CS 124/224 Final Exam June 12, 2017

Name & UID : _____

Question	Points	Score
1	10	
2	30	
3	15	
4	20	
5	10	
Total:	85	

1. (10 points) Use Clark's algorithm to phase the following 4 genotypes. Show your work.

```
11111
00111
11000
22000
```

G h
11111 11000
00111
11000
-22000-

G h
-11111- 11000
00111 00111
11000
-22000-

G h
-11111- 11000
-00111- 00111
-11000- 00000
-22000-

1. Add haplotypes that must be present (22000 => 11000)

2. Remove genotypes that are already explained by the set of haplotypes already present

3. Add one or more haplotypes to the output set, explaining the genotypes in order (the input data order) Choose arbitrarily if multiple options

4

2. You have a population with the following 3 genotypes:

```
10122
21222
00011
```

(a) (10 points) For each genotype, write out the possible pairs of haplotypes that could have generated it.

10122: (10111, 00011);
(10011, 00111)

21222: (11111, 10111)

00011: (00000, 00011),
(00010, 00001)

(b) (10 points) How would we set this problem up to solve using expectation-maximization? In particular, what is our input data, D, the parameters which we are trying to solve for, θ_t, and the latent variables Z.

Input Data (D): Genotypes

Parameters we want (Theta_t): Population haplotype frequency/probability

Latent Variables (Z): Frequency/Probability of each pair of haplotypes that can generate one of the genotypes in D

(c) (10 points) Start by assuming that all haplotypes are equally likely; do one complete E and M step of this algorithm, and find a better prediction about the distribution of haplotypes.

	Pairs	Z
10122	10111, 00011	0.5
	10011, 00111	0.5
21222	10111, 11111	1
00011	00000, 00011	0.5
	00010, 00001	0.5

h	pvs;	theta_t
00000	1/8	0.5/6 = 1/12
00001	1/8	0.5/6 = 1/12
00010	1/8	1/12
00011	1/8	(.5 + .5)/6
00111	1/8	1/12
10011	1/8	1/12
10111	1/8	(1+.5)/6
11111	1/8	1/6

Z_{h_i,h_j} = P(h_i)*P(h_j) / Sum (P(h_i)*P(h_j) for all h_i, h_j such that h_i + h_j = g

3. You are given a standardized matrix of genotypes X, and a phenotype vector Y.

   (a) (5 points) Write an equation for the effect of **all** SNPs on the phenotype Y, assuming that the error $\epsilon \sim N(0, \sigma^2 I)$.

$$Y = \mu + X\beta + \epsilon$$

$$Y = m*1 + Xb + e$$

(Y = n x 1; X = n x p; b = p x 1, e = n x 1)

$$Y = m + \text{Sum}(X_i\, b_i) + e$$

   (b) (10 points) Derive the equation for a test on a single SNP by combining the effects of all other SNPs in to a single term.

$$Y = m + X_k * b_k + \text{Sum}(i =/= k)\ \{X_i * b_i\} + e$$

$$Y = m + X_k * b_k + g + e$$

Assume that most of the true effects are small

$$\text{Var}(g) = \text{Var}(bX) = (bX)^T bX = X^T b^T b X =$$

$$s^2_g XX^T = s^2_g K$$

$$Y = m + b_k * X_k + g + e$$

$$Y \sim N(m + b_k X_k,\ s^2_e I + s^2_g K)$$

$$V = s^2_e I + s^2_g K; \text{ compute } V^{(-1/2)}$$

$$V^{(-1/2)}Y \sim N(V^{(-1/2)}[m + b_k X_k],\ V^{(-1/2)}(V)V^{(-1/2)})$$

$$V^{(-1/2)}Y \sim N(V^{(-1/2)}[m + b_k X_k],\ I)$$

Because the matrix has variance equal to the identity, standard regression can be used

4. Naive Bayes

(a) (10 points) Use Bayes' rule to write out an equation for $P(\theta|D)$

P(thetaID) = P(DItheta)P(theta)/P(D)

P(thetaID) = posterior
P(DItheta) = likelihood (of the data given the params)
P(theta) = prior probability
P(D) = probability of dataset [hard to compute]
= Sum(theta) P(DItheta) P(theta)/sum(P(theta)

Understand connection between EM slides and Bayes' Law

(b) (10 points) Assume you have a null hypothesis of the form $H_0 : Y \sim \mathcal{N}(0, \sigma^2)$ and $H_1 : Y \sim \mathcal{N}(\mu, \sigma^2)$. What is the equation for the Bayes factor comparing H_0 and H_1?

$$\frac{P(H\_1 \mid D)}{P(H\_0 \mid D)} = \frac{\text{Product}(N(\mu - x\_i, s^2))}{\text{Product}(N(x\_i, s^2))}$$

5. Principal Components Analysis

   (a) (10 points) What is the relationship between the principal components of a matrix X and eigenvalues of $X^T X$?

The principal components of X are the eigenvectors corresponding to the eigenvalues of X X^T.

The 1st PC is the eigenvector corresponding to the largest eigenvalue of XX^T, and it is the 1 dimension about which the variance of X is the largest