Takehome Final

STAT 409 - Data Science II

Due at 6/21/2023, 10:00pm

Submission Rules (Important!)

- You must sumbit your work (report & code) by e-mail (`sjshin4TA@gmail.com`, and cc to your email).
- **Due date: 6/21 (Wed) 10:00pm**
  - If I get your mail after 6/21 (Wed) 10:00pm, you will lose 30% of the credits you earned.
  - If I get your mail after 6/21 (Wed) 10:15am, NO credit!
- Additional rules:
  - Subject line of the email: `STAT409_Final_StuduentID` (ex: `STAT509_Final_2020150001`)
  - File name of your report: `STAT409_Final_StuduentID.pdf`
  - File name of your code: `STAT409_Final_StuduentID.R`
  - All functions and codes must be included in a single file.
- If you do NOT strictly follow these rules above, you additionally lose 5% of your credits.

Problems

[Notice] **You should be very precise and detailed to earn full credit.** 1. For a given $\mathbf{X} = \mathbf{x}$, the "Bayes" classification boundary, denoted with $f^*(\mathbf{x})$, is defined as

$$f^*(\mathbf{x}) = \underset{f(\mathbf{x}) \in \mathbb{R}}{\operatorname{argmin}} \; P\{Y f(\mathbf{x}) < 0\}$$

Suppose $P(Y = 1) = 1 - P(Y = -1) = 1 - \pi$ for a given $\pi \in (0, 1)$, show that

$$\operatorname{sign}\{f^*(\mathbf{x})\} = \operatorname{sign}\{p(\mathbf{x}) - \pi\}$$

where $p(\mathbf{x}) = P(Y = 1 \mid \mathbf{X} = \mathbf{x})$. (Hint, Note that $f^*(\mathbf{x})$ is a constant for a given $\mathbf{x}$, not a function.)

2. Consider a mean estimation problem under normality:

$$X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, 1).$$

(a) Show that MLE of μ is $\bar{X}_n$.
(b) Given $\bar{X}_n = 12$ with $n = 100$, please report 95% confidence interval of μ.
(c) One can consider a Bayesian inference. Toward this, we assume the following prior distribution on μ:

$$\mu \sim N(0, 100^2)$$

Please derive the posterior distribution of μ given $X_1, \cdots, X_n$. (Hint, This is the Normal-Noraml conjugate model)
(d) Bayesian interval estimator (known as credible interval) can be readily obtained from the posterior distribution, i.e., 95% credible interval of μ is $[c_l, c_u]$ such that

$$P(c_l \leq \mu \leq c_u \mid X_1, \cdots, X_n) = 0.95.$$

Given $\bar{X}_n = 12$ with $n = 100$, please report 95% Bayesian credible interval of μ.

3. (Binary Classification) You can download a pair of training and test data sets with a binary y and 30-dimensional predictors.

```
setwd("your_path_to_the_download_data")
train <- read.csv("train.csv")
test  <- read.csv("test.csv")
```

Your job is to compare the classification performance of various binary classification methods including:

- Logistic Regression
- LASSO-penalized Logistic Regression
- Linear SVM
- Gaussian Kernel SVM
- Classification Tree
- Random Forest
- Logit Boosting

Please report test accuracy of all methods you trained to choose the best model. You must submit your ready-to-run code that reproduces your work.

4. (Clustering after Dimension Reduction) You can download a set of data with $n = 500$ and $p = 100$.

```
setwd("your_path_to_the_download_data")
train <- read.csv("data_usv.csv")
```

Your job is to conduct clustering analysis after the dimension reduction.

(a) Apply principal component analysis (PCA) and t-SNE to reduce the data dimension from $p = 10$ and 2.

(b) Apply various clustering method to the data on the reduced space you obtained from PCA and t-SNE in (a), respectively. Popular clustering methods we covered in the class include

- k-means clustering
- Hierarchical clustering
- Gaussian mixture model
- dbSCAN

Please visualize your result (eq. scatter plots on the reduced space with assigned cluster with different colors) to compare the eight approaches $\{\text{PCA}, \text{t-SNE}\} \times \{k - \text{Means}, \text{HClust}, \text{GMix}, \text{dbSCAN}\}$. For your convenience the number of cluster is given as $k = 3$. You must submit your ready-to-run code that reproduces your work.