weekly assignment 03

Xiang Li

2024/2/26

```
library(MASS)
```

1.

```
gen_data = function(n) {
    p = 15
    n1 = n2 = n/2
    cov_1 = diag(rep(1, p)) + 0.2
    x_class1 = mvrnorm(n1, mu = rep(3, p), Sigma = cov_1)
    x_class2 = mvrnorm(n2, mu = rep(2, p), Sigma = cov_1)
    x = rbind(x_class1, x_class2)
    y = rep(c(1, 2), c(n1, n2))
    df = as.data.frame(cbind(x, y))
    names(df) = c(paste0("x", 1:p), "y")
    return(df)
}
```

Either with small or big training set, LDA is expected to perform better than logistic regression, because the ture model is the model assumed by LDA.

And with big training set, both of models will perform better than small training set, because the variance of models is smaller.

2.

```
set.seed(519)
test_set = gen_data(10000)
cal_acc = function(n_train, test) {
    n_reps = 100
    acc_mat = matrix(data = 0, nrow = n_reps, ncol = 2)
    colnames(acc_mat) = c("logistic regression", "LDA")
    for (i in 1:n_reps) {
        train = gen_data(n = n_train)
        logis_model = glm(as.factor(y) ~ ., train, family = binomial)
        logis_prob_pre = predict(logis_model, newdata = test, type = "response")
        logis_y_pre = rep(1, nrow(test))
```

```
        logis_y_pre[logis_prob_pre > 0.5] = 2
        acc_mat[i, 1] = mean(logis_y_pre == test$y)
        lda_model = lda(y ~ ., train)
        lda_y_pre = predict(lda_model, newdata = test, type = "response")$class
        acc_mat[i, 2] = mean(lda_y_pre == test$y)
    }
    return(acc_mat)
}
acc_mat_50 = cal_acc(50, test_set)
acc_mat_10000 = cal_acc(10000, test_set)
acc = matrix(c(apply(acc_mat_50, 2, mean), apply(acc_mat_10000, 2, mean)), nrow = 2,
    byrow = TRUE)
colnames(acc) = c("logistic regression", "LDA")
rownames(acc) = c("n = 50", "n = 10000")
acc
```

```
##           logistic regression      LDA
## n = 50               0.730330 0.759068
## n = 10000            0.830698 0.830821
```

The obtained numbers in line with my expectations.