STAT115 Homework 4

(your name)

2017-02-25

Part I– Hidden Markov Models

CpG islands are stretches of CG-rich sequences in the genome. They are often of functional importance, as 50% of the human genes have a CpG island around 500bp upstream of the transcription start site. Of course, CpG island sequence is not only CG, and non-CpG island sequences could still contain some CG. Therefore, we could use HMM to predict CpG islands by looking at a long stretch of DNA. Now as a HMM practice, suppose that we have a short sequence AGGCGT.

Initial probability: 0.3 of CpG and 0.7 of non-CpG (abbreviated as N).
Transition probability: P(CpG to CpG) = 0.8; P(non-CpG to non-CpG) = 0.7.
Emission probability: P(A, C, G, T | CpG) = (0.1, 0.4, 0.4, 0.1), P(A, C, G, T | non-CpG) = (0.3, 0.2, 0.2, 0.3).

Algorithm	P_1	P_2	P_3	P_4	P_5
	α_1(CpG)	α_2(CpG)	α_3(CpG)	α_4(CpG)	α_5(CpG)
Forward	α_1(N)	α_2(N)	α_3(N)	α_4(N)	α_5(N)
	β_1(CpG)	β_2(CpG)	β_3(CpG)	β_4(CpG)	β_5(CpG)
Backward	β_1(N)	β_2(N)	β_3(N)	β_4(N)	β_5(N)
	γ_1(CpG)	γ_2(CpG)	γ_3(CpG)	γ_4(CpG)	γ_5(CpG)
Forward-backward	γ_1(N)	γ_2(N)	γ_3(N)	γ_4(N)	γ_5(N)

1. Given the data described above, compute the probabilities in the table above. Which set of results is most interpretable?

```
library(HMM)
```

2. Using the Viterbi algorithm, infer the Hidden states path that is most likely to give the sequence observations.

G1. [Graduate Students] Adjust the initial probabilities to a range of values and evaluate what effect the initial probabilities have on the outcome of the forward-backward inference. Show a plot to support your evaluation.

For each of the following problems, what are the observed and hidden states? Concisely interpret the transition and emission probabilities under these examples and give a brief example of each:

3. Inferring protein secondary structure from an amino acid sequence.

4. Inferring the functional annotation (e.g. poised enhancer, promoter) of the genome from ChIP-Seq data.

5. Inferring genome copy number variation from whole genome sequencing data.

G2. [Graduate Students] In 5-6 sentences, provide an overview of the chromHMM model and some of the key results of their findings. Specifically, report what the hidden and observed

states are and summarize which observed states contribute the most to the various hidden states.

Part II– Python programming

For the following, show your code (it doesn't need to compile at run time) and answer the questions to verify that you've completed the assignment.

6. From the `/n/stat115/hws/4/histone` folder, note the H3K27me3, H3K27ac, H3K4me3, H3K36me3 ChIP-seq data in BAM format. Use MACS to remove the redundant reads for each set of reads. How many reads does each .bam file have before and after the removal of redundant reads?

7. From UCSC download page, take the human RefSeq annotation table (the file refGene.txt.gz for hg38; present in Git repository). An annotation of the columns is available here. Similar to what you did in lab 3, write a program to get the promoters of each gene on chromosome 22. The promoters can be defined as +/- 1,000 bp around the transciprtion start site. What is the total number of base pairs included in this file?

8. Using the de-duplicated .bam files from 6, construct a table of promoters x histone modification counts by counting the number of reads that overlap each promoter. This again should be similar code to Homework 3. Make summary plots to show A) how these different modifications are correlated with each other at the promoter and B) how many of each feature is present at the promoters.

G3. [Graduate Students] Repeat 7 and 8 using the full gene bodies (from txStart to txEnd) on chromosome 22. In particular, note the total number of base pairs, the correlations between the modifications, and the number of counts for each feature in each gene body. Noting the difference in total base pairs between gene bodies and promoters, how do the counts of each histone modification feature change between the gene bodies and the promoters?

Part III– Feature selection and regression, epigenetic regulation

One could imagine that the programs in Part II define a subset of data needed to fully model epigenetic regulation of gene expression. Consider a more exhaustive approach that examines many different histone marks and their read distribution over many defined regions in the genome.

For each of the following 10 histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K9ac, H3K27me3, H3K27ac, H3K79me2, H3K36me3, H4K20me1) from the K562 cell line, we already parsed out the following read counts for each RefSeq sequence (in the file `data/histone_marks_read_count_table.txt`): distal promoter [-5KB, -1KB], proximal promoter [-1kb, +1kb] from TSS, gene body (from transcription start to end, including all exons and introns), transcript (concatenate all the exons), first 1/3 of transcript (concatenate all the exons, length-wise), middle 1/3 of transcript, last 1/3 of transcript, all the introns (concatenate all the introns). The table has one line for each RefSeq, and 81 columns (RefSeq ID, 10 histone marks, each with 8 features, so 1 + 10 * 8), the value is log read count for each feature. We also have the expression of the RefSeq (`data/k562expr.txt`).

9. Using LASSO regression, identify features that are predictive of gene expression from the files described above. Report A) the total number of features selected, B) the 5 features with the strongest effect size, and C) the direct of the effect of the effect on gene expression. Hint: use the `caret` package.

```r
exprs <- read.table("data/K562expr.txt", header = TRUE)
histmods <- read.table("data/histone_marks_read_count_table.txt", header = TRUE)
# merge ...
```

G4. [Graduate Students] In addition to LASSO, try another feature selection method (easily done using `caret`). Evaluate the performance of your selected model in terms of speed and performance using cross validation.

Submission

Please submit your solution directly on the canvas website. Please provide your code (.Rmd) and a pdf file for your final write-up. Please pay attention to the clarity and cleanness of your homework. Page numbers and figure or table numbers are highly recommended for easier reference.

The teaching fellows will grade your homework and give the grades with feedback through canvas within one week after the due date. Some of the questions might not have a unique or optimal solution. TFs will grade those according to your creativity and effort on exploration, especially in the graduate-level questions.