Crash course introduction to prediction computation

Paul Yousefi, PhD, MPH

June 2019

Before we start

This lab will performed in R and will use the following packages:

Add downloads

We'll be using data from Tibshirani et al. that is publicly available on the gene expression ombibus (GEO) website

All course material, including the data and code used for this lab practical, is available for you to download here:

url for where to download data

Goals

- Partitioning data into training and testing sets
- Evaluating performance of risk scores
- Fitting models in training data
- Predicting outputs from those models in the testing data
- Quantifying model prediction performance

Getting started

To start, I'll load our data into active memory and have a look at what's available:

```
load("dataset.rda")
```

```
ls()
```

[1] "meth" "samples"

So we have two data objects:

- `meth` with DNA methylation data
- `samples` with other phenotype information on the participants of this study

Let's get a better sense of the variables available in `samples`:

```
str(samples)
```

```
> 'data.frame': 464 obs. of  6 variables:
>  $ gsm       : chr  "GSM1225377" "GSM1225378" "GSM1225379" "GSM1225380" ...
>  $ gse       : chr  "GSE50660" "GSE50660" "GSE50660" "GSE50660" ...
>  $ age       : num  50 56 49 64 51 50 47 46 50 56 ...
>  $ sex       : chr  "male" "male" "female" "male" ...
>  $ smoking   : chr  "former" "never" "former" "former" ...
>  $ ever.smoke: num  1 0 1 1 1 1 1 1 1 0 ...
```

```
summary(samples)
```

```
>      gsm                gse                age            sex
>  Length:464         Length:464         Min.   :38.00   Length:464
>  Class :character   Class :character   1st Qu.:50.00   Class :character
>  Mode  :character   Mode  :character   Median :56.00   Mode  :character
>                                        Mean   :55.39
>                                        3rd Qu.:61.00
>                                        Max.   :67.00
>    smoking            ever.smoke
>  Length:464         Min.   :0.0000
>  Class :character   1st Qu.:0.0000
>  Mode  :character   Median :1.0000
>                     Mean   :0.6142
>                     3rd Qu.:1.0000
>                     Max.   :1.0000
```

```
table(samples$smoking)
```

```
>
> current  former   never
>      22     263     179
```

```
table(samples$ever.smoke)
```

```
>
>   0   1
> 179 285
```

The `smoking` variable has 3 categories, but it's easiest to begin with a binary outcome so let's focus on the `ever.smoke` variable that collapses the **current** and **former** subjects into a single category

- When I talk about predicting smoking going from now on I'll be referring to this `ever.smoke`variable

Applying risk scores

The simplest type of risk score we can use for prediction is just a single individual variable. The site `cg05575921` in the *AHRR* gene has consistently been the CpG with methylation showing the strongest association with smoking in several studies looking broadly across the genome.

Perhaps the methylation levels of this site would be sufficient to predict whether someone has been a smoker. To see, let's begin by adding this CpG site as a variable to our phenotype data object `samples`:

```
samples$ahrr <- meth["cg05575921", ]
```

We can use a package called `pROC` to see how well different values of our `ahrr` variable explain smoking status:

```
## load the pROC package
library("pROC")

## use the formula-based syntax of the package
roc(ever.smoke ~ ahrr, data = samples)
```

```
>
> Call:
> roc.formula(formula = ever.smoke ~ ahrr, data = samples)
>
```

```
> Data: ahrr in 179 controls (ever.smoke 0) > 285 cases (ever.smoke 1).
> Area under the curve: 0.851
```

```
roc.out <- roc(ever.smoke ~ ahrr, data = samples)
plot.roc(roc.out)
```