MATH 504

1. The genome of SARS-COV-2, a.k.a covid-19, is roughly 30,000 nucleotides long, meaning that it is a sequence of 30,000 A,C,G, and T's. Samples of genomes sequences from patients are being posted to NCBI (the bioinformatics wing of NIH). I downloaded 241 sequences on which this problem will be based. Attached you will find 4 files. All of the files are in csv format and do not have a header row.

   - `matrix_sequences.csv` contains a matrix with 241 rows and 29903 columns. Each row of the matrix is a SARS-COV-2 sequence pulled from a patient.
   - `matrix_countries.csv` contains a matrix with 241 rows and 1 column. Each row is the country of origin for the corresponding sequence in `matrix_sequences.csv`.
   - `matrix_reference.csv` contains a matrix with 1 row and 29903 columns. This is the so-called SARS-COV-2 reference sequence. It has been selected by the W.H.O. as the sequence to use as a reference from which to compare other sequences.
   - `ncbi_sequences_cleaned.fasta` is a text file that contains the sequenes in a format that is easier to look at. Each sequence is shown as a string of A,C,G,T preceded by a header line. This file is not part of this problem, but I'm including it in case you want to see the sequences in a more convenient format than a matrix.

     (Biological Note: I altered the data to make all the sequences align to 29903 columns. In the process some important variation was lost, in particular insertions and deletions.)

   Convert the matrix in `matrix_sequences.csv` to a matrix of 0's and 1's by comparing each sequence to the reference. If a sequence has the same letter at a given position as the reference, set the matrix entry to 1, otherwise set it to 0. Perform 2-d PCA on the resultant $0, 1$ matrix. (Don't forget to center!). The covariance matrix will be 29903 by 29903. **That's too big to use eigen and too big to form in reasonable space and time!** Instead, apply a power iteration approach and compute $X^T X v$ for a vector v without forming $X^T X$. Determine the fraction of the variance captured by the 2-d PCA. (Hint: You don't need all the eigenvalues of the covariance matrix to know the total variance.) Plot the projection of the sequences and color the points according to the country. Comment on the implications for how the virus spread.

2. In the lecture, I discussed kernel machines in the context of linear regression. Here, you will develop the same ideas in the context of logistic regression.

   (a) Suppose that we define features through the mapping $\phi(x) : \mathbb{R}^n \to \mathbb{R}^m$ and that we assume the logistic regression model (see homeworks 4 and 7),

   $$P(y = 1 \mid w, \phi(x)) = \frac{1}{1 + \exp[-w^T \phi(x)]}. \tag{1}$$

   What is the dimension of w in this model? Show that the log-likelihood under this model is given by,

   $$\log L(w) = \sum_{i=1}^{N} (1 - y_i)(-w^T \phi(x^{(i)})) - \log(1 + \exp(-w^T \phi(x^{(i)}))). \tag{2}$$

(b) Let B be the model matrix for the features of the sampled data, i.e. the ith row of B is $\phi(x^{(i)})^T$.

i. Show that the w that maximizes the log-likelihood, w^*, is in the span of the $\phi(x^{(i)})$ for $i = 1, 2, \ldots, N$ (where N is the number of samples), and that we can therefore express w^* as $w^* = B^T a^*$ for some $a^* \in \mathbb{R}^N$.

ii. Show that we can solve for a^* by solving the optimization

$$\max_{a \in \mathbb{R}^N} \sum_{i=1}^{N} (1 - y_i)(-a^T k^{(i)}) - \log(1 + \exp(-a^T k^{(i)})) \tag{3}$$

where K is a $N \times N$ kernel matrix defined by $K_{j\ell} = \phi(x^{(j)}) \cdot \phi(x^{(\ell)})$ and $k^{(i)}$ is the ith column of K.

iii. Suppose we have fitted our model by finding a^* and are given an $x \in \mathbb{R}^n$. Show that under our fitted model,

$$P(y = 1 \mid a^*, \phi(x)) = \frac{1}{1 + \exp[-(a^*)^T \tilde{k})]}. \tag{4}$$

where $\tilde{k}$ is a N-dimensional vector with $\tilde{k}_i = \phi(x^{(i)})^T \phi(x)$.

(c) Part (b) shows that we can replace (1) and (2) by (4) and (3). Under what circumstances would this replacement be advantageous?

(d) Consider now a penalized logistic regression in which we fit the model (1) by maximizing the following penalized log-likelihood for some $\lambda > 0$.

$$\log L(w) = \sum_{i=1}^{N} (1 - y_i)(-w^T \phi(x^{(i)})) - \log(1 + \exp(-w^T \phi(x^{(i)})) - \lambda \|w\|^2. \tag{5}$$

i. Show that this function is concave. (You may use results from hw 7.)

ii. Show that the penalized form of (3) is given by

$$\sum_{i=1}^{N} (1 - y_i)(-a^T k^{(i)}) - \log(1 + \exp(-a^T k^{(i)})) - \lambda a^T K a. \tag{6}$$

3. Consider the dataset from homework 13, `nn.txt`. In this problem we will fit the data using a kernel machine, as opposed to the neural net of homework 13. Recall that the data is composed of samples $x^{(i)} \in \mathbb{R}^2$ and a response $y \in \{0, 1\}$.

(a) For $x \in \mathbb{R}^2$, define

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1 x_2, x_2^2). \tag{7}$$

$\phi(x)$ contains all monomials of degree less than or equal 2 formed by the two coordinates of x. Show that if $x, x' \in \mathbb{R}^2$ then,

$$\phi(x) \cdot \phi(x') = (1 + x \cdot x')^2. \tag{8}$$

(b) Consider the kernel $K(x, x') = (1 + x \cdot x')^3$. What are the feature vectors for this kernel?

(c) Using the feature vector from part (a), fit the penalized logistic model (5) for different λ using a Newton's method approach. However, don't fit the model by directly optimizing w, instead use (6) and compute the optimal a. To make the optimization a bit easier, base your fit on 500 of the samples. (In this toy model, the dimension of w is only 6, while the dimension of a will be N, the number of samples, which we will set to 500. In this context we would optimize for w, but here I want you to go through the process of fitting for a in a setting where you can visualize the fit. If $x \in \mathbb{R}^n$ for $n \gg 2$, then the dimension of w would exceed that of a.)

(d) As in part (g) of homework 13, pick a cutoff p and visualize the corresponding classifiers for the different λ. Use (4) to make predictions. At no point should you compute w. Given that the features are quadratic functions of x, explain the shape of the region for which your classifier predicts $y = 1$.

(e) What do you think will happen if you use the kernel $K(x, x') = (1 + x \cdot x')^3$? Don't actually do the fitting, just discuss how this kernel might improve on the results in (d).