CS 541 Artificial Intelligence: Midterm Exam

Instructor: Jie Shen

10/19/2021, 6:30 - 9:00 pm EST

Instructions:

- Open book exam, feel free to use any lecture notes;
- Discussion is not permitted;
- Always give your answer and explain it (guaranteed 5 point for nonempty answer);
- 20 points per problem, totally 110 points (20 * 5 + 10).

0. Your name (10 pts)

1. Give a real-world AI problem which, from your point of view, cannot be solved in five years. If you are going to work on it, what is your first step to approach the main challenges?

2. An outlier is a data point that behaves abnormally, that its magnitude is out of control. For example, consider the age of a patient in his transcript: it is an outlier if it shows age = 2000, or if age = 25.5. Even if we observe a "normal" value of 85 is may also be an outlier provided that his actual age is 8.

Outlier may incur for a variety of reasons, e.g. human mistakes. In this case, we have to take such dirty data into algorithmic design. Consider the simple regression model:

$$y_i = \langle \boldsymbol{a}_i, \boldsymbol{w}_{\text{true}} \rangle + e_i, \quad \|\boldsymbol{w}_{\text{true}}\|_0 \leq k, \quad i = 1, \ldots, n. \tag{0.1}$$

In the above expression, $\boldsymbol{a}_i \in \mathbb{R}^d$ is the feature vector, $\boldsymbol{w}_{\text{true}}$ is the groundtruth model we aim to estimate, $e_i \in \mathbb{R}$ is the possible outlier for the ith sample and $y_i \in \mathbb{R}$ is the corrupted response.

- Given $\{\boldsymbol{a}_i, y_i\}_{i=1}^n$ for sufficiently large n, say $n \to \infty$, is it possible to learn the true model without any prior knowledge/condition on $\{e_i\}_{i=1}^n$?
- Now suppose that only n_1 samples are corrupted by outliers where $n_1 \ll n$. In other words, among $\{e_i\}_{i=1}^n$, $n - n_1$ of them are zeros (but we do not know which of them have zero values). Give a proper formulation under which it is possible to simultaneously recover $\boldsymbol{w}_{\text{true}}$ and detect the outliers.

3. Both principal component analysis (PCA) and random projection (RP) are widely used tools for dimension reduction. From a unified perspective, the only difference between them lies on the projection matrix. Suppose the data matrix $\boldsymbol{Y} \in \mathbb{R}^{d \times n}$. PCA projects $\boldsymbol{Y}$ onto $\boldsymbol{U}_{1:r}^{\top}\boldsymbol{Y}$ where $\boldsymbol{U}_{1:r}$ consists of the first r columns of $\boldsymbol{U}$ which is produced by singular value decomposition. In contrast, RP transforms the data into $\boldsymbol{AY}$ where $\boldsymbol{A} \in \mathbb{R}^{r \times d}$ is a random matrix (typically Gaussian).

- Give an example where PCA might be a better choice than RP;
- Give another example where RP is preferred;
- What are the main drawbacks of both methods, and what is your solution?

4. Suppose we have the following data from n patients:

	Age	Weight	Height	Gender	Blood Pressure	$\cdots$	Sharp Pain
Patient 1	z_{11}	z_{12}	z_{13}	z_{14}	**?**	$\cdots$	z_{1m}
Patient 2	z_{21}	z_{22}	z_{23}	z_{24}	z_{25}	$\cdots$	z_{2m}
Patient 3	z_{31}	z_{32}	z_{33}	z_{34}	**?**	$\cdots$	z_{3m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Patient n	z_{n1}	z_{n2}	z_{n3}	z_{n4}	z_{n5}	$\cdots$	z_{nm}

where some entries in the column Blood Pressure are missing (represented by the symbol "?"), and other columns are fully observed. Our goal is to estimate these missing values based on the current data matrix. Formulate it as a machine learning problem and state your solution.

5. Suppose we have the following data from n patients:

	Age	Weight	Height	Gender	Blood Pressure	$\cdots$	Sharp Pain
Patient 1	?	z_{12}	z_{13}	z_{14}	?	$\cdots$	z_{1m}
Patient 2	?	z_{22}	?	z_{24}	z_{25}	$\cdots$	?
Patient 3	z_{31}	?	z_{33}	?	?	$\cdots$	z_{3m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Patient n	?	z_{n2}	z_{n3}	?	z_{n5}	$\cdots$	?

where for each column and each row there are some missing entries (represented by the symbol "?").

- State how we can estimate all the missing values.
- Now consider that the above data matrix is fully observed, but some of its entries are corrupted by outliers. Give a proper formulation under which we are able to recover the clean data matrix.