數據科學方法期末考

總計 160 分。超過 100 分者均以 100 分計。

1. Consider the dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where the x_i are nonrandom and the y_i are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where α and β are unknown parameters to be estimated, ϵ_i's are i.i.d. random errors with $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2 < \infty$. Let $\hat{\alpha}$ and $\hat{\beta}$ be an estimator of α and β, which is the solution of

$$\min_{a,b} [y_i - a - bx_i]^2 + \lambda \cdot b^2,$$

where λ is a hyperparameter.

   (a) (20 分) Find explicit formulations of $\hat{\alpha}$ and $\hat{\beta}$.

   (b) (10 分) Find $\mathrm{E}[\hat{\beta}] - \beta$ as a function of λ.

   (c) (10 分) Find $\mathrm{E}[\hat{\alpha}] - \alpha$ as a function of λ.

   (d) (10 分) Find $\mathrm{Var}[\hat{\beta}]$ as a function of λ.

   (e) (10 分) By $\mathrm{MSE} = \mathrm{bias}^2 + \mathrm{Var}$, find $\mathrm{MSE}(\hat{\beta})$ as a function of λ.

2. The **Students performance in exams** dataset (StudentsPerformance.csv) contains students performance in three exams (math, reading, and writing) as long as their demographic and socioeconomic information.

   (a) (15 分) Investigate the dataset by some appropriate explorative data analysis. Are math, reading, and writing scores correlated with each other?

   (b) (15 分) Develop a regression model to predict their math, reading, and writing scores by the other variables. Evaluate your regression model by 10-fold cross-validation. Notice that most of the predictors are categorical.

   (c) (10 分) Does your model underfit or overfit? Explain why.

3. The **Credit card customers** dataset (BankChurners.csv) consists of 10000 customers mentioning their age, salary, marital_status, credit card limit, credit card category, etc. The purpose of this dataset is to predict whether a customer is gonna get churned

("Attrition_Flag") by the other 19 features (some of them are correlated), *so the bank can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction.* Notice that the dataset is unbalanced: we have only 16.07% of customers who have churned.

(a) (20 分) Build a binary classification model to predict who is going to leave their credit card services.

(b) (25 分) Compute the accuracy, precision, recall, F_1-score (the harmonic mean of precision and recall), and AUC by stratified 10-fold cross-validation of your model obtained in (a).

(c) (15 分) Among the above metrics, which one would you use to evaluate your model according to the purpose of this dataset? Explain your reason.