IIT BOMBAY

Indian Institute of Technology Bombay

CS 419 Quiz 2

Date: November 6, 2024 **Max Marks: 20** **Duration: 90 minutes**

Instructions:

- Answer all questions.
- Show all work clearly, be as clear and precise as possible.

Question 1 : LOGISTIC REGRESSION

Logistic Regression and SVM may seem simple, but they provide a solid foundation for more complex machine learning challenges, offering key insights into classification and decision boundaries. These form the technical foundations of many larger machine learning applications. We will now explore one such application: document retrieval.

In the context of document retrieval, we define:

- **Query** ($\mathbf{q} \in \mathbb{R}^k$): A vector of dimension k that represents the user's search intent in the vector space.
- **Document** ($\mathbf{d} \in \mathbb{R}^k$): Each document $d^{(i)}$ in the collection is represented as a k-dimensional vector.
- **Corpus** ($\mathcal{D}$): The set of all documents, $\mathcal{D} = \{\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \ldots, \mathbf{d}^{(n)}\}$

The goal of document retrieval is to find the top-k document vectors $\mathbf{d}^{(i)}$ from the corpus that are most similar to the query vector $\mathbf{q}$, typically using similarity measures between the query and document. The similarity between the query vector $\mathbf{q}$ and a document vector $\mathbf{d}$ is denoted as $sim(\mathbf{d}, \mathbf{q})$, where higher values of $sim(\mathbf{d}, \mathbf{q})$ indicate greater relevance of the document to the query, while lower values of $sim(\mathbf{d}, \mathbf{q})$ indicate lower relevance of the document to the query.

In document retrieval, datasets are often structured as triples $(\mathbf{q}, \mathbf{d}_p, \mathbf{d}_n)$, where:

- $\mathbf{q}$: The query vector representing the user's search intent.
- $\mathbf{d}_p$: A positive document vector that is relevant to the query $\mathbf{q}$.
- $\mathbf{d}_n$: A negative document vector that is not relevant to the query $\mathbf{q}$.

For example, if $\mathbf{q}$ represents the query *machine learning*, then $\mathbf{d}_p$ could be a document vector for an article about *supervised learning techniques*, while $\mathbf{d}_n$ could represent a document about *traditional art techniques*, which is irrelevant to the query. The aim is to maximize $\text{sim}(\mathbf{q}, \mathbf{d}_p)$ and minimize $\text{sim}(\mathbf{q}, \mathbf{d}_n)$ to improve retrieval accuracy.

(a) Given the triple $(\mathbf{q}, \mathbf{d}_p, \mathbf{d}_n)$, a positive margin hyperparameter ($\alpha > 0$) we can define a margin loss function to optimize the similarity scores. We take inspiration from the SVM framework and define the loss as :

$$\mathcal{L} = \max(0, \text{----------})$$

The loss aims to ensure that the positive document is more similar to the query than the negative document by a specified margin α. Fill in the blank above only in terms of $sim(\mathbf{q}, \mathbf{d}_p)$, $sim(\mathbf{q}, \mathbf{d}_n)$, α. Explain the loss design choice properly.
[2 marks]

(b) What is the interpretation of the margin hyperparameter ($\alpha > 0$) in this problem?
[1 mark]

(c) Another very famous way of formulating the loss is in terms of a binary classification framework, taking inspiration from Logistic Regression. Let us define

$$P(\mathbf{d}_p|\mathbf{q}) = \frac{e^{\text{sim}(\mathbf{d}_p,\mathbf{q})}}{e^{\text{sim}(\mathbf{d}_p,\mathbf{q})} + e^{\text{sim}(\mathbf{d}_n,\mathbf{q})}}$$

$$P(\mathbf{d}_n|\mathbf{q}) = \frac{e^{\text{sim}(\mathbf{d}_n,\mathbf{q})}}{e^{\text{sim}(\mathbf{d}_p,\mathbf{q})} + e^{\text{sim}(\mathbf{d}_n,\mathbf{q})}}$$

Formulate a Negative log-likelihood loss $\mathcal{L}_{\text{logistic}}$ that can be used to train the model, using the framework of Logistic Regression.
[2 marks]

(d) Once the loss in part (c) is formulated, we will use gradient descent-based techniques to train the model. Suppose we design a very simple model to calculate similarity as:

$$sim(\mathbf{d}, \mathbf{q}) = \mathbf{q}^T \mathbf{W} \mathbf{d}$$

Calculate the gradient of $\mathcal{L}_{\text{logistic}}$ with respect to $\mathbf{W} \in \mathbb{R}^{k \times k}$.
[5 marks]

Question 2 : SUPPORT VECTOR MACHINES

Part (a) [5 marks]

Support vector machines have a linear decision boundary. Non-linear data cannot be separated by it. For instance, consider the dataset in fig 1 where each point $x \in \mathbb{R}^2$. The points denoted by $+$ sign are positive samples and the ones denoted by $\circ$ are negative samples. Consider the dashed circle has radius $r = 1$. A transformation function has to be applied so as to make the data linearly separable.

Provide a transformation $\phi : \mathcal{V} \rightarrow \mathcal{U}$ (refer to $\mathcal{V}$ in fig 1) such that the data becomes linearly separable in $\mathcal{U}$. Note that $\mathcal{U}$ must also be $\mathbb{R}^2$.
Provide clear form of the transformation function ϕ, and mention any assumptions you make while answering the question.

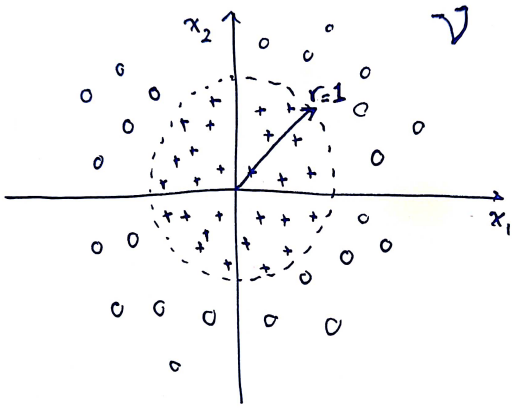


Figure 1

Part (b) Provide proper reasoning for your choices. (No marks for guesses.)

(1) Soft margin SVM always converges on datasets which are not linearly separable. State True or False. [1 mark]

(2) There is a slack penalty parameter C in the objective of soft-margin SVM and the optimization objective is given below [2 marks]

$$||w||^2 + C\sum_{i=1}^{n}\zeta_i$$

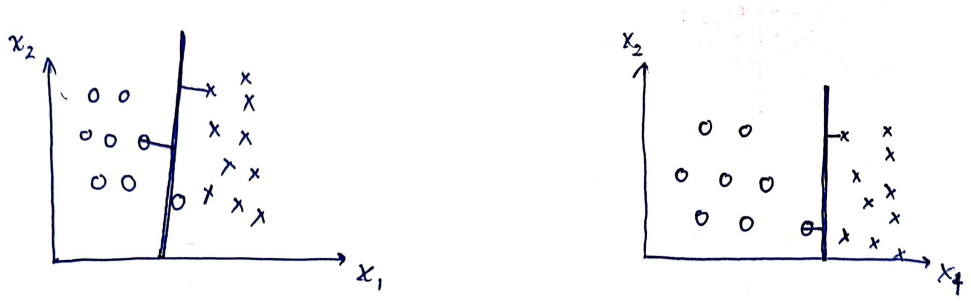
Choose correct option(s) from the following. Assume that fig 2 (a) and (b) both represent the same dataset. The thick line in both the figures represents the decision boundary for soft margin SVMs trained with different choice of parameter C.

a. $C = \infty$ is the same as Hard margin SVM.

b. $C = 0$ is same as Hard margin SVM.

c. Fig 2a corresponds to a small value of C.

d. Fig 2a corresponds to a large value of C.



(a) (b)

Figure 2

(3) Consider (w, b) defines the decision boundary of a hard margin SVM model. What is the range of values of $(w^T x + b)$ for a point x inside the margin defined by the support vectors? [2 marks]