CSE Ph.D. Qualifying Exam, Spring 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. **Random Forests**

   a. [3 pts] *Random forests* is a modification over *bagging* decision trees. The random forests improves variance reduction (over bagging) by reducing correlation among trees. Briefly explain how this correlation reduction ("de-correlation") among trees is achieved when growing the trees.

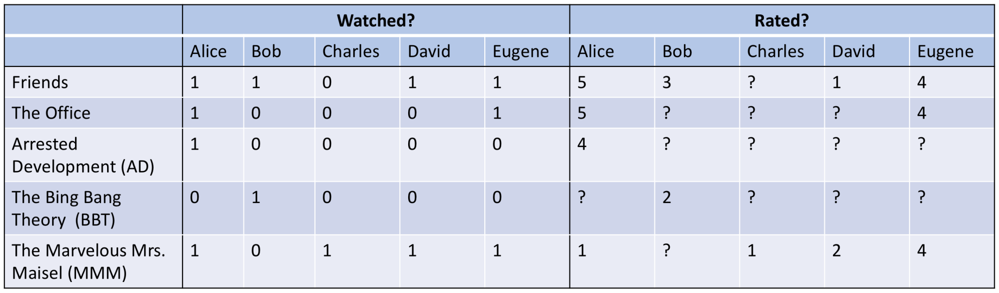
   b. [3 pts] Random forests are generally easy to implement and to train. It can be fit in one sequence, with cross validation performed along the way (almost identical to performing N-fold cross-validation, where N is the number of data instances), through the use of *out-of-bag* (OOB) samples. Explain why using OOB samples eliminates the need for setting aside a test set for evaluating a random forest, and how this leads to more efficient training.

   c. [2 pts] List the *model hyperparameters* and *model parameters* of a random forest.

   d. [2 pts] Alice and Bob are data scientists debating whether a random forest is an "interpretable" model. Alice argues that it is interpretable, while Bob argues that its interpretability is limited. Briefly discuss why they may both be correct.

2. **Recommendation Systems**

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



	Watched?					Rated?				
	Alice	Bob	Charles	David	Eugene	Alice	Bob	Charles	David	Eugene
Friends	1	1	0	1	1	5	3	?	1	4
The Office	1	0	0	0	1	5	?	?	?	4
Arrested Development (AD)	1	0	0	0	0	4	?	?	?	?
The Bing Bang Theory (BBT)	0	1	0	0	0	?	2	?	?	?
The Marvelous Mrs. Maisel (MMM)	1	0	1	1	1	1	?	1	2	4

Figure 1: TV Shows Rating Matrix.

   (a) (4 points) To generate recommendations, you adopt the following policy: "if a user U likes item X, then U will also like item 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 item X is at least 3.

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

(b) (3 points) 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 term and $b(i)$ as an item'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 $b(i)$ to be the remainder value per item.

You rank the items according to their bias scores (higher bias score is ranked higher). Which item, 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) (3 points) 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.

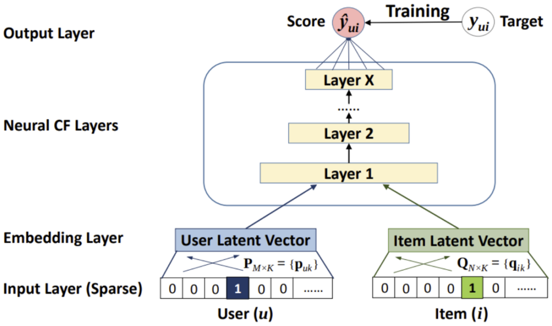


Figure 2: The Neural Collaborative Filtering (NCF) model.

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.

3. **Bayesian Linear Regression and Regularization**

Linear regression is a model of the form $P(y|\mathbf{x}) \sim N(\mathbf{w}^{\mathrm{T}}\mathbf{x}, \sigma^2)$ from a probabilistic point of view, where $\mathbf{w}$ is a d-dimensional vector. In ridge regression, we add an l-2 regularization term to our least squares objective function to prevent overfitting. Given data $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$, our objective function for ridge regression is then:

$$J(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^2 + \lambda \mathbf{w}^{\mathrm{T}}\mathbf{w}.$$

We can arrive at the same objective function in a Bayesian setting, if we consider a maximum a posteriori probability (MAP) estimate and assume $\mathbf{w}$ has the prior distribution $N(0, f(\lambda, \sigma)\mathbf{I})$.

(a) [3 pts] Write down the posterior distribution of $\mathbf{w}$ given the data.

(b) [7 pts] What $f(\lambda, \sigma)$ makes this MAP estimate the same as the solution to optimizing $J(\mathbf{w})$?

4. **Gaussian statistics**

You were hired to accompany an expedition to study the legendary *mathematodon*, an enormous amphibian mammal living exclusively on the *shepherd's islands*, hundreds of nautical miles southwest of Australia. After arriving on the archipelago, you begin collecting data at each adult mathematodon sighting, including the size of its hoofs and its height. After collecting a *large* number N of measurements, you gather them into an $N \times 2$ matrix $\mathbf{A}$, with the first column corresponding to the diameter of the mathematodon's forehoofs and the second to its height.

(a) (1 Point) How would you compute the mean $\mathbf{m}$ and covariance $\mathbf{C}$ of the joint distribution of the diameter of the forehoof and height of a mathematodon.

(b) (3 Points) You discover the imprint of a mathematodon forehoof of diameter d. You were unable to observe the mathematodon itself, but assume that hoof size and height are jointly Gaussian. Use the maximum likelihood criterion to estimate the parameters of their joint Gaussian distribution. What is the optimal estimate for the height of the unseen mathematodon (as a function of $\mathbf{A}$ and d). What is the mean squared error of this estimate?

(c) (1 Points) In the last problem, would it have been enough to assume that the two variables are marginally Gaussian to arrive at the same conclusion? Explain your answer.

(d) (2 Points) The *Bregmann rule* suggests that the variability of both hoof size and height varies from north to south. For $s \in [-1, 1]$ denoting the north-south position within the archipelago, consider the parametric form of the Bregmann rule given by the mean $\mathbf{m}$ and covariance $\mathbf{C} + \alpha s \mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $\alpha \in \mathbb{R}$ is a parameter. Let the vector $\mathbf{s} \in \mathbb{R}^N$ contain the s-values associated to the data collected in $\mathbf{A}$. What is the range of values for α that specify a valid Gaussian Process model throughout the entire archipelago? Formulate a maximum likelihood criterion for determining α.

(e) (3 Points) After returning from the expedition, your colleagues point out that the forehoof-sizes and heights of male and female mathematodons are likely following different distributions. *DARN! Beginner's mistake!* You forgot to note down whether you observed male or female mathematodons! Thankfully you still have your raw data **A**. Describe how you could try to recover the missing information by modeling your data as a Gaussian mixture model and derive an Expectation-Maximization algorithm to fit this model to the data given by **A**.