Problem Set 1

EN 600.438/638

February 14, 2019

Due date: February 20, 2019 by midnight
Submission: Please submit your write-up pdf files and codes to gradescope.
Reminders: You may discuss problems in small groups but must complete your own write-up and code. You may not copy any of your work or code from others, including but not limited to any resources you may find on the internet. List ALL people with whom you collaborated on your submission. You have 5 total late days for the semester, and can use a maximum of 3 on this assignment. Days will be rounded up; for example, if you submit your assignment 3 hours late, you will use 1 full late day. If you submit your assignment 25 hours after the deadline, you will use 2 full late days.
Programming component: You must submit all source code in addition to answering the questions below (in the template). Instructors should be able to run your code, without exceptions, on gradx or ugradx **with the command specified in each question**. Please note that there will be a text file called `install_modules.txt` that provides all the packages you can use for the programming questions. And we will import only the modules from this text file to run your code. If instructors are not able to run your code you will not receive full credit for the problem in question. If you have questions or trouble setting up your code this way, please get in touch with one of the instructors.

Analytical Section - 45 points

Exercise 1. Central Dogma (9 points)
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.

TACACGTTAGACAT**GCTACG**CTGGCAAC**GGGTACATC**

(a) Provide the RNA sequence that would result in transcription of this region of DNA, and the amino acid sequence that would then result from translation. (2 points)

(b) Starting with the sequence above, provide modified DNA sequences along with corresponding amino acid sequences reflecting (4 points)

i a frameshift insertion or deletion

ii a non-frameshift insertion or deletion

iii a synonymous single nucleotide change and

iv a non-synonymous single nucleotide change

(c) A defect in the human beta globin gene (Gene symbol: HBB) causes sickle cell anemia. Using UCSC genome browser online (GRCh38/hg38 build), locate SNP rs334 on the beta globin gene and answer the following: (3 points)

i On what chromosome and at what approximate position is HBB located?

ii What region of the gene is SNP rs334 found in? (a) Intronic or (b) Exonic

iii What type of mutation does SNP rs334 cause? (a) non-synonymous (b) synonymous or (c) frameshift

Exercise 2. Genetic variation in the human genome (6 points)
Answer the following questions with "True" or "False"

(a) A SNP located in a non-coding region can be associated with complex diseases.

(b) If you know the genotype on the maternal allele at a particular genomic coordinate, you also know the genotype on the paternal allele at that genomic coordinate.

(c) A genotype array is useful experiment to identify rare genetic variants.

(d) Genome wide association studies (GWAS) can definitely establish causal relationships between genetic variants and complex traits.

(e) SNP A and SNP B are in high linkage disequilibrium. If you know the genotype of SNP A for a particular individual, you can infer the genotype of SNP B for that individual confidently.

(f) Linear regression is a suitable model to test the association between a SNP and height.

Exercise 3. Modeling genotype effects (15 points)

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.

(a) Provide three possible outcomes of heterozygous genotype (AT) at SNP X_1 on petal color. For each, define a data encoding and linear model you could use to capture the effect of genotype on petal color. Give a brief description of the model and justify. (7 points)

(b) We have also find that the genotype on X_1 has an effect on the flower's ability to endure low temperature. Flowers with AA on X_1 can live in super cold enrivonment, while flowers with TT cannot endure low temperature. Qualitatively compare the proportion of red flowers in Minnesota and the proportion of red flowers in Florida, and explain why. (5 points)

(c) Suppose another SNP X_2 also has an effect on petal color, and that there is a nonlinear (i.e. non-additive) interaction between SNP X_1 and SNP X_2 which affects petal color. How might we encode an interaction effect in a linear model ? (3 points)

Exercise 4. Maximum Likelihood Estimation (15 points)

Scientists performed RNA-sequencing on 10 individuals and used this data to investigate expression levels of the SOX2 gene and the TNNT2 gene. The normalized expression levels of both genes in all 10 individuals is summarized in the table below:

Individual #	SOX2 expression	TNNT2 expression
1	10.9	15.6
2	7.6	15.1
3	11.7	19.4
4	11.8	17.1
5	12.5	17.2
6	8.7	15.5
7	10.8	17.6
8	10.3	17.6
9	10.9	14.4
10	9.3	14.6

You can assume that SOX2 expression and TNNT2 expression follow a bivariate normal distribution.

(a) To begin to understand the relationship between SOX2 and TNNT2 in these individuals, compute the pearson correlation between SOX2 expression and TNNT2 expression. Report the correlation coefficient and the pvalue (two-tailed). (5 points)

(b) Compute the maximum likelihood estimates (MLE) of parameters defining the bivariate normal distribution (ie return the mean vector and covariance matrix using the observed data). (5 points)

(c) Another individual has a SOX2 expression level of 10.2. Using your MLE estimates from part b, what is that individual's expected TNNT2 expression level? (5 points)

Programming Section - 55 points

All data for the assignment have been posted as a zipped folder on piazza - ***problem_set_1.zip***. This folder contains 3 sub-directories, exercise 5, exercise 6 and exercise 7. Each of these have the corresponding files you will need to complete the programming section of the assignment.

Exercise 5. Logistic regression (15 points)
Data: Please download data for exercise 5 from Piazza. There are two gene expression

files, (*train_expression.csv*) and (*test_expression.csv*). The rows of these matrices represents gene expression measurements for individuals in each column. The phenotype files (*phen_train.csv*) and (*phen_test.csv*) include simulated meta-data information for each individual represented belonging to group 0 and group 1. In this case, the groups represent indviduals diagnosed with luminal or basal subtypes of breast cancer, respectively.
Code: Please name the code exercise5.py. We should be able to run your code as:

```
python3 exercise5.py --X train_expression.csv --Y train_phen.csv
--testX test_expression.csv --testY test_phen.csv
```

(a) Use logistic regression to build a a model to predict group 0 and group 1 breast cancer patients using all genes. Please use `sklearn.linear_model.LogisticRegression` and the default setting. Train logistic regression parameters using only *train_expression.csv* and *phen_train.csv*. Evaluate your model by making predictions on *test_expression.csv* and evaluating those predictions with *phen_test.csv*. Report the precision and recall on the test data using a probability threshold of .5? (5 points)

(b) Repeat part a using only the first 10 genes in *train_expression.csv* and *test_expression.csv*. Again, report the precision and recall on the test data using a probability threshold of .5. (5 points).

(c) Is the performance different between the two models? What do you think is the reason? (3 points)

(d) How could you better select the 10 genes to be used in part b to increase model performance? (2 points)

Exercise 6. Genome-wide association studies (30 points)
Data: Please download data for exercise 6, which includes genotype and associated phenotype data, from piazza. This genotype file (*genotype.csv*) contains genotype information for 9088 SNPs from 279 individuals. Genotypes are encoded as 0, 1, 2 and 3, representing homozygous for one allele (AA), heterozygous (AB), homozygous for alternate allele (BB) and missing genotype respectively. The phenotype file (*phenotype.csv*) includes trait information for individuals.
Code: Please name the code exercise6.py. We should be able to run your code as:

```
python3 computeMAF.py --X genotype.csv --output Q6_a_output.txt

python3 logistic_regression.py --X genotype.csv --Y phenotype.csv
--output Q6_b_output.txt
```

(a) Write a function called `computeMAF.py` to compute the Minor Allele Frequency (MAF) of each SNP. Save MAF for all the SNPs in the output file (See the format below). How many SNPs have MAF greater than 0.03, 0.05 and 0.1? (7 points)
output file format:
`SNP1 MAF1`
`SNP2 MAF2`
...

(b) Write a function called `logistic_regression.py` to run logistic regression to test association of each SNP *independently* with the given trait, restricting to SNPs with MAF of at least 0.05. We need to adjust the p-values to account for the number of hypothesis tests performed in order to control for Type I Error. Use Benjamini Hochberg correction to control for FDR, and save the significant SNPs associated with the trait at FDR threshold of 0.05 and their uncorrected p-value to the output file. You may want to use the `statsmodels.discrete.discrete_model.Logit` module to get the p-values (Hint: use the `summary` attribute). (8 points)
output file format:
`SNP1 p-value1`
`SNP2 p-value2`
...

(c) What is the interpretation of the logistic regression parameter in terms of disease risk? (2 points)

(d) What is the regression parameter β learned for SNP 10? Compute the logistic loss function for a range of values near this parameter setting, and provide a plot demonstrating that the learned β is in fact the optimal Maximum Likelihood value. (8 points)

(e) Confounding Factors (638 level only)
Factors such as ancestry and age can have confounding effects in genome wide association studies, leading to both false positives and false negatives. How can you account for this in a linear model? Specify an appropriate null hypothesis to test for association of a SNP that would account for a confounding variable. (5 points)

Exercise 7. Regularized regression (638 only, 10 points)
Data: Please download data for exercise 7 from Piazza in the folder: *exercise7*
Code: Please name the code exercise7.py. We should be able to run your code as:

```
python3 exercise7.py --X train_expression.csv --Y phen_train.csv
--testX test_expression.csv --testY phen_test.csv
--output Q_7_output.txt
```

You would like to predict the phenotype using expression levels of all genes, and you choose to use multivariate linear regression. Because of the large number of genes, you decide to use ridge regression to help reduce the probability of over-fitting. Please use the module `sklearn.linear_model.Ridge`. In ridge regression, you need to choose the penalty parameter α. To do so, we have split the dataset into training data and test data. For each value of α, you should use the training data to learn the regression parameters β, and use the test data to evaluate the performance of the learned β, by measuring the sum of squared error. Then select the value of α that yields the smallest sum of squared error.

Please implement `exercise7.py` to apply ridge regression to the data in *exercise7*. *exercise7/train_expression.csv* and *exercise7/test_expression.csv* are training and test gene expression data, respectively. *exercise7/train_phen.csv* and *exercise7/test_phen.csv* are training and test phenotype data, respectively. Test the penalty parameter α for a range of settings

of α = [0, 0.1, 1, 10, 100]. Your code should be able to learn the optimal α in the given range. Then use the optimal α to run ridge regression on training data and save the learned parameters to output file.
output file format (assuming k genes):
β_1
β_2
...
β_k

(a) Which setting of α would you prefer? Why?