亞洲大學 博士班資格考 試題

文件編號：AC-0054
機密等級：2 限閱

考試日期：111 年 8 月 3 日

考試科目	資料科學	系所別	人工智慧博士學位學程	命題教師	

1. Suppose your classifier needs to detect diabetes in human patients. “Positive” means the patient has diabetes. “Negative” means that the patient is healthy. Which is more important, precision or recall? Why?(10 points)

2. A p-value of 0.05 does not mean that there is a 95% chance that a given hypothesis is correct. Instead, it signifies that if the null hypothesis is true, and all other assumptions made are valid, there is a 5% chance of obtaining a result at least as extreme as the one observed. Does a smaller p-value indicate a more important research finding? Please explain your reasons with an example. (10 points)

3. For data science, the correctness of the data is very important. Therefore, it is necessary to confirm the distribution and integrity of the data through data preprocessing. **For numerical data, please describe the work of data preprocessing.** (20 points)

4. For the analysis work of data science, the larger the number of features (or independent variables) of numerical data, the results may be bad. The reason may cause overfitting or collinearity. Therefore, we need to analyze the relationship between independent variables through statistical methods, **please explain those statistical methods that can be used for analysis and the possible benefits.** (20 points)

5. Please explain the differences between feature selection and feature extraction. Give as examples two renowned algorithms for both tasks. (20 points)

6. Suppose that you have two variants of feature extraction algorithm *E* and three variants of classification algorithm *C* for developing a classifier. Please explain in details how you can experimentally compare the overall performance of *C* and *E* taking all their variants into account.
   Note: You have to show the **theoretical foundation** of your established comparison. (20 points)