Question 1: What are the other topics you want to add to Chapter 3: ML Basics?

I would like to have a chapter dedicated to handle large and real data set at some specific healthcare companies.

Question 2: What are the typos and improvements you found from these chapters?

I did not find the typos in the chapter. I would like to see the improvement of handling large data set into the chapter as well.

Question 3: What other resources (e.g., book, blog, online tutorial) do you recommend on these topics?

One of the classes that I took at UIUC has a profound impact on my understanding of ML. Here is the link to the class:

https://publish.illinois.edu/liangf/teaching/stat-542/stat-542-lectures/

Question 4: If you have a classification problem on 500 10-dimensional data points, what algorithms would you try first? What algorithms would you try last?

1. Prediction target: the prediction target is a discrete variable, which is also known as a classification problem. Since the matrix has at least 50 number of observations per dimension, the first algorithm that I would try is logistic regression.

2. Cohort Construction: The patient cohort can be determined based on one of the attributes such as age or patients who are at higher risk of developing the disease.

3. Feature Construction: Feature construction is already selected since there are only 10 dimensions.

4. Feature Selection: Feature selection can be done with lasso or ridge regression and then compare the performance with the logistic regression with the full 10 dimensions.

5. Predicitve Model:

- Logistic regression: it is a simple model and can achieve a base line that other models can rely on.
- Lasso / ridge logistic regression: lasso reduces the number of dimensions while ridge only shrinks some of the coefficient magnitudes. Lasso may pick a different model than ridge so it is better to try them both.
- Non-linear model such as a decision tree: Logistic regression is not as flexible as a decision tree and tend to underfit the training data set. A simple decision tree may outperform the logistic regression model.
- Non-linear model such as a random forest tree: A decision tree may overfit the training data set so a random forest can help to eliminate the overfitting problem and de-correlate the tree by randomly choosing a subset of the training data at each split (without replacement).

6. Performance Evaluation Depending on the problem statement, an accuracy or F1 score on the test set can be used as the performance evaluation metric. A good threshold can be chosen based on the AUC curve or Precision Recall Curve, which depends on the trade-off between the TPR and FPR.

Question 5: If you have to cluster a large dataset (e.g., 1billion points), what algorithms would you use? what steps would you try to speed up the process?

Since K-Means takes O(k*ndi) where k is the number of clusters, d is the dimension of the data set, n is the number of training examples (1 billion) and i is the number of iterations till convergence, K means is* *not** the approriate choice. One of the scalable methods being mentioned in the CS412 Introduction to Data Mining is STING (A Statistical Information Grid Approach). Each cell at a high level contains a number of smaller cells of the next lower level. Statistical information of each cell is calculated and stored beforehand. Since patient's statistics of each subgroup is more similar to each other, such as age ranges or cholesterol variances, the STING approach could try to cluster different subgroups of patients based on those statistics. I can try to increase the number of grids or reduce the number of layers in order to speed up the process.