CSE 527, Fall 2017
Problem Set #2

(Due Nov 13th 11:59pm)

1. [**25 points**] **Quantitative trait loci analysis for cholesterol levels**

   We are given the genotype data from 334 mouse individuals produced by the backcross experiment. The genotype data measure the 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/cse527-au17/datasets`.

   (a) [**10 points**] Implement a function that computes the LOD score for each marker. The formula for calculating the LOD score is shown on page 10 in lecture 7 and the probability density function of the Gaussian distribution (see `http://en.wikipedia.org/wiki/Normal_distribution`). What is the maximum LOD score across all genetic markers?

   (b) [**10 points**] Let the LOD threshold be the 95th percentile of the distribution of genomewide max LOD when there is no QTL anywhere. Compute the LOD threshold by the permutation tests. Show the distribution of the max LOD scores.

   (c) [**5 points**] List the genetic markers that have a score greater than the threshold.

2. [**40 points**] **Multi-marker model for QTL analysis**

   Let's apply the regularized linear regression methods to the data set used in Q1. Denoting the cholesterol level of mouse i by Y_i, we model it as a linear combination of genotype values on the genetic markers: $Y_i = \mu_0 + \Sigma_j w_j X_{ij} + \epsilon$. To fit a linear model, we are going to use two different regularization penalties: L1 (lasso) and L2 (ridge). Lasso penalty is a sparsity-inducing penalty that shrinks most of the weights to zero. Ridge generally learns non-zero values for all of the weights, but keeps the total L2 norm of the weight vector small.

   You are encouraged to use existing packages for L1/L2-regularized linear regression available for the programming language of your choice. For R, please use glmnet package (you can get it through CRAN). For MatLab, you can use functions from Statistics Toolbox, or try SLEP (`http://www.public.asu.edu/~jye02/Software/SLEP/`). For Python, please use scikit-learn (`http://scikit-learn.org/stable/`)

   For this problem, we will split the 334 samples in the data set used in Q1 into training set and test set. Please use the first 250 samples as training set and the rest of the samples as test set. You should normalize all predictor variables (i.e., genetic markers) to have zero mean and variance one before learning or predicting. The normalized genotype can be computed as $X_{ij} = \frac{X_{ij}-\mu_j}{\sigma_j}$, where μ_j (σ_j) means the average (standard deviation) of the 334 genotypes for the j-th marker.

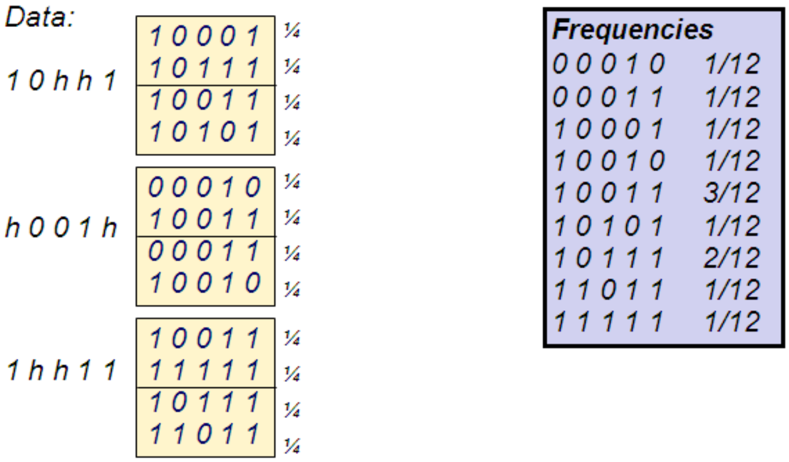
   (a) [**12 points**] Fit the Lasso regression to the training set and compute the mean squared error (MSE) for the predictions that your model makes on the test set. Generate a plot of MSE (y-axis) for varying regularization parameter C (x-axis). Find a range of regularization parameters that give you a U-shaped graph.

(b) **[12 points]** Now repeat what you did in (a) for a L2 regularized linear regression model. Are the MSEs from L2 regularization generally higher or lower than those from L1 regularization?

(c) **[16 points]** Run leave-one-out cross validation (LOOCV) on the training set to estimate the best regularization parameters for L1 and L2 regressions. Compute the test set MSE as you did in (a) and (b) using this chosen parameters. Which of L1 and L2 methods does better? How far is the MSE using the regularization parameters chosen by LOOCV from the minimal test MSE from the graphs you generate in (a) and (b)?

3. **[35 points] Implementation of the EM-based haplotype reconstruction algorithm**

Let's consider the following toy example of a haplotype reconstruction problem. You are given the genotype data on 5 markers from 3 individuals: ({10hh1}, {h001h}, {1hh11}). Given the initial haplotype frequencies listed below, we want to reconstruct haplotypes in this population.



Data:

1 0 h h 1	1 0 0 0 1	¼
	1 0 1 1 1	¼
	1 0 0 1 1	¼
	1 0 1 0 1	¼
h 0 0 1 h	0 0 0 1 0	¼
	1 0 0 1 1	¼
	0 0 0 1 1	¼
	1 0 0 1 0	¼
1 h h 1 1	1 0 0 1 1	¼
	1 1 1 1 1	¼
	1 0 1 1 1	¼
	1 1 0 1 1	¼

Frequencies	
0 0 0 1 0	1/12
0 0 0 1 1	1/12
1 0 0 0 1	1/12
1 0 0 1 0	1/12
1 0 0 1 1	3/12
1 0 1 0 1	1/12
1 0 1 1 1	2/12
1 1 0 1 1	1/12
1 1 1 1 1	1/12

(a) **[5 points]** As discussed in class, this problem can be solved by the expectation-maximization (EM) algorithm. This means that we need to estimate the parameters and hidden variables. Explain what the hidden variables and parameters are in this problem.

(b) **[10 points]** Given the haplotype frequencies listed above, explain how the next E-step works. Implement the E-step and write down the results.

(c) **[10 points]** Given the result of the E-step you obtained in part (b), explain how the M-step works. Implement the M-step and write down the results.

(d) **[10 points]** Run the EM algorithm you just implemented on these data, and show the estimated hidden variables and parameters, after convergence.