CSCI 4364/6364 Machine Learning, Midterm Exam

Section 81

Instructor: John Sipple

Wednesday, November 2, 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**.

Good luck!

Page intentionally left blank.

Design Question. Rare diseases, such as cystic fibrosis, with fewer than 5 patients per 10,000 can be difficult to diagnose because of similar symptoms with common illnesses and the low frequency of occurrence. In a survey published in 2013, on average, it takes more than five years, eight physicians, and two to three misdiagnoses until a rare disease receives the correct diagnosis. To aid in more rapid diagnosis of a single rare disease, occurring in 1 out of 10,000 patients, a large hospital network has provided you with a sample of one million patient health and treatment records. Your objective is to train and evaluate a classifier to identify patients with this rare disease.

[D1, 10 points] Describe how you might convert the patient records into a training and test sample. What are the rows, columns (with examples), and labels for the test and training set? How would you handle ordinal (such as age group or risk level) and categorical features (such as ethnicity or gender)?

[D2, 10 points] Decision Trees (DT) are commonly used as diagnostic aids because they are easily interpreted by medical professionals. However, they are prone to overfitting. Describe two methods to reduce overfitting a DT.

[D3, 10 points] Ensemble methods, like Random Forests (RF), yield higher accuracy than DTs, but are more difficult to interpret. Describe a method of ranking features by their importance (i.e., most to least important features) as an aid to understanding the model.

[D4, 10 points] Describe how you may enhance the positive (with disease) examples using dataset augmentation. How would you avoid creating examples that overlap with the negative (healthy) population?

[D5, 10 points] Design a feedforward neural network that predicts the rare disease. In your design, propose a suitable loss function, an architecture (input, hidden, and output layers), optimizer, and regularization techniques.

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] As you increase the regularization coefficient on Lasso or Ridge Regression, explain what happens to variance, bias, and irreducible errors in the model against the test set.

[C2, 10 points] In class we discussed that random forests are trained with a random subset of the features at each split. Describe the advantage this technique has over using all the features on each split.

[C3, 10 points] Suppose you are working with a Support Vector Machine using the Radial Basis Function (RBF) kernel. How would adjust the RBF gamma and C (misclassification cost) parameters to reduce underfitting**?**

[C4, 10 points] Applying the concept of multi-task learning from Goodfellow 7.7, rewrite the forward propagation algorithm (Algorithm 6.3) with two tasks that predict $\hat{y}_{(1)}$ and $\hat{y}_{(2)}$ for targets $y_{(1)}$ and $y_{(2)}$ instead of just predicting single task of predicting $\hat{y}$ for target y. Assume the input x passes through l shared layers, and then splits into $l_{(1)}$ and $l_{(2)}$ task-specific layers, with weights ${W_{(1)}}^{(k)}$ and ${W_{(2)}}^{(k)}$ and biases ${b_{(1)}}^{(k)}$ and ${b_{(2)}}^{(k)}$.

Algorithm 6.3 Forward propagation through a typical deep neural network and the computation of the cost function. The loss $L(\hat{\boldsymbol{y}}, \boldsymbol{y})$ depends on the output $\hat{\boldsymbol{y}}$ and on the target $\boldsymbol{y}$ (see section 6.2.1.1 for examples of loss functions). To obtain the total cost J, the loss may be added to a regularizer $\Omega(\theta)$, where θ contains all the parameters (weights and biases). Algorithm 6.4 shows how to compute gradients of J with respect to parameters $\boldsymbol{W}$ and $\boldsymbol{b}$. For simplicity, this demonstration uses only a single input example $\boldsymbol{x}$. Practical applications should use a minibatch. See section 6.5.7 for a more realistic demonstration.

Require: Network depth, l
Require: $\boldsymbol{W}^{(i)}, i \in \{1, \ldots, l\}$, the weight matrices of the model
Require: $\boldsymbol{b}^{(i)}, i \in \{1, \ldots, l\}$, the bias parameters of the model
Require: $\boldsymbol{x}$, the input to process
Require: $\boldsymbol{y}$, the target output

$\boldsymbol{h}^{(0)} = \boldsymbol{x}$
for $k = 1, \ldots, l$ **do**
  $\boldsymbol{a}^{(k)} = \boldsymbol{b}^{(k)} + \boldsymbol{W}^{(k)} \boldsymbol{h}^{(k-1)}$
  $\boldsymbol{h}^{(k)} = f(\boldsymbol{a}^{(k)})$
end for
$\hat{\boldsymbol{y}} = \boldsymbol{h}^{(l)}$
$J = L(\hat{\boldsymbol{y}}, \boldsymbol{y}) + \lambda \Omega(\theta)$

[C5, 10 points] When applying principal components analysis (PCA) to reduce dimensionality before executing a classifier algorithm, how can the explained variance help you choose the number of components to balance between efficiency (minimizing the number of features), and classification accuracy?