FINAL EXAM

INTRODUCTION

Please answer the question listed below and turn in a PDF of your answers. This is a portion of your final exam; the other portion is the applied coding exercise.

QUESTIONS

1) The gradient, the hessian, and optimization: Discuss the gradient and the hessian as they apply to optimization in machine learning. Be sure to address (at least) the following points in your discussion. Your answer should be at least a few paragraphs long, if not longer. If you use a source to support your answer, cite it.
    a. Define the gradient and the hessian in your own words, as they apply to optimization in machine learning. How does this relate to first and second order derivatives?
    b. How does optimization in machine learning differ from classic function optimization in mathematics?
    c. What is a local minimum versus a global minimum? What is a saddle point? Why are these problematic in ML optimization?
    d. Summarize the general steps an optimization algorithm takes (Usually the computer does this for you, so you don't code these steps)
    e. What are the tradeoffs in using gradient based optimization methods versus hessian based methods?
    f. Why has the field typically used Newtonian (2nd order derivative) optimization on unregularized GLM models? Under what specific circumstance do these second order methods become challenging to implement in practice? Why?
    g. Why do we tend to use first order methods like SGD or RMSProp in deep learning?
2) Machine learning versus inferential methods (like GLMs): Write a multiple paragraph discussion about ML versus inferential model like GLMs. Address all the points listed below:
    a. What is the overall goal of a machine learning prediction model?

b. Discuss the concept of training error versus generalization error, and how these two types of error are estimated in practice.
c. What different forms of cross validation can be implemented? Name a few different schemes for cross validation. Discuss their strengths and weaknesses.
d. Define model capacity. Is more capacity always better? Why or why not? How much capacity is optimal?
e. Define regularization – what are different ways we can regularize models in machine learning?
f. What is the difference between hyperparameters and model parameters? How do we optimize model parameters? How do we optimize model hyperparameters? If we implement an elastic net regression, what are the model parameters, and what are the model hyper parameters?
g. A clinic manager would like you to create a patient-no show risk model. The goal is to use the no-show risk to schedule extra patients…if a patient is at a high risk to no-show, we will book extra appointments that day, assuming some patients will no-show for their appointment. Should you use a machine learning approach here, or should you use classic inferential models like a GLM? Why? Justify your response.
h. A physician would like to understand whether drug A or drug B is more effective at controlling blood pressure. We have some data from last year where some patients were exposed to drug A, and some patients were exposed to drug B. Should you implement a machine learning approach to predict blood pressure here (like a GBM), or should you implement inferential models like GLM? Why? Justify your response.

3) Online learning / SGD –
   a. What is 'online learning'?
   b. How does implementing online learning in sk-learn allow us to fit models on data that is larger than memory?
4) Scikit-learn, Spark, Dask, and h2o: Based on your readings and experience in this class thus far, what do you think the strengths and weaknesses are of using these frameworks for data management and machine learning?
   a. For each technology, address its capabilities for working with data larger than memory, as well as if it can fit models in parallel (using multiple cores or clusters).
   b. Compare and contrast the available models for each framework. Can you fit all the same types of models in each framework?
   c. Compare and contrast the available optimization methods in each framework.

5) Deep learning: Write a few paragraphs discussing deep learning that address the points below.
    a. How does deep learning generally differ from other types of machine learning, like Elastic net, random forest, or GBM? How is it similar?
    b. Why has deep learning received so much attention in the past 8 years? What conditions in data science have contributed to this focus on deep learning?
    c. Do you think deep learning will reshape data science in healthcare in the coming years? Why or why not? What tasks will deep learning succeed or fail at?

