Genie Cancer Data

Smiti Kaul

Feb 2 - present, 2018

Load the Data

```
cancer <- read.csv("C:/Users/kauls15/Desktop/github/genie/data/derived/data_clinical_patient.txt",
    stringsAsFactors = FALSE, sep = "\t")
cancer <- cancer[-c(1, 2, 3), ]
colnames(cancer) = cancer[1, ]  # the first row will be the header
cancer <- cancer[-1, -5]
head(cancer)

sample <- read.csv("C:/Users/kauls15/Desktop/github/genie/data/derived/data_clinical_sample.txt",
    stringsAsFactors = FALSE, sep = "\t")
sample$AGE_AT_SEQ_REPORT[sample$AGE_AT_SEQ_REPORT == "<18"] <- 17
sample$AGE_AT_SEQ_REPORT[sample$AGE_AT_SEQ_REPORT == ">89"] <- 90
head(sample)
```

Map (most) Data to Discrete Variables and Merge Datasets

```
c <- cancer
# sort(unique(df$SEX))
c$SEX <- recode(c$SEX, Female = 0, Male = 1, Unknown = 2)
c$PRIMARY_RACE <- recode(c$PRIMARY_RACE, Asian = 0, Black = 1, `Native American` = 2,
    Other = 3, Undefined = 4, Unknown = 5, White = 6)
c$ETHNICITY <- recode(c$ETHNICITY, `Non-Spanish/non-Hispanic` = 0, `Spanish/Hispanic` = 1,
    Unknown = 2)
head(c)

s <- sample
s <- s[, c(-3, -6)]
s$SAMPLE_TYPE <- recode(s$SAMPLE_TYPE, Metastasis = 0, Other = 1, Primary = 2,
    Unspecified = 3)
head(s)

temp <- merge(c, s, by = "PATIENT_ID")
temp[, 8] = temp[6]
temp <- temp[, -6]
names(temp) <- c("patient_id", "sex", "primary_race", "ethnicity", "age", "cancer_type",
    "metastasis")
write.table(temp, "C:/Users/kauls15/Desktop/github/genie/data/derived/data_clinical_patient_and_sample.
    sep = "\t")
head(temp)
```

Begin Analysis

Only Clinical Data

```
d <- temp
```

Exploratory Data Analysis

Multivariate Linear Regression

```
m1 = lm(metastasis ~ sex + primary_race + ethnicity + age + cancer_type, data = d)
summary(m1)
mse <- function(sm) mean(sm$residuals^2)
mse(m1)
```

Multivariate Logistic Regression

```
set.seed(125)
cv_errors <- rep(0, 10)

df <- d[sample(nrow(d)), ]
folds <- cut(seq(1, nrow(df)), breaks = 10, labels = FALSE)

for (i in 1:10) {
    # segment data
    test_indexes <- which(folds == i, arr.ind = TRUE)
    test <- df[test_indexes, ]
    train <- df[-test_indexes, ]

    # method 1, 2, 3
    m1 = glm(metastasis ~ sex + primary_race + ethnicity + age + cancer_type,
        data = train, family = binomial)

    fold_errors <- rep(0, nrow(test))

    test <- test[-which(test$cancer_type == "Adenocarcinoma In Situ"), ]
    for (j in 1:nrow(test)) {
        # fold_errors[j] <- (predict(m1, test[j,])) - test$metastasis[j]
        predict.glm(m1, data.frame(metastasis = test$metastasis[5]), type = "resp")
        fold_errors[j] <- predict.glm(m1, data.frame(metastasis = test$metastasis[5]),
            type = "resp")

    }

    cv_errors[i] <- mean(fold_errors^2)
}

mse_m1 <- mean(cv_errors)
cat("test mse: ", mse_m1)
```