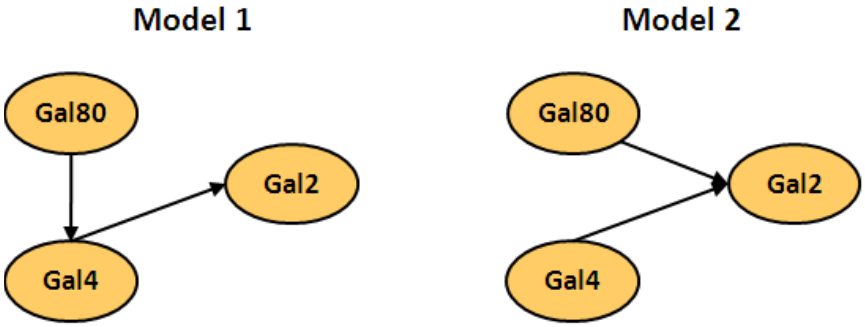
GENOME 560, Spring 2017
Problem Set #4

(Due Mar 24th 11:59pm)

---

1. **[70 points] Model selection to find the best regulatory network**

   In this question, we will implement an algorithm for selecting among various structures of the regulatory network. Specifically, we will focus on two possible models of the galactose regulatory network in *S. cerevisiae*. We will select a model based on the expression data on these three genes measured across *S. cerevisiae* individuals. We discussed this idea in Lecture #12.



   Let's assume that expression levels are binary values (high, low), and we use CPTs (conditional probability tables) for both networks in Model 1 and Model 2.

   (a) **[10 points]** Model 1 and Model 2 have different Bayesian network structures, and so they have different sets of parameters (θ). List all parameters and the CPTs in each of Model 1 and Model 2.

   (b) **[10 points]** Say that we are given the gene expression data D measuring binary expression levels of the 3 genes (Gal80, Gal4 and Gal2) across 112 samples. Write down the likelihood function $L(D|\theta)$ for Model 1 and Model 2.

   (c) **[10 points]** Write down the maximum likelihood estimation (MLE) solutions in Model 1 and Model 2.

   (d) **[30 points]** Download the binary expression data from `https://sites.google.com/a/cs.washington.edu/genome560-spr18/disc-gal80-gal4-gal2.txt?attredirects=0&d=1`, and implement the code that computes the log-likelihood function $\log(L)$ for Model 1 and Model 2. Please submit the code and the resulting values of the log-likelihood function for Model 1 and Model 2.

   (e) **[10 points]** Select between Model 1 and Model 2 based on the results in part (d).

2. **[30 points] MLE vs MAP estimation**

   Here, we will continue the R exercise in Lecture #14. The goal of this exercise is to understand the impact of hyperparameters α and β in a Bernoulli experiment (e.g., *Thumbtack* example). We will do that by comparing the shape of the distribution between the likelihood $P(D|p)$ and the posterior $P(p|D)$. We assume that our prior belief is that we get the same number of heads and tails (i.e., $p = 0.5$).

Consider the following sets of n_H, n_T, α, and β. Plot the likelihood and posterior functions over varying p in [0, 1] by using the R command *par(mfrow=c(2,1))* to compare multiple plots. What are the MLE and MAP estimations of p in each case?

(a) **[10 points]** Say that $n_H = 100$, $n_T = 50$, $\alpha = 5$, and $\beta = 5$.

(b) **[10 points]** Say that $n_H = 100$, $n_T = 50$, $\alpha = 30$, and $\beta = 30$. (Optional: Describe the difference of the results between (a) and (b) and the reason for the difference.)

(c) **[10 points]** Say that $n_H = 10$, $n_T = 5$, $\alpha = 30$, and $\beta = 30$. (Optional: Describe the difference of the results between (b) and (c) and the reason for the difference.)

3. **[70 points] Quantitative trait loci analysis for cholesterol levels**

We are given the genotype and phenotype data from 334 mouse individuals. The genotype data measure binary genotype values of 1333 genetic markers for each mouse, and the phenotype data measure the normalized blood cholesterol levels. Given these data, we want to find the quantitative trait loci (QTLs) that contribute to elevated cholesterol level. The genotype and phenotype data can be downloaded from `https://sites.google.com/a/cs.washington.edu/genome560-spr18/homework`.

We will perform a single predictor regression. This means that we will model the phenotype based on the linear regression model with only one marker . For each genetic marker j, we model the measurement of blood cholesterol level Y as a linear combination of an intercept term and the genotype value X_j: $Y = \beta_0 + \beta X_j$.

(a) **[20 points]** Use "lsfit" or "lm" (Method II in R exercise of Lecture 15) to compute the β values. Implement a function that computes for each marker i, the squared error of prediction of Y by using the β values from a linear regression fit. What is the minimum squared error (SSE) across all 1333 genetic markers? What is the marker that has the minimum squared error?

(b) **[10 points]** What is the maximum SSE across all 1333 markers? What is the marker that has the maximum squared error?

(c) **[15 points]** Compute the coefficient of determination (R^2) for each of 1333 markers. Plot the R^2 values (y-axis) across all 1333 markers (x-axis).

(d) **[10 points]** What is the R^2 value of the marker you found in (a)? What is the R^2 value of the marker you found in (b)?

(e) **[15 points]** Generate a scatter plot of Y and X (e.g., slides 19-22 in Lecture 15) and add a fitted linear regression line for each of the markers you found in (a) and (b). Descibe how they are different in terms of goodness of fit.

4. **[30 points] Multi-marker model for QTL analysis**

Let's apply the regularized linear regression methods to the data set used in Q3. Denoting the cholesterol level of mouse i by Y_i, we model it as a linear combination of genotype values on the genetic markers: $Y_i = \mu_0 + \Sigma_j \beta_j X_{ij} + \epsilon$. To fit the linear model, we are going to use L2 (ridge) regularization. As we discussed in class, ridge generally learns non-zero values for all of the weights, but keeps the total L2 norm of the weight vector small. In this exercise, we will measure the test set error for varying lambda values.

(a) **[15 points]** Fit a ridge regression and compute the sum of the squared errors (SSE) (y-axis) for varying tuning parameter λ (between 1 and 20) (x-axis). What is the λ value that minimizes SSE?

(b) **[5 points]** Explain why the λ value you found in (a) minimizes SSE.

(c) **[10 points]** What is the R^2 for this λ value. Is this larger or smaller than the R^2 you computed in Q1 (d)? Explain why.