Lab 7a: Machine learning exercises

Contents

1 Exploratory hierarchical clustering with shiny **1**
- 1.1 A basic function . . . 1
- 1.2 Application to tissue discrimination . . . 1
- 1.3 Exercises. . . . 1

2 NMF with drosophila expression patterns **2**
- 2.1 The drosmap package . . . 2
- 2.2 Expression patterns . . . 3
- 2.3 Exercises . . . 3

1 Exploratory hierarchical clustering with shiny

1.1 A basic function

The R source program `dfHclust.R` is in the github repository for lab 7. Source it into your R session, and then verify that it works with the call

```
data(mtcars)
dfHclust(mtcars, labels=rownames(mtcars))
```

If it fails, add any missing libraries, and do what it takes to get it to work. Interrupt the shiny session to proceed.

1.2 Application to tissue discrimination

Set up inputs to `dfHclust` using the `tissuesGeneExpression` data.

```
library(tissuesGeneExpression)
data(tissuesGeneExpression)
df = data.frame(t(e))
no = which(tab$SubType == "normal")
df = df[no,]
tisslabel = tab$Tissue[no]
```

Use `dfHclust(df[,1:50], tisslabel)` as a check. Interrupt

1.3 Exercises.

1.3.1 Symbol mapping

Map the column names of `df` to gene symbols. Use `hgu133a.db`. Remove columns with unmappable symbols and rename the remaining columns with the symbols.

```
library(hgu133a.db)
nids = mapIds(hgu133a.db, keys=
   sub("^X", "", colnames(df)), keytype="PROBEID", column="SYMBOL")
## 'select()' returned 1:many mapping between keys and columns
```

```
bad = which(is.na(nids))
if (length(bad)>0) {
  df = df[,-bad]
  nids = nids[-bad]
  colnames(df) = nids
  }
dim(df)
## [1]    85 21112
```

Interrupt the shiny session and use the new `df` as input. Note that the clustering is based on three genes by default. Other default choices for the clustering are - the object:object distance used - the agglomeration algorithm - the height at which the tree is cut to define clusters

Shift the view to the `silhouette` plot. With the default settings, the average silhouette value for five clusters is 0.35.

Increase the `height for cut` value to 8. How many clusters are declared, and what is the average silhouette value?

Add the gene CCL5 to the feature set used for clustering. Now how many clusters are declared?

Interrupt the shiny session to proceed.

1.3.2 Alphabetizing the selection options

Modify `df` so that the column names are in alphabetical order. Use `dfHclust(df[,1:50], tisslabel)` for the new ordering. What is the average silhouette value for the default choices of `dfHclust` settings?

Change the clustering method to `ward.D2`. What is the new average silhouette value?

Interrupt the shiny session to proceed.

1.3.3 Clustering with a gene set

We have used an arbitrary selection of genes for these illustrations. Consider the idea that steady-state expression pattern of genes that are used to perform splicing is important for tissue differentiation. We can get a list of relevant genes on the hgu133a array as follows.

```
# using GO.db
            GOID                                  TERM
22097 GO:0045292 mRNA cis splicing, via spliceosome
```

```
splg = select(hgu133a.db, keys="GO:0045292", keytype="GO", columns="SYMBOL")
## 'select()' returned 1:many mapping between keys and columns
tokeep = intersect(splg$SYMBOL, colnames(df))
dfsp = df[, tokeep]
```

You should have 7 genes available after these operations. Use `dfsp` with dfHclust, and select all genes for clustering.

As you add spliceosome-annotated genes into the clustering, does the appearance of the clustering tree improve?

2 NMF with drosophila expression patterns

2.1 The drosmap package

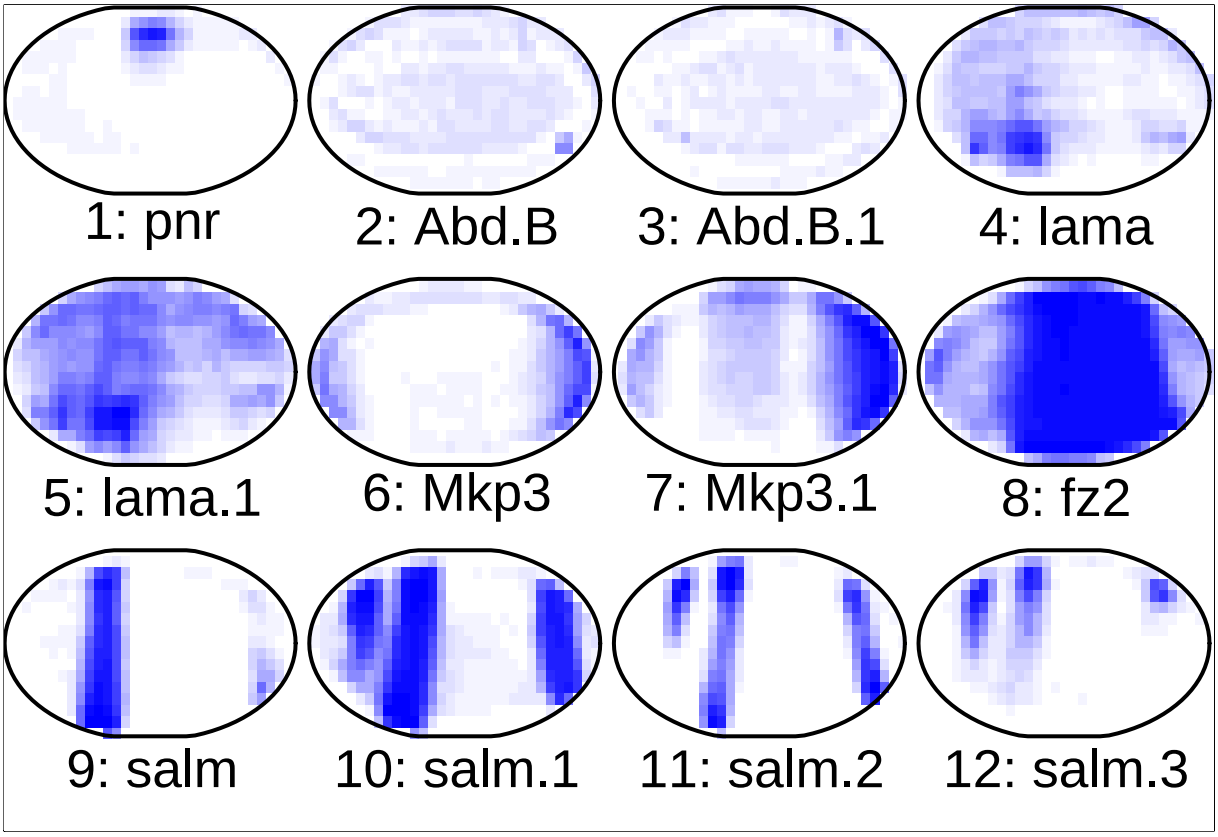
Install and attach the `drosmap` package. This is a simple repackaging of code and data provided at BDGP.

```
library(BiocInstaller)
biocLite("vjcitn/drosmap")
library(drosmap)
```

2.2 Expression patterns

A data.frame of spatially recorded gene expression patterns derived from blastocyst samples is available. We'll display some examples.

```
library(drosmap)
data(expressionPatterns)
data(template)
imageBatchDisplay(expressionPatterns[,1:12],
  nrow=3, ncol=4, template=template[,-1])
```



2.3 Exercises

2.3.1 Comparing non-negative matrix factorizations of the expression pattern matrix

We'll reduce the data matrix (for convenience) to 701 unique genes

```
data(uniqueGenes)
uex = expressionPatterns[,uniqueGenes]
```

We'll begin with a factorization using a basis of rank 10.

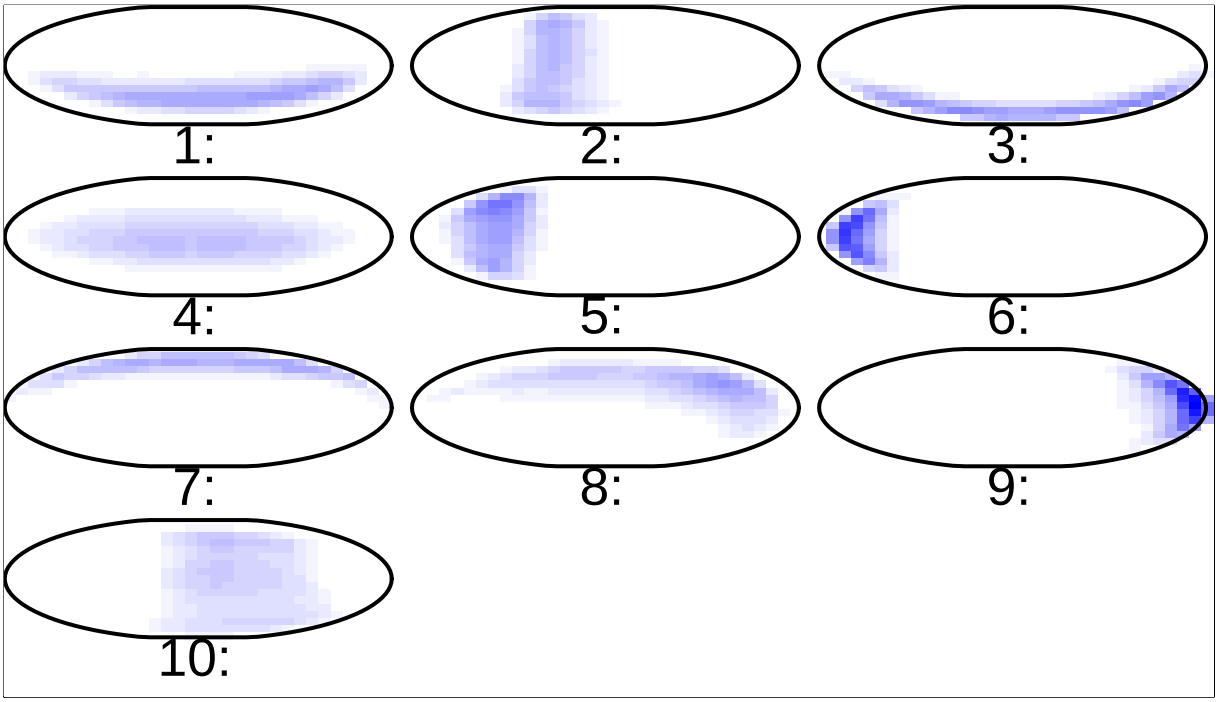
```
set.seed(123)
library(NMF)
m10 =nmf(uex, rank=10)
m10
## <Object of class: NMFfit>
##  # Model:
##   <Object of class:NMFstd>
##   features: 405
##   basis/rank: 10
##   samples: 701
##  # Details:
##   algorithm:  brunet
##   seed:  random
##   RNG: 403L, 624L, ..., 2099891502L [e38d032700af470a3a1013304e0fcab6]
##   distance metric:  'KL'
##   residuals:  4901.298
##   Iterations: 2000
##   Timing:
##      user  system elapsed
##    44.356   4.256  51.406
```

The authors of the Wu et al. 2016 PNAS paper justify a rank 21 basis.

```
set.seed(123)
library(NMF)
m21 =nmf(uex, rank=21)
```

To visualize the clustering of the expression patterns with the rank 10 basis, use

```
imageBatchDisplay(basis(m10), nrow=4,ncol=3,template=template[,-1])
```



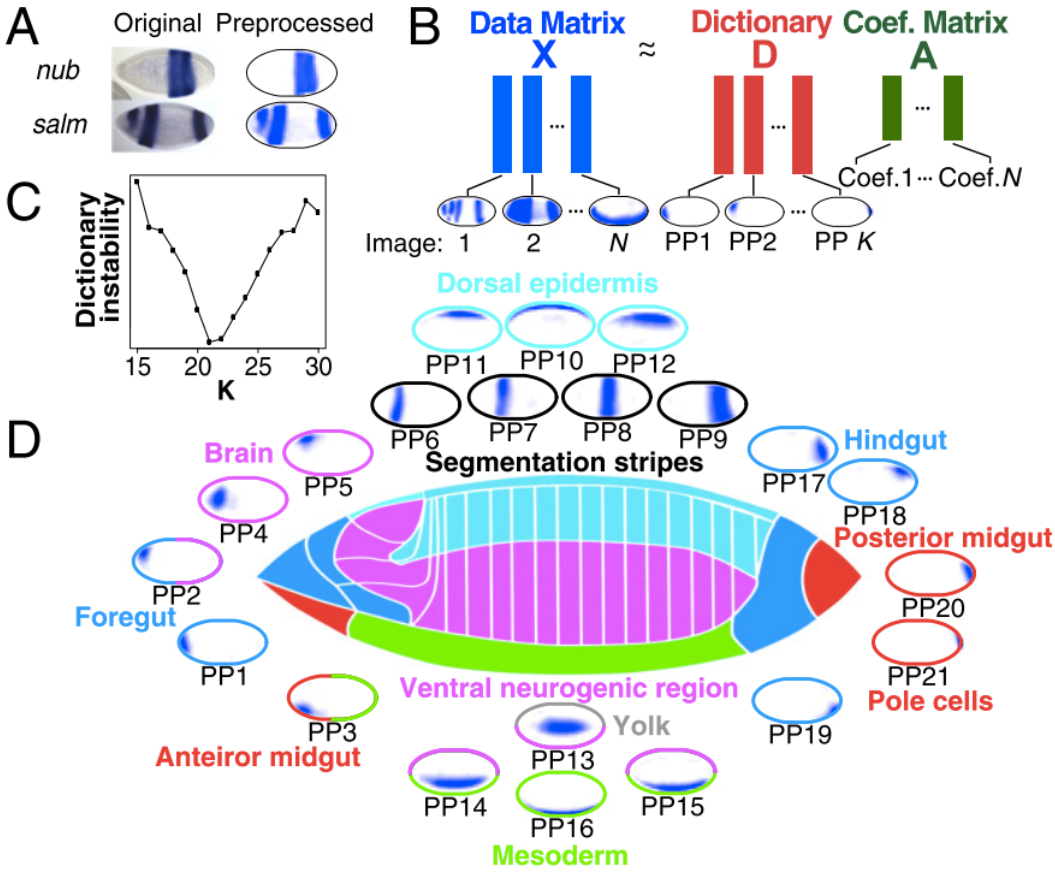
The 'predicted' matrix with the rank 10 basis is

```
PM10 = basis(m10)%*%coef(m10)
```

Compare the faithfulness of the rank 10 and rank 21 approximations.

2.3.2 Comparison to a cell fate schematic

Produce the display of the `m21` basis with `imageBatchDisplay` and check that the constituents are similar to those shown as principal patterns below (from the Wu et al. paper).



Can any of the patterns found with the rank 10 basis be mapped to key anatomical components of the blastocyst fate schematic?