EN 600.438 Problem Set 1

Tony Melo
Computational Genomics

February 25, 2017

Problem 1. The following DNA sequence is the template strand (the template strand of DNA is complementary to transcribed RNA). In the sequence below, the exonic regions of the DNA are in bold. The non-bold are intronic regions. **TACACG**TTAGACAT**GCTACG**CTGGCAAC **GGGTACATC**

1. Corresponding RNA sequence:
   AUG|UGC|CGA|UGC|CCC|AUG|UAG|
   Amino acid sequence:
   Met|Cys|Arg|Cys|Pro|Met|STOP|

2. i. **Result of frameshift insertion**: Insert an T into the second exonic region (**GCTACG** becomes **GCTACTG**) such that the new RNA sequence becomes CGA|UGA. Since UGA translates to a stop codon, the rest of the DNA sequence would not be transcribed into RNA.
   New DNA sequence: TACACGTTAGACATGCTACTGCTGGCAACGGGTACATC
   New RNA sequence: AUG|UGC|CGA|UGA| New amino acid sequence: Met|Cys|Arg|STOP|

   ii. **Non-frameshift insertion/deletion**: delete GCT from second exonic region
   New DNA sequence: TACACGTTAGACATACTGCTGGCAACGGGTACATC
   New RNA sequence: AUG|UGC|UGC|CCC|AUG|UAG| New amino acid sequence: Met|Cys|Cys|Pro|Met|STOP|

   iii. **Synonymous single nucleotide change**: Change G at the end of first exonic region to an A
   New DNA sequence: TACACATTAGACATGCTACTGCTGGCAACGGGTACATC
   New RNA sequence: AUG|UGU|CGA|UGC|CCC|AUG|UGA| New amino acid sequence: Met|Cys|Arg|Cys|Pro|Met|STOP|

   iv. **Non-synonymous single nucleotide change**: Change that same G to a C
   New DNA sequence: TACACCTTAGACATGCTACTGCTGGCAACGGGTACATC
   New RNA sequence: AUG|UGG|CGA|UGC|CCC|AUG|UGA| New amino acid sequence: Met|Trp|Arg|Cys|Pro|Met|STOP|

3. i. HBB is on chromosome 11 on approximately position chr11:5227002-5227002
   ii. SNP rs334 is on exonic region 1
   iii. non-synonymous

Problem 2. Genotype at a particular SNP X_1 is associated with color of flower petals. Genotype AA gives rise to red petals and TT gives white petals.

1. Three possible outcomes of heterozygous genotype (AT) at SNP X_1 are that A could be the dominant gene, and the presence of an A leads to red petals, or T could be the dominant gene in that the presence of one T could lead to white petals, lastly, A and T could be codominant alleles where the presence of both A and T would lead to blending, giving a pink color. If A is dominant, then a binary classifier ($C = 2$) where the presence of an A allele would be encoded as a 1 and a homozygous recessive genotype (TT) would be encoded as a 0. Conversely, if encoding T as a dominant allele, we perform the same binary classification, except now we encode the presence of a T allele as a 1 and

Problem 3. Doctors collected the anonymized visit records to the maternity ward of the biggest hospital in a small town. Each year, they counted the number of babies born with known birth defects. For the first 5 years of their study, the number of documented birth defects found per year is shown below: (not going to copy the data but yeah)

1. We can model the given data with a Poisson distribution with mean λ

2. Given a Poisson distribution with mean λ, we find the maximum likelihood estimator by beginning with the function that gives the probability that x takes a certain value

$$P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Now the likelihood function is given as

$$L(x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$L(x_i|\lambda) = \prod_{i=1}^{n} \frac{1}{x_i!} \cdot \lambda^{\sum_{i=1}^{n} x_i} e^{-\sum_{i=1}^{n} \lambda}$$

$$log[L(x_i|\lambda)] = -\sum_{i=1}^{n} log(x_i!) + (\sum_{i=1}^{n} x_i) log(\lambda) - n\lambda$$

Now to find the value $\hat{\lambda}$ that maximizes this likelihood function, take the derivative and set it to 0

$$\frac{dlog[L(x_i|\lambda)]}{d\hat{\lambda}} = \frac{\sum_{i=1}^{n} x_i}{\hat{\lambda}} - n = 0$$

$$n = \frac{\sum_{i=1}^{n} x_i}{\hat{\lambda}}$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

So we see that the maximum likelihood estimator for the Poisson distribution is simply the population mean for the entire data set, which in this case, $\hat{\lambda} = 15$

3. Given that $\hat{\lambda} = 15$, the probability that a year has 5 birth defects is equal to:

$$P(x = 5) = \frac{15^5 e^{-15}}{5!} = 0.002$$

Problem 4. How would you select a SNP to drop from the model such that it will have the least effect on the residual sum of squares? An increase in residual sum of squares will have a negative impact on your model. Justify your answer.

1. To have the least effect on the residual sum of squares, backward stepwise regression removes the data point with the lowest Z-score, which in this case would be SNP A.

Problem 5. Use logistic regression to build a 1) a model to predict group 0 and group 1 breast cancer patients from all genes jointly, and 2) a model to predict group 0 and group 1 Breast cancer patients only from the first 10 genes in the file (b). (Parameters should be trained using train expression.csv and phen train.csv only). Test each model on test data provided. Return precision and recall for your predictions from both models on the test dataset(test expression.csv, phen test.csv). If the performance differed between the two models, what do you think is the reason?

Precision for model using all genes: 0.59999
Precision for model using 10 genes: 0.79166
Recall for model using all genes: 0.5
Recall for model using 10 genes: 0.79166
The reason the performance differed between models is due to the fact that using every gene overtrains the model since the most complex model will be picked in order to minimize the loss on the training set. This leads to poor generalization to test data.

b) We could improve performance by partitioning the data into a larger training set while still withholding a portion of it to use as test data for the model.