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 = 2, Primary = 1,
    Unspecified = 3)
s$CANCER_TYPE <- recode(s$CANCER_TYPE, `Adenocarcinoma In Situ` = 0, `Adrenocortical Carcinoma` = 1,
    `Ampullary Carcinoma` = 2, `Anal Cancer` = 3, `Appendiceal Cancer` = 4,
    `Bladder Cancer` = 5, `Bladder/Urinary Tract Cancer, NOS` = 6, `Blastic Plasmacytoid Dendritic Cell
    `Blood Cancer, NOS` = 8, `Bone Cancer` = 9, `Bone Cancer, NOS` = 10, `Bowel Cancer, NOS` = 11,
    `Breast Cancer` = 12, `Breast Sarcoma` = 13, `Cancer of Unknown Primary` = 14,
    `Cervical Cancer` = 15, `Choroid Plexus Tumor` = 16, `CNS Cancer` = 17,
    `CNS/Brain Cancer, NOS` = 18, `Colorectal Cancer` = 19, `Embryonal Tumor` = 20,
    `Endometrial Cancer` = 21, `Esophageal/Stomach Cancer, NOS` = 22, `Esophagogastric Cancer` = 23,
    `Gastrointestinal Neuroendocrine Tumor` = 24, `Gastrointestinal Stromal Tumor` = 25,
    `Germ Cell Tumor` = 26, `Gestational Trophoblastic Disease` = 27, Glioma = 28,
    `Head and Neck Cancer` = 29, `Head and Neck Cancer, NOS` = 30, `Hepatobiliary Cancer` = 31,
    `Histiocytic Disorder` = 32, Histiocytosis = 33, `Hodgkin Lymphoma` = 34,
    Leukemia = 35, `Lung Cancer, NOS` = 36, Mastocytosis = 37, Melanoma = 38,
    Mesothelioma = 39, `Miscellaneous Brain Tumor` = 40, `Miscellaneous Neuroepithelial Tumor` = 41,
    `Multiple Myeloma` = 42, Myelodysplasia = 43, `Myeloproliferative Neoplasm` = 44,
    `Nerve Sheath Tumor` = 45, `Non-Hodgkin Lymphoma` = 46, `Non-Small Cell Lung Cancer` = 47,
```

```
    `Other Cancer, NOS` = 48, `Ovarian Cancer` = 49, `Ovarian/Fallopian Tube Cancer, NOS` = 50,
    `Pancreatic Cancer` = 51, `Pancreatic Cancer, NOS` = 52, `Penile Cancer` = 53,
    Pheochromocytoma = 54, `Pineal Tumor` = 55, `Prostate Cancer` = 56, `Renal Cell Carcinoma` = 57,
    Retinoblastoma = 58, `Salivary Gland Cancer` = 59, `Sellar Tumor` = 60,
    `Sex Cord Stromal Tumor` = 61, `Skin Cancer, Non-Melanoma` = 62, `Skin Cancer, NOS` = 63,
    `Small Bowel Cancer` = 64, `Small Cell Lung Cancer` = 65, `Soft Tissue Sarcoma` = 66,
    `Testicular Cancer, NOS` = 67, `Thymic Tumor` = 68, `Thyroid Cancer` = 69,
    `Thyroid Cancer, NOS` = 70, `Uterine Cancer, NOS` = 71, `Uterine Sarcoma` = 72,
    `Vaginal Cancer` = 73, `Vulvar/Vaginal Cancer, NOS` = 74, `Wilms Tumor` = 75)
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)
```

Only Clinical Data

Create dataframe

```
d <- temp
head(d)
t <- subset(d, metastasis != 2)
t <- subset(t, metastasis != 3)
head(t)
d <- t
```

Exploratory Data Analysis

Multivariate Linear Regression

```
m1 = glm(metastasis ~ sex + primary_race + ethnicity + age + cancer_type, data = d)
summary(m1)
mse <- function(sm) mean(sm$residuals^2)
mse(m1)  # 0.008522365
```

Multivariate Logistic Regression

```
# library(caret)
set.seed(125)
trainIndex <- createDataPartition(d$metastasis, p = 0.75, list = FALSE, times = 1)

train <- d[trainIndex, ]
test <- d[-trainIndex, ]

# m1 = glm.fit(metastasis ~ sex + primary_race + ethnicity + age +
# cancer_type, data = train, family = binomial)
m1 = glm(train$metastasis ~ train$sex + train$primary_race + train$ethnicity +
```

```
    train$age + train$cancer_type, 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[j]),
        type = "resp") - test$metastasis[j]
}

error <- mean(fold_errors^2)
mse_m1 <- mean(error)
cat("test mse: ", mse_m1)  # 0.02379561
```

Multivariate Logistic Regression with 10-fold Cross Validation

```
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.fit(metastasis ~ sex + primary_race + ethnicity +
    # age + cancer_type, data = train, family = binomial)
    m1 = glm(train$metastasis ~ train$sex + train$primary_race + train$ethnicity +
        train$age + train$cancer_type, 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]
        fold_errors[j] <- predict.glm(m1, data.frame(metastasis = test$metastasis[j]),
            type = "resp") - test$metastasis[j]

    }

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

mse_m1 <- mean(cv_errors)
cat("test mse: ", mse_m1)  # 0.02253641
```