

Analysis of High Dimensional Data

Addendum to project assignment 2015-2016

Alternative Dataset for Project

Introduction

The dataset provided to you for the project assignment (part of American Gut project) may still be used for your project assignment, but do not expect to find good prediction models. Since students often get more fun out of building models that seem to work pretty well, we give you the option to use another microbiome dataset (see further). You are completely free to choose: either you continue working with the American Gut data, or you shift to the new data. It will have no effect whatsoever on your marks.

The original assignment remains unchanged; only the data and the outcome to be predicted may be exchanged with the data presented in this document.


New dataset: diabetes in infants

The new data comes from Kostic et al. (2015). You can find the paper on Minerva, but there is no need to read to full paper. A group of 33 infants who were genetically predisposed to develop type I diabetes (T1D) were followed over time. At several points in time the gut microbiome was characterised, resulting in 777 biological samples from these 33 infants. You are given the abundances of 2239 OTUs. The dataset also contains the following variables:

- T1D: defined as 0 when no T1D was diagnosed, and 1 when T1D was diagnosed (i.e. 0 refers to a control and 1 to a case)
- Age: age (in days) when the gut sample was taken

You may choose to either predict the binary T1D or the continuous Age variable.

You are given three datasets:

-  OTUTable.RData: data in OTUTable object. The rows refer to the 777 biological samples. The first 2239 columns refer to the OTUs, for which the abundances are given (counts). The last two columns are T1D and Age.
- OTUTableRel.RData: same as OTUTable.RData, but now for the 2239 OTUs the relative abundances are given (i.e. row sums are equal to 1). This is biologically more relevant for prediction, but perhaps that the original data can be used for use of some graphical exploration methods you find in the literature (this is not a hint; it's just for completeness).
- phyloD.RData: a phyloseq dataset (An R S4 object). This is the original complete dataset. If you are interested in the data, you find here more info on the taxonomy of the OTUs and some more info on the samples (see R code further down). You do not necessarily need this dataset; it's only for the interested student.

R code for reading the data

```
setwd("~/dropbox/education/AnalysisHighDimensionalData/project1516")

# read data
load("OTUTable.RData") # data in OTUTable
load("OTUTableRel.RData") # data in OTUTableRel

# load the original complete data in the phyloseq R package (need to installed from Bioconductor)
source("https://bioconductor.org/biocLite.R")
```

```
## Bioconductor version 3.2 (BiocInstaller 1.20.1), ?biocLite for help
```

```
biocLite("phyloseq")
```

```
## BioC_mirror: https://bioconductor.org
```

```
## Using Bioconductor 3.2 (BiocInstaller 1.20.1), R 3.2.4 (2016-03-10).
```

```
## Installing package(s) 'phyloseq'
```

```
##
```

```
## The downloaded binary packages are in
```

```
## /var/folders/h_/33kgb2tx55z24dl_mhcz9jz80000gn/T//RtmpNekWKP/downloaded_packages
```

```
## Old packages: 'car', 'digest', 'doBy', 'evaluate', 'formatR', 'glmnet',
## 'htmltools', 'knitr', 'latticeExtra', 'lava', 'lme4', 'mclust',
## 'MCMCpack', 'multcomp', 'mvtnorm', 'nleqslv', 'nlme', 'pbkrtest',
## 'permute', 'quantreg', 'Rcpp', 'RcppEigen', 'RCurl', 'rJava',
## 'rmarkdown', 'S4Vectors', 'SuppDists', 'TH.data', 'vegan', 'VGAM', 'XML'
```

```
library(phyloseq)
```

```
## Warning: replacing previous import 'BiocGenerics::Position' by
## 'ggplot2::Position' when loading 'phyloseq'
```

```
load("phyloD.RData")
```

```
# sample information
```

```
SampleData<-phyloD@sam_data
```

```
head(SampleData)
```

```
## Sample Data:      [6 samples by 65 sample variables]:
##      G_id Subject_ID Case_Control Gender Delivery_Route T1D_Diagnosed
## G37016 G37016      E016924   control female          vaginal          false
## G36918 G36918      T014292   control female          vaginal          false
## G37044 G37044      E029817   control  male          vaginal          false
## G37009 G37009      T026177   control female          vaginal          false
## G37029 G37029      E029817   control  male          vaginal          false
## G37035 G37035      T012808   control female          vaginal          false
##      Post_T1D_Diag HLA_Risk_Class AAB_positive AAB_Post_LastNeg
```

##	G37016	false	2	true	false
##	G36918	false	2	false	false
##	G37044	false	3	false	false
##	G37009	false	3	false	false
##	G37029	false	3	false	false
##	G37035	false	2	false	false
##	AAB_Post_FistPos AbxExposureAbsolute PostAbxExposure				
##	G37016	false	false	false	
##	G36918	false	true	true	
##	G37044	false	false	false	
##	G37009	false	true	true	
##	G37029	false	false	false	
##	G37035	false	true	false	
##	AbxAtCollection AbxPreCollection IllnessAtCollection IAA_Level				
##	G37016	no_abx	no_abx	no_illness	0.790
##	G36918	no_abx	Amoxicillin	no_illness	0.560
##	G37044	no_abx	no_abx	no_illness	0.000
##	G37009	no_abx	Amoxicillin	no_illness	0.950
##	G37029	no_abx	no_abx	no_illness	0.370
##	G37035	no_abx	no_abx	no_illness	0.445
##	GADA_Level IA2A_Level ZNT8A_Level ICA_Level IAA_Positive				
##	G37016	0.00	0.030	0.140	0 false
##	G36918	0.21	0.130	0.070	0 false
##	G37044	0.00	0.190	0.080	0 false
##	G37009	0.05	0.110	0.100	0 false
##	G37029	0.00	0.140	0.080	0 false
##	G37035	0.00	0.061	0.063	0 false
##	GADA_Positive IA2A_Positive ZNT8A_Positive ICA_Positive Flowcell				
##	G37016	false	false	false	false A2G30
##	G36918	false	false	false	false A2G30
##	G37044	false	false	false	false A2G30
##	G37009	false	false	false	false A2G30
##	G37029	false	false	false	false A2G30
##	G37035	false	false	false	false A2G30
##	Total_Reads Read_Depth_Class DNA_Concentration DNA_Yield_Class				
##	G37016	85792	1	46.3173409	2
##	G36918	38494	1	2.4840753	1
##	G37044	20977	0	0.7353876	0
##	G37009	13404	0	1.3903483	0
##	G37029	40192	1	4.3934765	1
##	G37035	40085	1	1.4215868	0
##	Age_at_Collection Container_1 Container_2 Tech_rep_exists Country				
##	G37016	547	C0-6616550	C0-6426557	false Finland
##	G36918	530	C0-6616307	C0-1694630	true Estonia
##	G37044	184	C0-6613160	C0-6426557	false Finland
##	G37009	541	C0-6616550	C0-6426557	false Estonia
##	G37029	520	C0-6613160	C0-6426557	false Finland
##	G37035	251	C0-6613160	C0-6426557	false Estonia
##	Collection_Location Exclusive_BF BF Infant_Formula				
##	G37016	Jorvi	false	false	true
##	G36918	Tarto	false	true	false
##	G37044	Jorvi	false	true	false
##	G37009	Tarto	false	false	true
##	G37029	Jorvi	false	false	false

```

## G37035          Tarto          false false          true
##      Fruits_Berries  Corn Rice Wheat  Oat Barley  Rye Buckwheat_Millet
## G37016          true  true true  true true   true true          false
## G36918          true false true  true true   false true          false
## G37044          true false true  true true   true true          false
## G37009          true  true true  true true   true true          true
## G37029          true false true  true true   true true          false
## G37035          true  true true false true   false true          true
##      Cereal Root_Veg  Veg  Eggs Soy_Prod Milk_Prod Meat  Fish Solid_Food
## G37016  true      true true  true      true      true true  true      true
## G36918  true      true true  true      false      true true  true      true
## G37044  true      true true  false     false      true true  true      true
## G37009  true      true true  false     false      true true  false     true
## G37029  true      true true  true      false      true true  true      true
## G37035  true      true true  false     false      false true  false     true
##      IAA_Level_Bin GADA_Level_Bin IA2A_Level_Bin ZNT8A_Level_Bin
## G37016          neg          zero          neg          neg
## G36918          neg          neg          neg          neg
## G37044          zero          zero          neg          neg
## G37009          neg          neg          neg          neg
## G37029          neg          zero          neg          neg
## G37035          neg          zero          neg          neg
##      ICA_Level_Bin BF_Exclusive_Duration BF_Exclusive_Positive
## G37016          zero          0          false
## G36918          zero          0          false
## G37044          zero          87          true
## G37009          zero          0          false
## G37029          zero          87          true
## G37035          zero          2          true
##      BF_Long_Term
## G37016          false
## G36918          true
## G37044          true
## G37009          true
## G37029          true
## G37035          false

```

OTU taxonomy information

```

TaxData<-phyloD@tax_table
TaxData[1:10,]

```

```

## Taxonomy Table:      [10 taxa by 7 taxonomic ranks]:
##      Kingdom      Phylum      Class
## 4333897 "Bacteria" "Proteobacteria" "Gammaproteobacteria"
## 190162  "Bacteria" "Firmicutes"   "Clostridia"
## 134726  "Bacteria" "Firmicutes"   "Bacilli"
## 679245  "Bacteria" "Firmicutes"   "Bacilli"
## 289734  "Bacteria" "Firmicutes"   "Clostridia"
## 302049  "Bacteria" "Firmicutes"   "Clostridia"
## 197991  "Bacteria" "Bacteroidetes" "Bacteroidia"
## 3903651 "Bacteria" "Firmicutes"   "Clostridia"
## 184922  "Bacteria" "Firmicutes"   "Clostridia"
## 193946  "Bacteria" "Firmicutes"   "Clostridia"
##      Order      Family      Genus      Species

```

## 4333897	"Enterobacteriales"	"Enterobacteriaceae"	" "	" "
## 190162	"Clostridiales"	"Lachnospiraceae"	"Blautia"	" "
## 134726	"Lactobacillales"	"Lactobacillaceae"	"Lactobacillus"	" "
## 679245	"Lactobacillales"	"Lactobacillaceae"	"Lactobacillus"	" "
## 289734	"Clostridiales"	"Lachnospiraceae"	" "	" "
## 302049	"Clostridiales"	"Lachnospiraceae"	"Blautia"	" "
## 197991	"Bacteroidales"	"Bacteroidaceae"	"Bacteroides"	" "
## 3903651	"Clostridiales"	"Ruminococcaceae"	"Oscillospira"	" "
## 184922	"Clostridiales"	"Veillonellaceae"	"Dialister"	" "
## 193946	"Clostridiales"	"Lachnospiraceae"	"Blautia"	" "

References

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A. M., ... & Lähdesmäki, H. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell host & microbe*, 17(2), 260-273.