CSCI 4364/6364 Machine Learning, Midterm Exam

Section 81

Instructor: John Sipple

Thursday, October 27, 2022

Time limit: 1:15

Instructions:

This exam consists of 10 questions, and each question is worth 10 points. This exam is divided into two parts. The **design questions** are related to a single, specific real-world problem that you may encounter in practice. Each answer should be about a paragraph or two in length, and should relate to the original problem statement. The **concept questions** are about important topics that we discussed in lecture and were part of the required readings. Each answer should be about a paragraph or two in length.

Complete your answers in the exam notebook. Clearly mark your answer with the question ID (e.g., D1, C1, C2…) in your exam booklet, and please write as legibly as possible.

The exam is open book, and you may use electronic devices such as laptops, ipads, etc. However, during the exam, you may not communicate with anyone remotely (e.g., chat) or in the class. If you have any questions, please raise your hand and I will try to clarify.

Good luck!

Page intentionally left blank.

Design Question. Some companies get flooded with resumes and job applications, overwhelming the HR departments. These HR departments often employ human resume screeners who scan resumes, and reject them or pass them on to recruiters. Since they have only a minute or two for each resume and application, these screeners have a high false negative error rate (i.e., they reject many qualified applicants). Perhaps we can design a machine learning model to improve performance.

[D1, 10 points] Given a sample of 100,000 resumes of rejected applicants (negatives), and 10,000 resumes of candidates who received job offers (positives), propose a method of converting the raw text sample onto a structured table for training and testing a binary classifier. (You may assume you have a one-word, or two-word vocabulary to create a feature vector.)

[D2, 10 points] Assuming that your training and test data are very sparse (i.e., features are mostly consisting of zeros), propose (and describe) a transformation that reduces the input vector into a much denser feature vector in lower dimensionality.

[D3, 10 points] Design a multitask feedforward neural network that predicts both applicant qualification and gender (male vs. female). In your design, propose suitable loss functions, an architecture, optimizer, and regularization techniques.

[D4, 10 points] In 2018, Amazon abandoned their candidate screener model because many of the positive examples were dominated by men, and the model downgraded applicants with terms like "*women's*" as in "*women's chess club*", and favored applicants with verbs commonly found in male candidates resumes, like "*executed*" or "*captured*". Considering the binary cross-entropy loss function and general gradient descent algorithm, explain how a model might learn a bias favoring male applicants.

[D5, 10 points] HR departments may ask what words (or word pairs) influenced the model to select or not select a candidate. Explain why a simple algorithm like logistic regression or decision tree would more readily provide an answer than a sophisticated neural network with many layers.

Concept Questions. Please answer the following questions in your own words. Your answers should be detailed, and contained in a paragraph or two.

[C1, 10 points] In class we discussed decision trees (DT) and their vulnerability to overfitting to the training data. Explain why DTs overfit and describe at least two methods that reduce the overfitting problem.

[C2, 10 points] Describe the condition number of the Hessian matrix. How might a large condition number affect basic gradient descent?

[C3, 10 points] Describe the significance of the kernel function, such as the radial basis function, in Support Vector Machines.

[C4, 10 points] Extending the basic gradient descent algorithm (Goodfellow, Algorithm 8.1), write a modified gradient descent algorithm that replaces the constant learning rate ϵ with a function of the dot product of the current and previous gradient vector. In other words, the adaptive learning rate will be greatest when the dot product of the last two gradients is highest.

Algorithm 8.1 Stochastic gradient descent (SGD) update

Require: Learning rate schedule $\epsilon_1, \epsilon_2, \ldots$
Require: Initial parameter $\boldsymbol{\theta}$
  $k \leftarrow 1$
  while stopping criterion not met **do**
    Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
    Compute gradient estimate: $\hat{\boldsymbol{g}} \leftarrow \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$
    Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon_k \hat{\boldsymbol{g}}$
    $k \leftarrow k + 1$
  end while

[C5, 10 points] Consider a binary classifier that outputs a prediction between 0 (class A) and 1 (class B) with an associated ROC-AUC curve. What effect does the decision threshold have on precision and recall for class B?