CS 541-A Artificial Intelligence: Mid-Term Exam

Instructor: Jie Shen

10/25/2022, 18:30 – 21:00 EST

Instructions:

- Open book exam, feel free to use any resource but no electronics;
- Discussion is not permitted;
- Always give your answer and explain it;
- 20 points per problem, totally 110 points (20 * 5 + 10).

0. Your name. (10 pts)

1. Discuss one practical challenge in modern AI applications, and propose three potential approaches.

2. Give one example and one counterexample to each of the following statements, where X_1 and X_2 are random variables.

- $\mathrm{E}[X_1 X_2] = \mathrm{E}[X_1] \cdot \mathrm{E}[X_2]$
- $\mathrm{Var}[X_1 X_2] = \mathrm{Var}[X_1] \cdot \mathrm{Var}[X_2]$
- $\mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2]$

3. Consider that there is such a learning algorithm $\mathcal{A}$ that by loading n labeled samples into memory, it returns a hypothesis h_n with error rate less than $1/\sqrt{n}$.

- Assume we hope to learn a hypothesis with error rate $\epsilon = 2^{-200}$. Find the minimum sample size n.
- Suppose that the samples lie in $\mathbb{R}^d$ with $d = 2^{20}$ and we use double-precision floating-point format (64 bits) to represent a real. What is the memory cost to store the samples as you calculated above.
- Is it realistic to learn such a model on a single computer? If not, propose an alternative approach.

4. The classic PAC learning model of Valiant'84 made two fundamental assumptions: 1) the distributions of the training data and testing data are the same; and 2) all instances are labeled correctly. The goal of PAC learning is to find a hypothesis whose error rate, i.e. the probability that it misclassifies a new sample, is upper bounded by $\epsilon \in (0, 1)$. Give two respective examples to illustrate that if either assumption is violated, PAC learning becomes impossible.

5. Consider two functions $F_1(\boldsymbol{w})$ and $F_2(\boldsymbol{w})$: both of them are strongly convex, but F_1 is smooth and F_2 is non-smooth. Suppose we apply GD to optimize these two functions. The following figure shows two convergence curves: a solid line and a dashed line. One is for $F_1(\boldsymbol{w}^t)$ and another for $F_2(\boldsymbol{w}^t)$.

- Explain which curve may correspond to F_1.
- Plot a possible convergence curve when applying stochastic GD to optimize F_1 in the same figure.
- Now recall that SGD will randomly select a sample to compute the stochastic gradient in each iteration. Use another figure to plot the convergence curves of SGD with 5 different trials, where one trial means we restart SGD with the same initial point and same step size.
- Finally, recall that the convergence guarantee of SGD is phrased in terms of expectation, but why do we often run it for one trial and still observe certain convergence property?