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

Single variable 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:

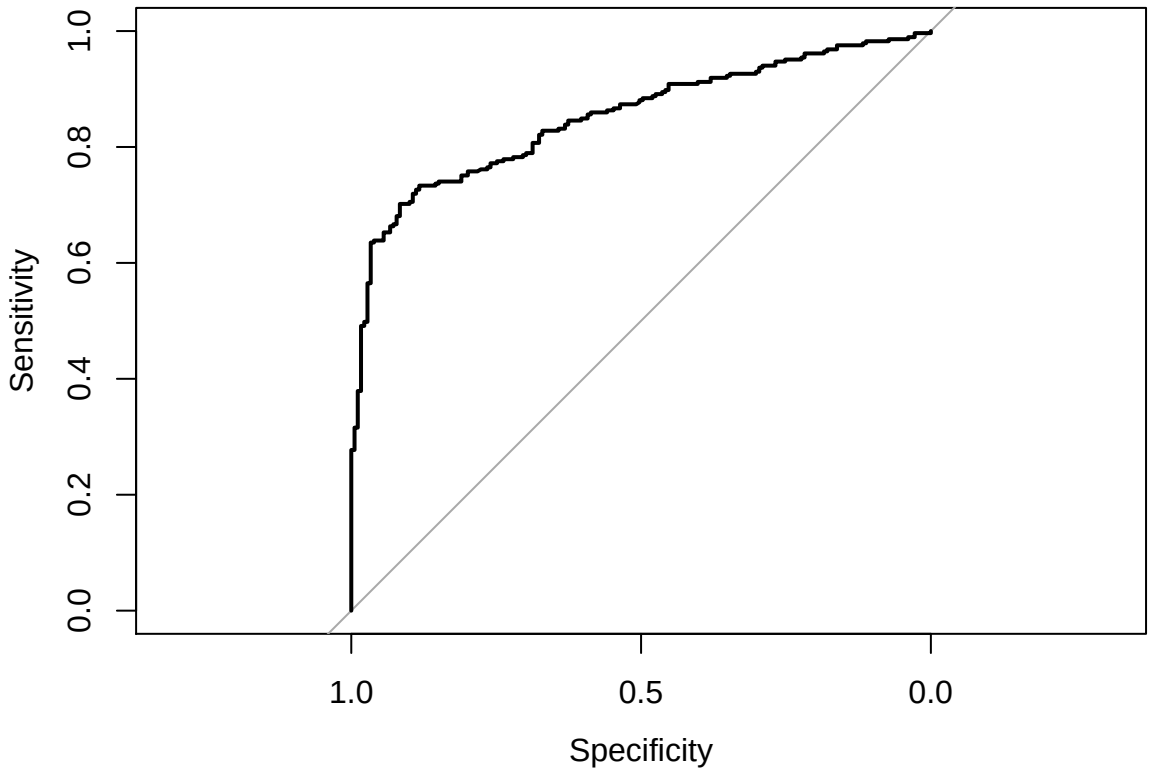
```
## load the pROC package
library("pROC")

## 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
```

We can also visualize our results by using the `pROC` package's `plot.roc()` function on the saved output

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



Weighted risk scores from published coefficients

Another common approach in omic prediction is to apply a risk score using information from multiple loci or omic measures weighted by previously reported magnitudes of association observed between those features and the outcome of interest. This has perhaps been most often performed using genetic data to compose 'polygenic risk scores', but can easily be extended to other types of data input.

For example, in the context of DNA methylation data we can define and subsequently apply a smoking score derived from the published coefficients of the largest blood epigenome wide association study (EWAS) meta-analysis to date in Joehanes et al. 2016:

> Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet 2016;

- The coefficients from their models were distributed in their Supplemental Table 2 that I previously saved and load into R below:

```
load("joehanes2016_st2_bonf.rda")
```

The `joehanes` object has summary information on the 2617 CpGs that were significant at a Bonferroni p-value threshold in the original meta-analysis and that were available in our methylation dataset

```
str(joehanes)
```

```
> 'data.frame': 2617 obs. of  14 variables:
>  $ probe.id            : chr  "cg16145216" "cg19406367" "cg05603985" "cg14099685" ...
>  $ infinium.design.type: chr  "I" "II" "II" "II" ...
>  $ chromosome          : chr  "1" "1" "1" "11" ...
>  $ location..hg19.     : chr  "42,385,662" "66,999,929" "2,161,049" "47,546,068" ...
>  $ strand              : chr  "R" "R" "F" "R" ...
>  $ gene.symbol         : chr  "HIVEP3" "SGIP1" "SKI" "CUGBP1" ...
>  $ effect              : num  0.0298 0.0175 -0.0122 -0.0124 -0.0262 -0.0182 -0.0166 -0.0163 -0.0134 -0
>  $ std..error          : num  0.002 0.0013 0.0009 0.0009 0.002 0.0014 0.0013 0.0013 0.0011 0.0067 ...
>  $ z.value             : num  14.5 13.9 -13.8 -13.7 -13.4 ...
>  $ p.value             : num  6.7e-48 7.0e-44 1.8e-43 1.5e-42 6.1e-41 ...
>  $ fdr                 : num  3.3e-42 1.7e-38 2.8e-38 1.8e-37 5.9e-36 ...
>  $ previously.seen     : int  1 1 1 1 1 0 1 1 1 1 ...
>  $ small.molecules     : int  0 0 0 0 0 0 0 0 0 0 ...
>  $ sign                : chr  "++++++++++-++++" "++++++++++++++++" "----------+--" "----------
```

Let's restrict our big methylation data object, `meth`, to just the CpGs that are in the `joehanes` list. This keeps the CpGs that we expect to be most related to smoking behavior, in the correct format, while reducing the size of the data were working with:

```
X <- meth[joehanes$probe.id, ]
## transpose to make columns = methylation site variables,
##                  rows = subjects/observations
X <- t(X)
```

To compute an individual's risk score using these coefficients, we need to take the sum of each coefficient multiplied by the participant's value at the corresponding variable, for example:

$$\hat{Y}_{JoehanesScore} = \sum_{i}^{2617 CpGs} \hat{\beta}_i X_i$$

- Where $\hat{\beta}_i$ are the individual previously estimated Joehanes et al. coefficients and X_i are their corresponding variables (CpG site measurements in this case).

Equivalently, we can compute the exact same quantity, $\hat{Y}_{JoehanesScore}$, more simply using matrix multiplication:

$$\hat{Y}_{JoehanesScore} = X\hat{\beta}$$

- Where X is an NxP matrix of all P variables being used and $\hat{\beta}$ is a corresponding vector of all Joehanes et al. coefficients.

To implement this, lets start by making a named vector of the `joehanes` coefficients:

```
coefs <- joehanes$effect
names(coefs) <- joehanes$probe.id
```

We can then use matrix multiplication against our observed DNA methylation values to get our $\hat{Y}_{JoehanesScore}$ values:

```
y.hat <-   X %*% coefs
```

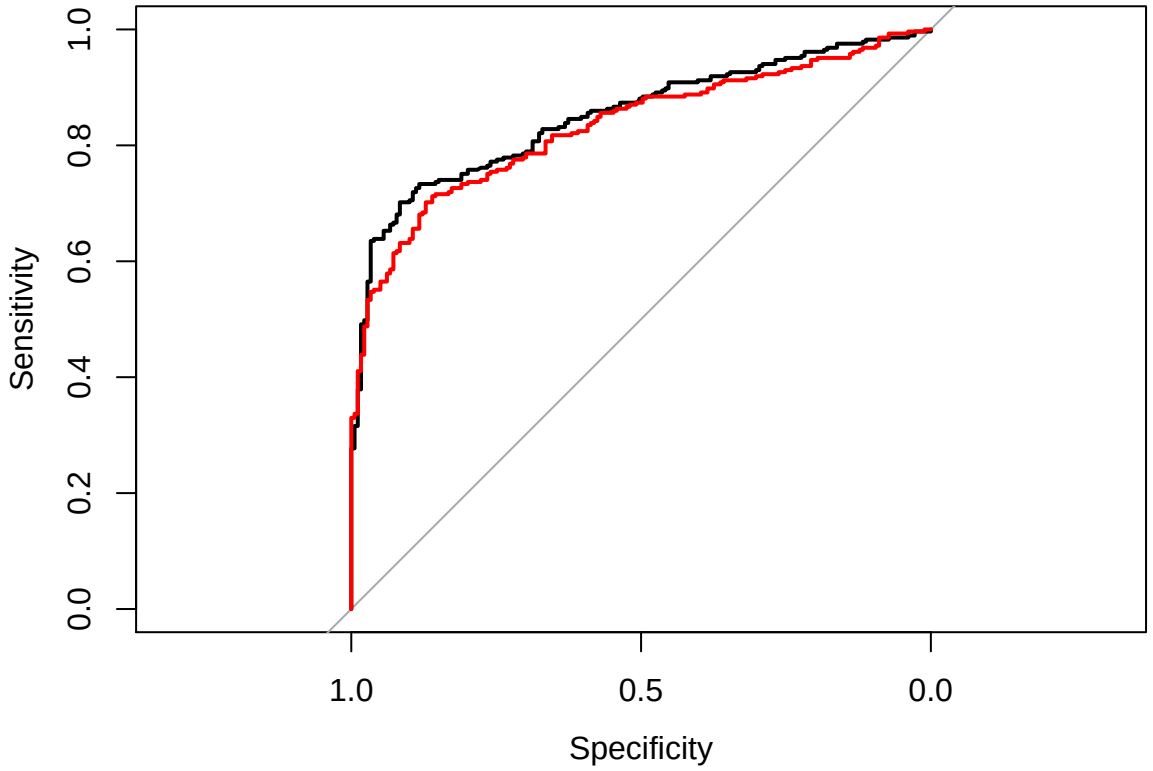
By adding this output as a variable to our `samples` data, we can again use the `pROC` package to evaluate and visualize the prediction performance of this score:

```
samples$y.hat <- as.vector(y.hat)
```

```
roc.out.again <- roc(ever.smoke ~ y.hat, data = samples)
roc.out.again
```

```
>
> Call:
> roc.formula(formula = ever.smoke ~ y.hat, data = samples)
>
> Data: y.hat in 179 controls (ever.smoke 0) < 285 cases (ever.smoke 1).
> Area under the curve: 0.832
```

```
plot.roc(roc.out)
lines.roc(roc.out.again, col="red")
```



Does the joehanes score predict never/ever smoking better than just cg05575921 alone?

```
roc.out$auc
```

```
> Area under the curve: 0.851
```

```
roc.out.again$auc
```

```
> Area under the curve: 0.832
```

No, it doesn't appear to!

Training a novel predictor

```
library(caret)

set.seed(138)
in.train <- createDataPartition(
  y = samples$ever.smoke,
  ## the outcome data are needed
  p = .75,
  ## The percentage of data in the
  ## training set
  list = FALSE
)
```

```
training <- samples[ in.train,]
testing  <- samples[-in.train,]

nrow(training)
```

```
> [1] 348
```

```
nrow(testing)
```

```
> [1] 116
```