Spring 2017

COMPUTER SCIENCES DEPARTMENT
UNIVERSITY OF WISCONSIN–MADISON
PH.D. QUALIFYING EXAMINATION

Artificial Intelligence

Monday, January 30, 2017

GENERAL INSTRUCTIONS

1. This exam has 6 numbered pages.
2. Answer each question in a separate book.
3. Indicate on the cover of each book the area of the exam, your code number, and the question answered in that book. On one of your books, list the numbers of all the questions answered. Do not write your name on any answer book.
4. Return all answer books in the folder provided. Additional answer books are available if needed.

SPECIFIC INSTRUCTIONS

You should answer all four questions.

POLICY ON MISPRINTS AND AMBIGUITIES

The Exam Committee tries to proofread the exam as carefully as possible. Nevertheless, the exam sometimes contains misprints and ambiguities. If you are convinced that a problem has been stated incorrectly, mention this to the proctor. If necessary, the proctor can contact a representative of the area to resolve problems during the first hour of the exam. In any case, you should indicate your interpretation of the problem in your written answer. Your interpretation should be such that the problem is nontrivial.

760 – MACHINE LEARNING

760-1 Supervised Learning

1. Define "overfitting" in the context of supervised learning.
2. Define "overfitting" in the context of graphical model structure learning.
3. For each of the following types of learning, describe what you believe to be the most effective special-purpose method to combat overfitting. By "special-purpose" we mean a method that does not easily generalize to most types of learning; bagging would be one example of a method that is *not* special-purpose.
   (a) Decision trees
   (b) Support vector machines
   (c) Logistic regression
   (d) Bayesian network structure learning
   (e) Deep neural networks

760-2 Randomization in Supervised and Reinforcement Learning

Randomization is often used in both supervised and reinforcement learning settings.

1. Several supervised learning approaches employ randomization, including *bagging*, *random forests*, and *dropout*. For each of these three approaches, describe what is randomly selected in the approach, and when during the learning process the random decisions are made.

2. What is the common rationale for using randomization in the three approaches listed above?

3. Describe how randomization is typically used in reinforcement learning and why it is important in this setting.

761 – ADVANCED MACHINE LEARNING QUESTIONS

761-1 Learning from Averages

In standard supervised learning, a training set consists of n training items of the form $(\mathbf{x}_i, y_i)$ for $i = 1 \ldots n$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector and $y \in \mathbb{R}$ is a continuous (for regression) or $\{0, 1\}$ (for binary classification) label.

However, in some real applications the label is only available in aggregated form over training subsets. For example, the feature vector may contain a patient's age, and y is that individual's dementia cognitive test score. A study may only publish the average score of different age groups, not individual scores. We still assume, however, that the feature vector is observed.

Formally, let $X_1, \ldots, X_K$ be a fixed and known K-way partition of the feature space: $\cup_{k=1}^{K} X_k = \mathbb{R}^d$. Let $\pi(\mathbf{x}) \in \{1, \ldots, K\}$ be the partition index of item $\mathbf{x}$, namely $\mathbf{x} \in X_{\pi(\mathbf{x})}$. To the learner, each training item is represented by $(\mathbf{x}_i, \bar{y}_{\pi(\mathbf{x}_i)})$, where the true label is replaced by the average label in that partition:

$$\bar{y}_k = \frac{\sum_{i=1}^{n} y_i 1[\pi(\mathbf{x}_i) = k]}{\sum_{i=1}^{n} 1[\pi(\mathbf{x}_i) = k]}.$$

Here $1[z]$ is the indicator function taking value 1 if z is true, and 0 otherwise.

Your task is to learn from the partition $X_1, \ldots, X_K$ and the modified training set $(\mathbf{x}_i, \bar{y}_{\pi(\mathbf{x}_i)})$ for $i = 1 \ldots n$.

1. Sketch a picture of an example data set for linear regression (the original y is continuous), in the simple case of $d = 1$ and a modest K (around 4 or 5). Schematically show the original training set, the modified training set, and the linear fit.

2. Design an approach to perform linear regression on such data (not necessarily in 1D). You should clearly define your model, and discuss how you may estimate its parameters from the data. For the latter a full derivation is not necessary, but your answer should contain enough technical details to explain the key steps.

3. The same as question 2 but for binary classification where the original $y \in \{0, 1\}$. Note, in this case, $\bar{y}_k \in [0, 1]$ is the fraction of training items with label 1 in partition k.

761-2 Empirical Risk Minimization

Consider the following prediction problem. You are given training items $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where each $\boldsymbol{x}_i \in \mathbb{R}^d$ and each $y_i \in \{-1, 1\}$. The samples are independent and identically distributed from an unknown distribution $\mathcal{D}$. This problem will analyze the performance of linear classifiers of the form

$$\widehat{y} \ = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}) \ .$$

Define the risk as

$$R(\boldsymbol{w}) \ = \ \mathbb{P}(y \neq \text{sign}(\boldsymbol{w}^T \boldsymbol{x})) \ ,$$

where the expectation is with respect to a random item $(\boldsymbol{x}, y) \sim \mathcal{D}$. Imagine we have p candidate classifiers $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$ and we want to pick the best one.

1. Propose an empirical risk minimization (ERM) procedure for this task.

2. Let $\widehat{\boldsymbol{w}}$ be the solution to the ERM procedure, and let

$$\boldsymbol{w}^\star \ = \ \arg \min_{k \in \{1,\ldots,p\}} \mathbb{P}(y \neq \text{sign}(\boldsymbol{w}^T \boldsymbol{x})) \ ,$$

   the weight vector that minimizes the true risk. Bound the generalization error $\mathbb{E}[R(\widehat{\boldsymbol{w}})] - R(\boldsymbol{w}^*)$. You may ignore constant factors.

This page intentionally left blank. You may use it for scratch paper. Please note that this page will NOT be considered during grading.