Week 3 Reflections

Course	CS 598 - Deep Learning for Healthcare
Submitted By	Rohit Bhagwat (rohitb2@illinois.edu)

General Summary

After having introduced us to basics on Health data, this week was designed to provide high level introduction to Machine Learning - and perhaps provide a direction for the students to start thinking about its application in Healthcare. The videos highlighted various types of learning, the high level steps involved in building a model and finally the metrics to evaluate the model.

Below is a high level summary of what we learned this week

- Supervised Learning
    - Typical Predictive Modeling Pipeline
        - Identifying the Prediction Target → Construct the Cohort → Feature Construction → Select appropriate Features → Build Predictive Model → Evaluate Performance → Start again
        - Key concepts: Types of Predictive Models → Regression, Classification; Splitting the Data into Training, Validation and Test is critical
- Unsupervised Learning
    - Dimensionality Reduction - why its important to work on a subset of dimensions instead of working with hundreds or thousands of dimensions
        - Two techniques
            - Scalar Vector Decomposition
            - Principal Component Analysis
    - Clustering - primarily K-means clustering
    - Key Concepts: How to visualize SVD, how it is fundamental for PCA and dimensionality reduction.
- Evaluation Metrics
    - Classification: Confusion Matrix: Accuracy, Prevalence, True Positive Ratio, Positive Predictive Value, F1 score etc.. ; Receiver Operating Characteristic (ROC)
    - Regression: Mean Absolute Error (MAE), Mean Square Error (MSE), R2
        - Clustering: Rand Index, Mutual Information, Silhouette Coefficient

Questions

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

As an introduction I thought the topics covered in this chapter were adequate, especially if the students have not taken a prior course on Machine Learning (Practical Statistical Learning or Applied Machine Learning as an example). I believe some of the advanced algorithms are going to be covered future weeks, ex: Deep Networks, Neural Networks etc..

If I had to suggest a topic to be covered, I'd say perhaps some discussion / references on Linear Algebra can be useful for students who have not taken Statistics

- Linear Algebra basics
- Standard Deviation, Variance, Covariance etc..
- Other Classification algorithms: Gradient Descent, Naive Bayes, Decision Tree, Random forest etc..

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

I couldn't find any typos (honestly I wasn't looking to find typos, was more keen in understanding the concepts presented in the slides and the chapter).

In my humble opinion, I thought during the weekly slides some real life examples on applications to Healtcare could've proved to be very engaging. As an example a case study, each on applying Regression and Cluster model and its results.

What other resources (eg., book, blog, online tutorial) do you recommend on these topics?

I would highly recommend the following course and its related book for Machine Learning -

- CS 498: Applied Machine Learning
- Books: Applied Machine Learning by David Forsyth, Probability and Statistics for Computer Science by David Forsyth
- YouTube videos that are my go to reference for various concepts around Machine Learning
  - StatQuest with Josh Starmer: https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw
  - Steve Brunton: https://www.youtube.com/channel/UCm5mt-A4w61lknZ9lCsZtBw

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

Given the fact that we have 500 data points with 10 dimensions I think we can try one of the following -

- Linear Regression: The dataset is small so Linear regression or Linear SVM is an algorithm we can go with.
- K-means clustering: The K-means algorithm is best suited for finding similarities between entities based on distance measures with small datasets - considering the dataset is small I think k-means algorithm may be applied here.

The decision will also depend upon what is the objective of the exercise - do we know the labels or not.

In this case I don't think Decision trees could be applied here.

If you have to cluster a large dataset (eg., 1 billion points), what algorithms would you use? What steps would you try to speed up the process?

Considering we're trying to cluster, I believe k-means clustering is a valid algorithm to be chosen. Since we have large data volume we may need to apply techniques to speed up the process. I'm not familiar with these processes, but it seems Mini Batch k-means clustering (https://scikit-learn.org/stable/modules/clustering.html#mini-batch-kmeans) offers a mechanism to speed up the process by processing mini-batches in order to reduce computation time.