Exercise

Session information

hmm homework

Matthew Stephens

5/19/2018

Last updated: 2018-05-19

workflowr checks: (Click a bullet for more information)

- ▶ ✓ **R Markdown file:** up-to-date
- ▶ ✓ **Environment:** empty
- ▶ ✓ **Seed:** `set.seed(20180411)`
- ▶ ✓ **Session information:** recorded
- ▶ ✓ **Repository version:** 225f991 (https://github.com/stephens999/stat34800/tree/225f991bff25554a3d7fae1b31f8917d1147696f)

▶ **Expand here to see past versions:**

Here we simulate a simple HMM with two states, $Z_t \in \{1,2\}$ that represent two different genetic populations. The data X_t is genetic data at locus (position) t, which we will assume are 0 or 1. So the emission distribution at position t in state k is Bernoulli(p_{kt}) where p_{kt} is the frequency of the 1 allele at position t in population k.

The transition matrix for the Markov chain is symmetric, with probability 0.9 of staying in the same state, and 0.1 of switching at each step.

Here is some code to simulate from this:

```
set.seed(1)
T = 1000
K = 2
P = rbind(c(0.9,0.1),c(0.1,0.9))

# simulate the matrix of allele frequencies in each of the K populations at each of T loci
p = matrix(runif(K*T),nrow=K,ncol=T)

# Simulate the latent (Hidden) Markov states
Z = rep(0,T)
Z[1] = 1
for(t in 1:(T-1)){
  Z[t+1] = sample(K, size=1, prob=P[Z[t],])
}

# Work out the corresponding bernoulli probability for each state
prob = rep(0,T)
for(i in 1:T){
  prob[i] = p[Z[i],i]
}

# Simulate the emitted/observed values
X= rbinom(n=T,size=1,prob=prob)
```

Exercise

- Implement the forward and backwards algorithm to "decode" this HMM: that is to compute $\Pr(Z_t = k|X_1, \dots, X_T)$ for each t. Note that because of the longer length of this Markov chain you will need to do some normalizing of the forwards and backwards probabilities to avoid numerical errors.

- Compute the error rate (compared with the truth) if you assign each Z_t to its most probable value.

- Now imagine you did not know the true values for Z. Explain how you would estimate the error rate directly from the output of the forward-backwards calculation. Compare this estimate with that obtained when you did know the truth. (Hint: this is a bit like estimating a False Discovery Rate from the Empirical Bayes Normal Means output)

Session information

```
sessionInfo()
```

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X El Capitan 10.11.6

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base     

loaded via a namespace (and not attached):
 [1] workflowr_1.0.1   Rcpp_0.12.16      digest_0.6.15    
 [4] rprojroot_1.3-2   R.methodsS3_1.7.1 backports_1.1.2  
 [7] git2r_0.21.0      magrittr_1.5      evaluate_0.10.1  
[10] stringi_1.1.7     whisker_0.3-2     R.oo_1.22.0      
[13] R.utils_2.6.0     rmarkdown_1.9     tools_3.3.2      
[16] stringr_1.3.0     yaml_2.1.18       htmltools_0.3.6  
[19] knitr_1.20       
```

This reproducible R Markdown (http://rmarkdown.rstudio.com) analysis was created with workflowr (https://github.com/jdblischak/workflowr) 1.0.1