CSE Ph.D. Qualifying Exam, Fall 2022

Data Analysis

Please answer three of the following four questions. All questions are graded on a scale of 10. If you answer all four, all answers will be graded and the three lowest scores will be used in computing your total. Show all your work and write in a readable way.

1. **Recommendation Systems**

   You have collected the following ratings of popular comedy TV shows from five users:

	Rating				
	Alice	Bob	Charles	David	Eugene
Friends (F)	5	3		2	4
The Office (TO)	4				4
Arrested Development (AD)	4			4	
The Bing Bang Theory (BBT)		2	1		
The Invincible (IV)	1		1	1	

Figure 1: TV Shows Rating Matrix.

(a) (3 pts) To generate recommendations, you adopt the following policy: "if a user U likes show X, then U will also like show Y". You implement this by *maximizing the cosine similarity* between the ratings of items X and Y. Your policy also states that you will only make a recommendation to user U if (a) U has not already watched or rated Y and (b) U's rating of show X is at least 3.

Using this policy, which TV show would be recommended to Eugene? Show the comparisons that you made.

(b) (3 pts) Next, you design a recommendation system to rank TV show to find the 'Best TV Shows of All Times', using the following formula: $ratings(i) = a + b(i)$. In this formula, you set a as a global average rating term and $b(i)$ as an show i's bias score. You first fit this model to calculate a as the mean of all ratings across the dataset, and in the process, you calculate $b(i)$ = average rating given to show i – global average rating a.

You rank the shows according to their bias scores (higher bias score is ranked higher). Which show, among the five shows shown in Table 1, would be the Best TV Show and which one would be the Worst TV show? Show your calculations.

(c) (2 pts) You come up with the idea of training a deep learning-based recommendation system model, namely the Neural Collaborative Filtering (NCF) model, on your large dataset to create better recommendation models. Your large dataset has 10 million ratings given by approximately 100,000 users to approximately 1,000,000 movies.

Your NCF model first generates 8-dimensional user and item embeddings. Then you pass the embeddings through two fully-connected neural CF layers with sizes 8x16 and 16x16 dimensions. Finally, this is passed through a 16x1 output layer with ReLU activation to produce a single prediction value of recommending an item to a user. You train the model for 10 epochs with back-propagation.

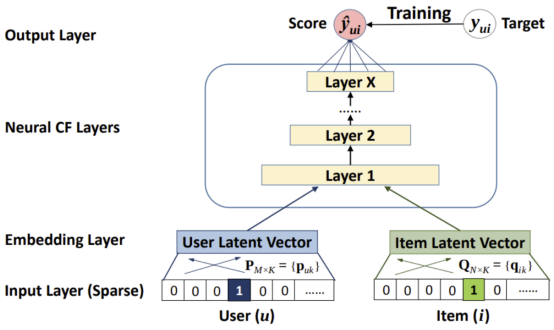

Figure 2: The Neural Collaborative Filtering (NCF) model.

After training the model, you find that the model does not perform well. What changes can you make to the model or parameters to potentially improve the performance? Give at least three options. Note that you cannot choose a different model now.

(d) (2 pts) Recommender systems are typically trained on a subset of training data, due to the large size of the entire dataset. A popular dataset sampling strategy is to take the interactions between the most active users and most interactive items. Specifically, all users with less than k interactions are removed and all items with less than k interactions are removed. The recommender model is trained on the remaining dataset. Describe a bias that this sampling strategy can introduce in a recommender model trained on this dataset.

2. **Learning Theory and VC Dimension**

Consider a binary classification problem with the hypothesis class of two-dimensional thresholds, $H = \{h_{a,b} : a \in \mathbb{R} \text{ and } b \in \mathbb{R}\}$ where:

$$h_{a,b}(\mathbf{x}) = \begin{cases} 1 & \text{if } x_1 \leq a \text{ and } x_2 \leq b \\ -1 & \text{otherwise} \end{cases}$$

and $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{-1, 1\}$.
(a) [3 pts] Describe an algorithm for computing the ERM for this class in the realizable case, assuming the $0-1$ loss is used. State the computational complexity of the algorithm in the context of a training data set of size m.
(b) [7 pts] What is the VC dimension of this hypothesis class? Provide a complete proof.

3. **Gaussian Discriminant Analysis**

Consider a two-class classification problem with data in $\mathbb{R}^d \times \{1, 2\}$. Gaussian discriminant analysis solves this problem by modeling the class-conditional distributions as a Gaussian:

$$p(X|Y = i) \sim \mathcal{N}(\mu_i, \Sigma_i), \quad p(Y = i) = \pi_i \quad \text{for} \quad i \in \{1, 2\}.$$

Here, $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal density with mean μ and covariance matrix Σ.

(a) [2 pts] Given a large set $(X_i, Y_i)_{1 \leq i \leq N}$ of independent data, derive the maximum likelihood estimates of the parameters$(\pi_i, \mu_i, \Sigma_i)_{i \in \{1,2\}}$.

(b) [2 pts] Once the parameters of the model have been obtained, we want to classify a new data point $\tilde{X}$ by maximizing the conditional probability $p(Y|X = \tilde{X})$. Formulate the resulting decision rule mathematically.

(c) [1 pt] Consider the special case where we make the restriction that $\Sigma_1 = \Sigma_2$. What is the maximum likelihood estimator for $(\pi_i, \mu_i, \Sigma_i)_{i \in \{1,2\}}$ in this case?

(d) [2 pts] Show that in the setting of (c), the decision boundaries are linear. In particular, show that log-probability-ratio is of the form

$$\log\left(\frac{p(Y = 1|X = x)}{p(Y = 2|X = x}\right) = c + v^\top x, \tag{1}$$

where c and v are independent of x.

(e) [2 pts] Consider the univariate case $d = 1$ and assume that our data set is such that $\mu_1 = -1$, $\mu_2 = 1$, $\Sigma_1 = \Sigma_2 = 1$. Now imagine we add the data points $(X_{N+1}, Y_1) = (-\lambda, 1)$ and $(X_{N+2}, Y_2) = (\lambda, 2)$ to our data set and reapply the methodology from (c). What happens, for a given input value x, to the left hand side of (1) as λ goes to infinity. What does this mean for the classifier?

(f) [1 pt] What would you expect if in (e), we were using logistic regression instead of the class-conditional Gaussian approach? Explain your prediction.

4. **Support Vector Machine**

You are provided with $m > 1$ data points $\{x_j \in \mathbb{R}^n\}_{j=1}^m$ of which at least d, with $1 < d \leq m$ are distinct. Let $X = [x_1, ..., x_m]$ and consider the one class SVM problem:

$$\begin{aligned} \min_{R \in \mathbb{R}, a \in \mathbb{R}^n, s \in \mathbb{R}^m} \quad & R^2 + C\mathbf{1}^\top s \\ s.t. \quad & ||x_i - a||_2^2 \leq R^2 + s_i, \quad i = 1, ..., m, \\ & s \geq \mathbf{0}. \end{aligned}$$

(a) [1 pts] Show that this is a feasible convex program and that strong duality holds.

[Hint: let $r = R^2$]
(b) [1 pts] Write down the KKT conditions.
(c) [2 pts] Show that $\alpha^* \neq \mathbf{0}$ and that if $C > 1/(d-1)$ then $(R^2)^* > 0$ (harder).
(d) [2 pts] What are the support vectors for this problem?
(e) [2 pts] Derive the dual problem.
(f) [2 pts] Assume $C > 1/(d-1)$. Given the dual solution, how should a and R^2 be selected?