تمرین سری چهارم – تحلیل آماری کاربردی

نکات مهم:

- برای تحویل تمارین گزارشی با استفاده از RMarkdown ایجاد نموده که در آن به صورت کامل صورت سوال، کد مربوطه، خروجی کد و توضیحات مربوطه آورده شود و سپس در سامانه آموزش مجازی بارگذاری شود. برای بارگذاری، ابتدا فایل word خروجی RMarkdown را تهیه کرده و سپس آن را pdf نموده و فایل pdf را به همراه فایل Rmd در سامانه بارگذاری کنید. صورت سوال را می توانید به صورت یک عکس در گزارش بیاورید. نام فایل ها حتما شامل شماره تمرین و شماره دانشجویی باشد (به عنوان مثال، HW2- 9913418023).
- دقت کنید که حتما برای پاسخ به تمامی قسمت های سوالات کدی نوشته و پاسخ را به صورت خروجی کد بیاورید.

برای این سری تمرین ممکن است به کتابخانه های زیر نیاز داشته باشید:

```
# general
library(MASS)
library(caret)
library(tidyverse)
library(knitr)
library(kableExtra)

# specific
library(e1071)
library(nnet)
library(pROC)
library(ellipse)
```

تمرین سری چهارم – تحلیل آماری کاربردی

مساله ۱:

برای انجام این مساله مجموعه داده های برآورد و اعتبارسنجی از دو متغیر پیش بینی کننده X_1 و X_2 و یک متغیر پاسخ کیفی Y دارای ۴ سطح A، B، C و D به صورت زیر تولید می شوند:

```
# setup parameters
num_obs = 1000

# means
mu_1 = c(12, 8.5)
mu_2 = c(22, 10)
mu_3 = c(12, 15)
mu_4 = c(12, 20)

# sigmas
sigma_1 = matrix(c(10, -4, -4, 8), 2, 2)
sigma_2 = matrix(c(5, -3, -3, 5), 2, 2)
sigma_3 = matrix(c(8, 3, 3, 8), 2, 2)
sigma_4 = matrix(c(8, 6, 6, 8), 2, 2)

# control randomization
set.seed(42)

# make estimation data
hw04_est = data.frame(
  # create response
  as.factor(c(rep("A", num_obs / 2), rep("B", num_obs),
              rep("C", num_obs * 2), rep("D", num_obs))),
  # create predictors
  rbind(
    mvrnorm(n = num_obs / 2, mu = mu_1, Sigma = sigma_1),
    mvrnorm(n = num_obs, mu = mu_2, Sigma = sigma_2),
    mvrnorm(n = num_obs * 2, mu = mu_3, Sigma = sigma_3),
    mvrnorm(n = num_obs, mu = mu_4, Sigma = sigma_4)
  )
)

# label variables
colnames(hw04_est) = c("y", "x1", "x2")

# make validation data
hw04_val = data.frame(
  # create response
  as.factor(c(rep("A", num_obs), rep("B", num_obs),
              rep("C", num_obs), rep("D", num_obs))),
  # create predictors
  rbind(
    mvrnorm(n = num_obs, mu = mu_1, Sigma = sigma_1),
    mvrnorm(n = num_obs, mu = mu_2, Sigma = sigma_2),
    mvrnorm(n = num_obs, mu = mu_3, Sigma = sigma_3),
    mvrnorm(n = num_obs, mu = mu_4, Sigma = sigma_4)
  )
)

# label variables
colnames(hw04_val) = c("y", "x1", "x2")
```

تمرین سری چهارم – تحلیل آماری کاربردی

ابتدا با استفاده از ابزار EDA و بر روی مجموعه داده برآورد، سعی کنید با استفاده از چند نمودار رفتار متغیرهای X_1 و X_2 را به طور جداگانه و همچنین نسبت به هم در کلاسهای مختلف متغیر پاسخ بررسی کنید. سپس مدلهای زیر را بر روی مجموعه داده برآورد برازش دهید:

- مدل رگرسیون لجستیک چند کلاسه با استفاده از هر ۲ متغیر (از تابع multinom در کتابخانه nnet استفاده کنید.)
- مدل رگرسیون لجستیک چند کلاسه با استفاده از هر ۲ متغیر و اثر متقابل و همچنین عبارات درجه دوم آنها (از تابع multinom در کتابخانه nnet استفاده کنید.)
- مدل LDA با استفاده از هر ۲ متغیر و تابع توزیع پیشین برآورد شده از داده ها
- مدل LDA با استفاده از هر ۲ متغیر و اثر متقابل و عبارات درجه دوم آنها و تابع توزیع پیشین برآورد شده از داده ها
- مدل LDA با استفاده از هر ۲ متغیر و تابع توزیع پیشین یکنواخت برای تمامی کلاسها (Flat Prior)
- مدل QDA با استفاده از هر ۲ متغیر و تابع توزیع پیشین برآورد شده از داده ها
- مدل QDA با استفاده از هر ۲ متغیر و تابع توزیع پیشین یکنواخت برای تمامی کلاسها (Flat Prior)
- مدل بیز ساده (Naïve Bayes) با استفاده از هر ۲ متغیر و تابع توزیع پیشین برآورد شده از داده ها
- مدل KNN با استفاده از هر ۲ متغیر (بدون نرمالسازی)
- مدل درخت طبقه بندی با استفاده از هر ۲ متغیر

نرخ خطای طبقه بندی نادرست (Misclassification) را برای تمامی مدلها و بر روی هر دو مجموعه برآورد و اعتبارسنجی محاسبه کنید و آنها را در قالب یک جدول مناسب گزارش کنید. کدام مدل را به عنوان مدل بهتر انتخاب می کنید؟ سعی کنید چند مقدار را برای پارامترهای تنظیمی KNN و درخت طبقه بندی بررسی کنید و بهترین مقادیر را در جدول مربوطه گزارش کنید. همچنین به سوالات مفهومی زیر نیز پاسخ دهید:

- تحلیل کنید که برای مدلهای LDA و QDA در نظر گرفتن تابع توزیع پیشین به صورت یکنواخت کارآیی بهتری دارد یا برآورد آن از روی داده ها؟
- کارآیی روش های LDA ، QDA و بیز ساده را با ذکر دلیل با هم مقایسه کنید.
- از بین ۴ کلاس متغیر پاسخ و با توجه به بهترین مدل انتخاب شده، کدام کلاس راحت تر طبقه بندی می شود؟

تمرین سری چهارم – تحلیل آماری کاربردی

مساله ۲ (تشخیص سرطان با رگرسیون لجستیک):

برای این مساله، مجموعه داده های برآورد (wisc-est.csv) و اعتبارسنجی (wisc-val.csv) در اختیار شما قرار داده خواهد شد. این مجموعه داده، اصلاحی از مجموعه داده سرطان سینه ویسکانسین است که از پایگاه داده UCI گرفته شده است. در واقع، فقط ۱۰ ویژگی اول مجموعه داده اصلی در این اصلاح مورد استفاده قرار گرفته است. برای کسب اطلاعات بیشتر در رابطه با این مجموعه داده می توانید از لینک های زیر استفاده کنید:

- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
- https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names

حتما اطمینان حاصل کنید که متغیر پاسخ مساله به صورت factor در نرم افزار R وارد شود.

- یک مدل رگرسیون لجستیک با همه متغیرها بر روی مجموعه داده برآورد برازش دهید. برای پیش بینی از نقاط برش (cutoff) متفاوتی (۰.۱، ۰.۳، ۰.۵، ۰.۷ و ۰.۹) استفاده کنید و معیارهای Accuracy، Sensitivity و Specificity را بر روی مجموعه داده اعتبارسنجی در قالب یک جدول مناسب گزارش کنید. کلاس M (malignant) را به عنوان کلاس مثبت در نظر بگیرید. همچنین در قالب یک نمودار این معیارها را بر حسب مقادیر مختلف نقاط برش رسم کنید. برای وضوح بیشتر، نقاط برش بیشتری (cutoff = seq(0.01, 0.99, by = 0.01)) را در نظر بگیرید و هر یک از معیارها را در قالب یک خط جداگانه با نوع و رنگ متفاوت رسم کنید. حتما برای نمودار عنوان کلی، عنوان برای محورها و همچنین راهنما (legend) داشته باشید. از بین نقاط برش (۰.۱، ۰.۳، ۰.۵، ۰.۷ و ۰.۹)، کدام نقطه برش را با توجه به صورت مساله ترجیح می دهید؟ کدام معیار در این مساله ارجحیت بیشتری دارد؟
- منحنی ROC را نیز با استفاده از مجموعه داده اعتبارسنجی رسم کنید. حتما توجه داشته باشید که مقدار AUC نیز بر روی نمودار رسم شده باشد.
- یک مدل رگرسیون لجستیک فقط با ۲ متغیر radius و symmetry بر روی مجموعه داده برآورد برازش دهید. سپس نموداری بر روی مجموعه داده اعتبارسنجی رسم کنید که این ۲ متغیر بر روی محورهای آن قرار دارند و رنگ نقاط با توجه به کلاس آنها تعیین شده است. در نهایت نیز با توجه به مدل برازش داده شده و در نظر گرفتن نقطه برش ۰.۵، مرز تصمیم را بر روی نمودار اضافه کنید. این کار را با اضافه کردن اثر متقابل ۲ متغیر به مدل تکرار کنید.

موفق باشید- شادمان