naivebayes_multi

Xingyu Wu

12/5/2021

```
data = read.csv("processed_counts.csv")
```

```
label = read.csv("annotation.csv")
```

```
library(sampling)
set.seed(6690)

train_id <- sample(label$ID, round(dim(label)[1]*0.75))

train_data <- data[data$ID %in% train_id, ]
test_data <- data[!(data$ID %in% train_id), ]

train_label <- label[data$ID %in% train_id, ]
test_label <- label[!(data$ID %in% train_id), ]
```

```
total_train = merge(train_data, train_label, by = "ID")
total_test = merge(test_data, test_label, by = "ID")
total_train = total_train[, -1]
total_test = total_test[, -1]
```

```
library(ggplot2)
library(lattice)
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:sampling':
##
##     cluster
```

```
total_train$Type = factor(total_train$Type)
total_test$Type = factor(total_test$Type)
control <- trainControl(method = 'repeatedcv', number = 10, repeats = 2)
model <- train(Type~., total_train,
               method = 'naive_bayes',
               preProcess = c('center', 'scale'),
               trControl = control)
model
```

```
## Naive Bayes
##
## 5730 samples
## 2916 predictors
##   16 classes: 'BLCA', 'BRCA', 'CESC', 'COAD', 'GBM', 'HNSC', 'LIHC', 'LUAD', 'LUSC', 'Normal', 'PRAD
```

```
##
## Pre-processing: centered (2916), scaled (2916)
## Resampling: Cross-Validated (10 fold, repeated 2 times)
## Summary of sample sizes: 5156, 5157, 5154, 5158, 5157, 5155, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa
##   FALSE      0.8561969  0.8450464
##    TRUE      0.8613379  0.8505019
##
## Tuning parameter 'laplace' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were laplace = 0, usekernel = TRUE
##  and adjust = 1.
```

```
truth <- total_test$Type
pred <- predict(model, newdata = total_test)
cm <- confusionMatrix(table(pred, truth))
cm
```

```
## Confusion Matrix and Statistics
##
##         truth
## pred     BLCA BRCA CESC COAD GBM HNSC LIHC LUAD LUSC Normal PRAD READ SKCM STAD
##   BLCA     74    1    0    0   0    0    0    2    0      1    0    1    0    0
##   BRCA      0  254    0    0   0    0    0    0    0      2    4    0    0    0
##   CESC     13    1   63    0   0   11    0    1    7      0    0    0    0    0
##   COAD      0    0    0   55   0    0    0    0    0      0    0    8    0    6
##   GBM       0    0    0    0  43    0    0    0    0      1    0    0    0    0
##   HNSC      5    1    4    0   0  108    0    0    8      4    0    0    1    1
##   LIHC      0    0    0    0   0    0   80    0    0     22    0    0    0    0
##   LUAD      0    0    0    0   0    0    1  124   10      0    0    0    0    0
##   LUSC      1    4    6    0   0   18    1    7  114      0    0    0    0    1
##   Normal    2    8    0    0   0    1    2    7    1    114    2    0    0    4
##   PRAD      0    0    0    0   0    0    0    0    0      4  121    0    0    0
##   READ      0    0    0   46   0    0    0    0    0      0    0   34    0    0
##   SKCM      0    0    0    0   0    0    0    0    0      0    0    0  108    0
##   STAD      2    2    0    6   0    0    1    0    1      2    0    1    0   84
##   THCA      0    0    0    0   0    0    0    0    0      7    0    0    0    0
##   UCEC      1    2    4    0   0    0    0    0    0      1    0    0    1    0
##         truth
## pred     THCA UCEC
##   BLCA      0    0
##   BRCA      0    0
##   CESC      0    1
##   COAD      0    0
##   GBM       0    0
##   HNSC      0    0
##   LIHC      0    0
##   LUAD      2    0
##   LUSC      0    0
##   Normal    1    2
##   PRAD      0    0
```

```
##   READ      0    0
##   SKCM      0    1
##   STAD      0    0
##   THCA    122    0
##   UCEC      0  143
##
## Overall Statistics
##
##                Accuracy : 0.8592
##                  95% CI : (0.8427, 0.8745)
##     No Information Rate : 0.1429
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8481
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: BLCA Class: BRCA Class: CESC Class: COAD Class: GBM
## Sensitivity              0.75510      0.9304     0.81818     0.51402    1.00000
## Specificity              0.99724      0.9963     0.98145     0.99224    0.99946
## Pos Pred Value           0.93671      0.9769     0.64948     0.79710    0.97727
## Neg Pred Value           0.98689      0.9885     0.99228     0.97175    1.00000
## Prevalence               0.05131      0.1429     0.04031     0.05602    0.02251
## Detection Rate           0.03874      0.1330     0.03298     0.02880    0.02251
## Detection Prevalence     0.04136      0.1361     0.05079     0.03613    0.02304
## Balanced Accuracy        0.87617      0.9634     0.89982     0.75313    0.99973
##                      Class: HNSC Class: LIHC Class: LUAD Class: LUSC
## Sensitivity              0.78261     0.94118     0.87943     0.80851
## Specificity              0.98646     0.98795     0.99265     0.97852
## Pos Pred Value           0.81818     0.78431     0.90511     0.75000
## Neg Pred Value           0.98313     0.99723     0.99041     0.98464
## Prevalence               0.07225     0.04450     0.07382     0.07382
## Detection Rate           0.05654     0.04188     0.06492     0.05969
## Detection Prevalence     0.06911     0.05340     0.07173     0.07958
## Balanced Accuracy        0.88453     0.96456     0.93604     0.89351
##                      Class: Normal Class: PRAD Class: READ Class: SKCM
## Sensitivity                0.72152     0.95276     0.77273     0.98182
## Specificity                0.98288     0.99776     0.97535     0.99944
## Pos Pred Value             0.79167     0.96800     0.42500     0.99083
## Neg Pred Value             0.97508     0.99664     0.99454     0.99889
## Prevalence                 0.08272     0.06649     0.02304     0.05759
## Detection Rate             0.05969     0.06335     0.01780     0.05654
## Detection Prevalence       0.07539     0.06545     0.04188     0.05707
## Balanced Accuracy          0.85220     0.97526     0.87404     0.99063
##                      Class: STAD Class: THCA Class: UCEC
## Sensitivity              0.87500     0.97600     0.97279
## Specificity              0.99173     0.99608     0.99490
## Pos Pred Value           0.84848     0.94574     0.94079
## Neg Pred Value           0.99337     0.99832     0.99772
## Prevalence               0.05026     0.06545     0.07696
## Detection Rate           0.04398     0.06387     0.07487
## Detection Prevalence     0.05183     0.06754     0.07958
```

```
## Balanced Accuracy        0.93337     0.98604     0.98384
```

```
importance <- varImp(model, scale = FALSE)
importance
```

```
## ROC curve variable importance
##
##   variables are sorted by maximum importance across the classes
##   only 20 most important variables shown (out of 2916)
##
##             BLCA   BRCA   CESC   COAD    GBM   HNSC   LIHC   LUAD   LUSC Normal
## GPM6A     0.6076 0.5790 1.0000 0.6388 0.5735 0.7620 0.6115 0.9292 0.9049 0.6314
## TARP      0.7377 0.7263 0.7263 0.7263 0.7263 0.7863 0.7263 0.7859 1.0000 0.7263
## KCNJ16    0.6762 0.6762 0.9896 0.6762 0.6762 0.7815 0.8398 0.8981 0.6825 0.6762
## ADCYAP1R1 0.8547 0.8547 1.0000 0.8547 0.8547 0.8547 0.8547 0.8985 0.8547 0.8547
## LOC145837 0.9244 0.9244 0.9244 0.9244 0.9438 0.9244 0.9244 0.9244 0.9999 0.9331
## KLK2      0.6299 0.6299 0.6647 0.6299 0.6299 0.6299 0.6299 0.6299 0.9999 0.6299
## HEPACAM   0.7287 0.7287 0.9999 0.7287 0.7500 0.7287 0.7287 0.8175 0.7287 0.7287
## KLHL14    0.7489 0.7489 0.7489 0.7489 0.7828 0.7489 0.7489 0.8323 0.8622 0.7489
## PMP2      0.6664 0.6664 0.9999 0.6664 0.7564 0.6664 0.6664 0.7502 0.8411 0.6664
## NWD1      0.7945 0.7198 0.9488 0.6947 0.7903 0.7649 0.7374 0.7987 0.9999 0.7357
## LINC00461 0.6208 0.5614 0.9998 0.5680 0.6683 0.6030 0.5913 0.5381 0.6213 0.6149
## AQP4      0.8435 0.8435 0.9998 0.8435 0.8435 0.9672 0.9455 0.9274 0.9116 0.8435
## PLP1      0.7201 0.7201 0.9998 0.7201 0.8305 0.7201 0.7201 0.8613 0.8723 0.7201
## LAD1      0.8952 0.8952 0.9998 0.8952 0.8952 0.8952 0.8952 0.8952 0.9274 0.8952
## SPDEF     0.9412 0.9412 0.9412 0.9412 0.9412 0.9412 0.9412 0.9412 0.9997 0.9412
## CMTM5     0.5853 0.6501 0.9997 0.5165 0.7212 0.5095 0.5395 0.8775 0.8513 0.5917
## CHRNA2    0.6354 0.6354 0.9070 0.6354 0.6354 0.6709 0.6897 0.7909 0.9996 0.6354
## ERGIC1    0.8975 0.9112 0.8975 0.8975 0.8975 0.8975 0.8975 0.8975 0.9996 0.9068
## PEBP4     0.7924 0.7924 0.9835 0.7924 0.7924 0.9556 0.8638 0.9298 0.9948 0.7924
## CHRM1     0.9523 0.9523 0.9715 0.9523 0.9523 0.9523 0.9523 0.9523 0.9996 0.9528
##             PRAD   READ   SKCM   STAD   THCA   UCEC
## GPM6A     0.5735 0.8366 0.8841 0.5856 0.6684 0.6076
## TARP      0.7263 0.7580 0.7263 0.7263 0.7263 0.7377
## KCNJ16    0.6762 0.8045 1.0000 0.8096 0.6762 0.6019
## ADCYAP1R1 0.8547 0.8547 0.8547 0.8890 0.8547 0.6160
## LOC145837 0.9244 0.9244 0.9244 0.9244 0.9244 0.8683
## KLK2      0.6299 0.6299 0.7711 0.6299 0.6903 0.5486
## HEPACAM   0.7287 0.8259 0.7287 0.7287 0.7287 0.5190
## KLHL14    0.7489 0.7489 0.9999 0.9921 0.7489 0.5276
## PMP2      0.7711 0.6664 0.7168 0.6664 0.6664 0.6056
## NWD1      0.7332 0.7082 0.6947 0.8799 0.6947 0.7945
## LINC00461 0.8859 0.5381 0.6820 0.9047 0.5706 0.6208
## AQP4      0.8435 0.8435 0.9719 0.8435 0.8435 0.5712
## PLP1      0.9901 0.7201 0.7201 0.7201 0.8312 0.6325
## LAD1      0.9516 0.8952 0.9344 0.8952 0.9448 0.5905
## SPDEF     0.9412 0.9412 0.9412 0.9412 0.9412 0.7831
## CMTM5     0.7984 0.5428 0.5449 0.5940 0.5620 0.6501
## CHRNA2    0.6354 0.7798 0.6901 0.6354 0.6354 0.5713
## ERGIC1    0.8975 0.8975 0.9397 0.9062 0.8975 0.9112
## PEBP4     0.7924 0.7924 0.9996 0.7924 0.8567 0.6755
## CHRM1     0.9523 0.9523 0.9523 0.9523 0.9523 0.9386
```

```
index <- importance$importance
for(i in 1:16){
```

```
  index1 <- head(index[order(index[,i],decreasing = TRUE),],n=20)
  print(colnames(index1[i]))
  print(rownames(index1))
}
```

```
## [1] "BLCA"
##  [1] "TCF21"     "C8orf85"   "ANKS1B"    "ST8SIA6"   "FOXF1"     "LOC400550"
##  [7] "SCGB1D2"   "SLC1A2"    "LRIG1"     "CHRM1"     "HPN"       "ANKRD30B"
## [13] "AFF3"      "C10orf99"  "UPK3B"     "COBL"      "RERG"      "CREB3L4"
## [19] "LRRN2"     "SPDEF"
## [1] "BRCA"
##  [1] "GPA33"    "CDX1"     "MYO1A"    "TCF21"    "PIP5K1B"  "EPCAM"
##  [7] "C9orf152" "CFTR"     "TDGF1"    "GUCY2C"   "MEP1A"    "TFF3"
## [13] "SLC6A7"   "NKX2.3"   "BTNL8"    "PCK1"     "IHH"      "C8orf85"
## [19] "FRMD1"    "ANKS1B"
## [1] "CESC"
##  [1] "GPM6A"     "ADCYAP1R1" "HEPACAM"   "PMP2"      "LINC00461" "AQP4"
##  [7] "LAD1"      "PLP1"      "CMTM5"     "CORO2B"    "JAM2"      "DBX2"
## [13] "GPR89C"    "TMEM59L"   "STMN4"     "ASTN1"     "BAIAP2L1"  "PPIAL4G"
## [19] "GRIA4"     "DPP6"
## [1] "COAD"
##  [1] "TCF21"     "C8orf85"   "ANKS1B"    "ST8SIA6"   "FOXF1"     "LOC400550"
##  [7] "SCGB1D2"   "SLC1A2"    "LRIG1"     "CHRM1"     "HPN"       "PPARG"
## [13] "ANKRD30B"  "AFF3"      "C10orf99"  "UPK3B"     "COBL"      "RERG"
## [19] "CREB3L4"   "LRRN2"
## [1] "GBM"
##  [1] "C8B"      "F11"      "APOB"     "CPB2"     "ACSM2A"   "APOH"
##  [7] "F10"      "TAT"      "HPN"      "CREB3L3"  "RBP4"     "SLC38A3"
## [13] "C6"       "DAO"      "SERPINA5" "ADH1A"    "SLC22A9"  "PKLR"
## [19] "STYK1"    "OTC"
## [1] "HNSC"
##  [1] "TCF21"     "SLC22A31"  "ARHGEF38"  "HPN"       "C8orf85"   "ANKS1B"
##  [7] "SFTA1P"    "ST8SIA6"   "PON3"      "FOXF1"     "C16orf89"  "LOC400550"
## [13] "GGTLC1"    "AQP4"      "ATP11A"    "SCGB1D2"   "KCNQ3"     "LOC643441"
## [19] "FMO2"      "SLC1A2"
## [1] "LIHC"
##  [1] "TCF21"     "C8orf85"   "ANKS1B"    "ST8SIA6"   "FOXF1"     "LOC400550"
##  [7] "SCGB1D2"   "SLC1A2"    "LRIG1"     "FMO2"      "CHRM1"     "HPN"
## [13] "ANKRD30B"  "AFF3"      "C10orf99"  "UPK3B"     "COBL"      "RERG"
## [19] "AQP4"      "RASSF9"
## [1] "LUAD"
##  [1] "TCF21"    "HLF"      "TMEM220"  "NUP85"    "C20orf20" "RCC2"
##  [7] "UBE2C"    "ORC6"     "TROAP"    "C1orf135" "CDC6"     "KIF2C"
## [13] "C16orf59" "RNASEH2A" "NUF2"     "HJURP"    "RCC1"     "RAD54L"
## [19] "C8orf85"  "HMGB3"
## [1] "LUSC"
##  [1] "TARP"      "LOC145837" "KLK2"      "NWD1"      "SPDEF"     "CHRNA2"
##  [7] "ERGIC1"    "CHRM1"     "CREB3L4"   "HOXB13"    "ARHGEF38"  "HPN"
## [13] "FEV"       "TRIM36"    "DEFB132"   "TMEFF2"    "CGNL1"     "BEND4"
## [19] "ABCC4"     "C9orf152"
## [1] "Normal"
##  [1] "GPA33"        "CDX1"         "MYO1A"        "TDGF1"        "EPCAM"
##  [6] "CFTR"         "FLJ32063"     "C9orf152"     "GUCY2C"       "MEP1A"
## [11] "PIP5K1B"      "TCF21"        "TFF3"         "IHH"          "ACY3"
```

```
## [16] "FRMD1"        "NKX2.3"       "BTNL8"        "LOC100505933" "ALPI"
## [1] "PRAD"
##  [1] "EPCAM"     "STYK1"     "TCF21"     "SPINT2"    "PLP1"      "ETV5"
##  [7] "SOX10"     "GRHL2"     "VIM"       "ST14"      "AP1M2"     "EDNRB"
## [13] "LOC257358" "PHACTR1"   "PLOD3"     "FOXD3"     "ROPN1"     "MAPK13"
## [19] "C1orf172"  "MAL2"
## [1] "READ"
##  [1] "TCF21"     "METTL24"   "LOC440563" "DPCR1"     "CLDN18"    "POTEE"
##  [7] "MYO1A"     "C8orf85"   "ANKS1B"    "ST8SIA6"   "FOXF1"     "LOC400550"
## [13] "PPIAL4G"   "SCGB1D2"   "PIP5K1B"   "ACY3"      "SLC1A2"    "SLC18A2"
## [19] "HSP90AA6P" "HOMER3"
## [1] "SKCM"
##  [1] "KCNJ16"   "KLHL14"   "PEBP4"    "C16orf89" "ACOX2"    "PADI3"
##  [7] "FAM189A2" "SRL"      "TRIP13"   "MGAT4C"   "PPAP2B"   "LRRC2"
## [13] "HOXA10"   "HN1"      "FGFRL1"   "PLK1"     "FAM83D"   "SHE"
## [19] "LONRF2"   "UBE2C"
## [1] "STAD"
##  [1] "KLHL14"    "TCF21"     "SCGB1D2"   "AQP5"      "SOX17"     "LRRTM1"
##  [7] "PRAME"     "C8orf85"   "ANKS1B"    "ST8SIA6"   "FOXF1"     "LOC400550"
## [13] "LOC643650" "STXBP6"    "SLC1A2"    "LRIG1"     "SMPDL3B"   "DACT2"
## [19] "SERTM1"    "CXorf57"
## [1] "THCA"
##  [1] "TCF21"     "C8orf85"   "ANKS1B"    "ST8SIA6"   "FOXF1"     "CTNND2"
##  [7] "LOC400550" "SCGB1D2"   "FOXQ1"     "GSTP1"     "SCN4A"     "SLC1A2"
## [13] "MPZL2"     "PRND"      "MAB21L1"   "LRIG1"     "AGTR1"     "PTGER3"
## [19] "COL7A1"    "MAGED2"
## [1] "UCEC"
##  [1] "GPA33"    "CDX1"     "MYO1A"    "PIP5K1B"  "EPCAM"    "C9orf152"
##  [7] "CFTR"     "TDGF1"    "GUCY2C"   "MEP1A"    "TFF3"     "SLC6A7"
## [13] "NKX2.3"   "BTNL8"    "PCK1"     "IHH"      "FRMD1"    "C6orf223"
## [19] "ACY3"     "TUSC3"
```

```
library(mltools)
mcc <- mcc(pred, truth)
mcc
```

```
## [1] 0.8486713
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
table <- data.frame(cm$table)
q <- ggplot(table, aes(truth, pred, fill= Freq)) +
geom_tile(aes(fill = Freq), colour = "black") +
geom_text(aes(label=Freq)) +
scale_fill_gradient(low="white", high="purple") +
```

```
labs(x = "Reference",y = "Prediction") +
scale_x_discrete(labels = colnames(index)) +
scale_y_discrete(labels = colnames(index)) +
ggtitle("Multi classification by naive bayes") +
theme(plot.title = element_text(hjust = 0.5),
axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 0.5))
q
```