Assignment 3 - Part 2 - Diagnosing Schizophrenia from Voice

Riccardo Fusaroli

October 17, 2017

Assignment 3 - Diagnosing schizophrenia from voice

In the previous part of the assignment you generated a bunch of "features", that is, of quantitative descriptors of voice in schizophrenia, focusing on pitch. In the course of this assignment we will use them to try to automatically diagnose schizophrenia from voice only, that is, relying on the set of features you produced last time, we will try to produce an automated classifier.

Question 1: Can you diagnose schizophrenia from pitch range only? If so, how well?

Build a logistic regression to see whether you can diagnose schizophrenia from pitch range only.

RESPONSE: Yes, we can significantly predict diagnosis based on range (p-value = 0.007).

Calculate the different performance measures (accuracy, sensitivity, specificity, PPV, NPV, ROC curve) on a logistic regression using the full dataset. Don't forget the random effects!

Then cross-validate the logistic regression and re-calculate performance on the testing folds. N.B. The cross-validation functions you already have should be tweaked: you need to calculate these new performance measures.

N.B. the predict() function generates log odds (the full scale between minus and plus infinity). Log odds > 0 indicates a choice of 1, below a choice of 0. N.B. you need to decide whether calculate performance on each single test fold or save all the prediction for test folds in one datase, so to calculate overall performance. N.B. Now you have two levels of structure: subject and study. Should this impact your cross-validation?

Question 2 - Which single acoustic predictor is the best predictor of diagnosis?

RESULT: From the point of view of diagnosing as many schizopherinic as possible so no schizophrenic goes without treatment, we aim for that our model has a relatively high sensitivity and accuracy. Based on this, the best simple feature model, we can produce, is diagnosis explained by interquartile range (IQR) and as random effects a random intercept for the effect of study together with a by-subject random slope for trial (diagnosis ~ iqr + (1+trial|Subject) + (1|study))

Question 3 - Which combination of acoustic predictors is best for diagnosing schizophrenia?

Now it's time to go wild! Use all (voice-related) variables and interactions you can think of. Compare models and select the best performing model you can find.

Remember: - Out-of-sample error crucial to build the best model! - After choosing the model, send Malte and Riccardo the code of your model

```
set.seed(3)
pitch_data$diagnosis<- as.factor(pitch_data$diagnosis)
folds <- createFolds(unique(pitch_data$Subject), 10)
pitch_data$Subject <- as.numeric(as.factor(pitch_data$Subject))
```

```
meanPred = rep(NA, nrow(pitch_data))


# Loop for mean feature
for (i in 1:length(folds)){
  f <- folds[[i]]
  train = filter(pitch_data,!(Subject %in% f))
  test = filter(pitch_data,(Subject %in% f))
  model = glmer(diagnosis ~ mean + (1+trial|Subject) + (1|study), train, family="binomial")
  pitch_data$mean_pred[pitch_data$Subject %in% f] = GMCM:::inv.logit(predict(model, test, allow.new.leve
  pitch_data$mean_diag[pitch_data$mean_pred>0.5]="1"
  pitch_data$mean_diag[pitch_data$mean_pred<=0.5]="0"
  pitch_data$mean_diag <- as.factor(pitch_data$mean_diag)
  accuracyTest <- accuracy(pitch_data$mean_diag[which(pitch_data$Subject %in% f)],  pitch_data$diagnosis
  sensitivityTest <- sensitivity(pitch_data$mean_diag[which(pitch_data$Subject %in% f)], reference = pi
  specificityTest <- specificity(pitch_data$mean_diag[which(pitch_data$Subject %in% f)], reference = pi
  ppvTest <- posPredValue(pitch_data$mean_diag[which(pitch_data$Subject %in% f)], reference = pitch_data
  npvTest <- negPredValue(pitch_data$mean_diag[which(pitch_data$Subject %in% f)], reference = pitch_data
  temp_df <- data_frame(accuracyTest = accuracyTest, sensitivityTest = sensitivityTest, specificityTest
  if (i == 1){
    result_df <- temp_df
  } else {
    result_df <- rbind(result_df, temp_df)
  }
}

mean_performance_means <- colMeans(result_df[-6])
mean_performance_means
```

```
##    accuracyTest sensitivityTest specificityTest         ppvTest
##       0.5152208       0.4410050       0.5951162       0.5137029
##         npvTest
##       0.5162881
```

```
#Previous best model
model = glmer(diagnosis ~ iqr + (1+trial|Subject) + (1|study), train, family="binomial")

#Models that did not fail to converge in cv:
model = glmer(diagnosis ~ range + iqr + (1+trial|Subject) + (1|study), train, family="binomial") # A: 0
model = glmer(diagnosis ~ iqr + sd + (1+trial|Subject) + (1|study), train, family="binomial") # A: 0.58
model = glmer(diagnosis ~ iqr + mad + (1+trial|Subject) + (1|study), train, family="binomial") # A: 0.5
model = glmer(diagnosis ~ iqr + max + (1+trial|Subject) + (1|study), train, family="binomial") # A: 0.5

#Models that fail to converge in cv:
#- with interaction:
model = glmer(diagnosis ~ range*iqr + (1+trial|Subject) + (1|study), train, family="binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.0125704 (tol =
## 0.001, component 1)
```

```
model = glmer(diagnosis ~ iqr*mean + range + (1+trial|Subject) + (1|study), train, family="binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.0109434 (tol =
```

```
## 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
model = glmer(diagnosis ~ range*iqr*mean + (1+trial|Subject) + (1|study), train, family="binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : unable to evaluate scaled gradient
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge: degenerate Hessian with 1 negative
## eigenvalues
```

```
model = glmer(diagnosis ~ iqr*sd + (1+trial|Subject) + (1|study), train, family="binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
model = glmer(diagnosis ~ iqr*mad + range + (1+trial|Subject) + (1|study), train, family="binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.00787973 (tol =
## 0.001, component 1)
```

```
model = glmer(diagnosis ~ iqr*coefvar + (1+trial|Subject) + (1|study), train, family="binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.00687516 (tol =
## 0.001, component 1)
```

```
model = glmer(diagnosis ~ iqr*max + (1+trial|Subject) + (1|study), train, family="binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.0144198 (tol =
## 0.001, component 1)
```

```
#- without interaction:
model = glmer(diagnosis ~ range + iqr + mean + (1+trial|Subject) + (1|study), train, family="binomial")
model = glmer(diagnosis ~ range + iqr + sd + (1+trial|Subject) + (1|study), train, family="binomial")
model = glmer(diagnosis ~ range + iqr + mad + (1+trial|Subject) + (1|study), train, family="binomial")
model = glmer(diagnosis ~ range + sd + (1+trial|Subject) + (1|study), train, family="binomial")
model = glmer(diagnosis ~ sd + iqr + mad + (1+trial|Subject) + (1|study), train, family="binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.00396505 (tol =
## 0.001, component 1)
```

```
model = glmer(diagnosis ~ sd + iqr + coefvar + (1+trial|Subject) + (1|study), train, family="binomial")
model = glmer(diagnosis ~ iqr + coefvar + (1+trial|Subject) + (1|study), train, family="binomial")
model = glmer(diagnosis ~ iqr + sd + mean + (1+trial|Subject) + (1|study), train, family="binomial")
model = glmer(diagnosis ~ iqr + min + (1+trial|Subject) + (1|study), train, family="binomial")
model = glmer(diagnosis ~ coefvar + sd + (1+trial|Subject) + (1|study), train, family="binomial")
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.0202255 (tol =
## 0.001, component 1)
```

RESULT: Using crossvalidation to test the performance of the different constructed models, we find that no other converging model tops the IQR-model from question 2. Thus, the best performing model is still

diagnosis explained by interquartile range (IQR) and as random effects a random intercept for the effect of study together with a by-subject random slope for trial (diagnosis ~ iqr + (1+trial|Subject) + (1|study))

Question 4: Properly report the results

METHODS SECTION: how did you analyse the data? That is, how did you extract the data, designed the models and compared their performance?

RESULTS SECTION: can you diagnose schizophrenia based on voice? which features are used? Comment on the difference between the different performance measures.

METHODS: The pitch data from the recordings of schizophrenic and control participants was quite comprehensive. In order to make this more manageable, we downsampled by extracting acoustic features of the recordings per participant per trial i.e. range, mean, max, min, iqr, etc.

To test diagnosis predicted by the acoustic features, we constructed a 10-fold crossvalidation, comparing models predicting diagnosis by acoustic features with similar random effects. Crossvalidation allows us to minimize the out-sample-error, so our model will deal better with new data.

To rate the performance of the model, we extracted measures such as accuracy, sensitivity, specificity, positive prediction values and negative prediciotn values. From the point of view of diagnosing as many schizopherinic as possible so no schizophrenic goes without treatment, we aim for that our model has a relatively high sensitivity and accuracy.

RESULTS: The model with the highest sensitivity and accuracy measure is the model, where diagnosis is explained by the interquartile range (IQR) and as random effects a random intercept for the effect of study together with a by-subject random slope for trial (diagnosis ~ iqr + (1+trial|Subject) + (1|study)). The model does significantly predict diagnosis based on the effect of IQR, however, based on the accuracy measure, we are not much better than chance diagnosing schizophrenia (accuracy = 0.59).

Bonus question 5

You have some additional bonus data involving speech rate, pauses, etc. Include them in your analysis. Do they improve classification?

Bonus question 6

Logistic regression is only one of many classification algorithms. Try using others and compare performance. Some examples: Discriminant Function, Random Forest, Support Vector Machine, etc. The package caret provides them.