} L_\lambda(\boldsymbol{w}, \mathcal{S} \cup z_k) - \min_{\boldsymbol{w}} L_\lambda(\boldsymbol{w}, \mathcal{S}) \ge \lambda ||\boldsymbol{w}^*(\mathcal{S} \cup k)||^2$$

3.e	/3	

4. **4.a** Consider the training objective for a linear regression problem:

$$\min_{\boldsymbol{w}} \sum_{(x,y)\in\text{training data}} (\boldsymbol{w}^T\boldsymbol{x} - y)^2 + \lambda \boldsymbol{w}^T\boldsymbol{w}$$

Select all that applies as we train different models for increasing value of λ. Mark the correct options with explanations.

- Models trained with larger λ have larger training error
- Models trained with larger λ have smaller training error
- Models trained with larger λ have smaller error on unseen test instances and generalize better.
- Models trained with larger λ have larger error on unseen test instances and generalize better.
- Models trained with larger λ, have smaller value of norm $\boldsymbol{w}$
- Models trained with larger λ, have large value of norm of $\boldsymbol{w}$

4.a	/3+2	

4.b Consider the surrogate of the ranking loss functions

$$L(\boldsymbol{w}) = \sum_{i,j\in D, y_i=-1, y_j=+1} [\boldsymbol{w}^\top\boldsymbol{x}_i - \boldsymbol{w}^\top\boldsymbol{x}_j + 1]^2 \cdot \mathbb{I}[\boldsymbol{w}^\top\boldsymbol{x}_i - \boldsymbol{w}^\top\boldsymbol{x}_j + 1 \geq 0] \tag{3}$$

$\mathbb{I}(\boldsymbol{x}) = 1$ if x is true and 0 otherwise. Compute the gradient of the above loss.

4.b	/3	

4.c What is the disadvantage of the above loss function (4.b) in comparison to simple classification loss?

4.c	/2	

Total: 55