Stats305a Problem Set 1
Due: Friday, October 8 by 11:59pm on Gradescope.

Question 1.1 (Block matrix inversion, linear algebra, and the Sherman-Morrison-Woodbury formula): In this question, we will work out a few formulae for inverting block-structured matrices, using them to develop various inversion formulas for structured matrices.

(a) Consider the matrix equation
\[
A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}
\]
where \(A\) and \(D\) are square matrices and \(B, C\) have appropriate sizes (this is not important for this question). Give a formula for \(x\) in terms of \(A, B, C, D, a,\) and \(b\); your formula, if correct, will involve inverses of some of these. You may assume that \(A\) and \(D\) are invertible and that \(A - BD^{-1}C\) is invertible. (Note: \(B\) and \(C\) may be rectangular, so don’t try to invert them alone.)

(b) We now consider inverting a matrix plus a (typically) low rank matrix. We wish to solve \((A + UCV^T)x = z\) for \(x\), where \(A \in \mathbb{R}^{n \times n}, U \in \mathbb{R}^{n \times k}, C \in \mathbb{R}^{k \times k},\) and \(V \in \mathbb{R}^{n \times k}\). Introducing the variable \(y = CV^T x\), or \(V^T x - C^{-1} y = 0\), solve
\[
\begin{bmatrix} A & U \\ -V^T & C^{-1} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} z \\ 0 \end{bmatrix}
\]
for \(x\). Using this, show that
\[
x = (A^{-1} - A^{-1} U (C^{-1} + V^T A^{-1} U)^{-1} V^T A^{-1}) z,
\]
i.e.,
\[
(A + UCV^T)^{-1} = A^{-1} - A^{-1} U (C^{-1} + V^T A^{-1} U)^{-1} V^T A^{-1}.
\]
(As an aside, if you know \(A^{-1}\) and \(C^{-1}\) already, the largest matrix you must invert to compute \((A + UCV^T)^{-1}\) is then \(k \times k\), which is smaller.)

Question 1.2 (A rank-one update to a linear regression solution): Consider a least-squares problem with data \(X \in \mathbb{R}^{n \times d}, Y \in \mathbb{R}^{n}\), where \(n \geq d\) and \(X\) has rank \(d\), and let
\[
\hat{\beta}_n = \arg\min_{\beta} \|X \beta - Y\|^2.
\]
Assume we get a new observation \((x_{n+1}, y_{n+1}) \in \mathbb{R}^d \times \mathbb{R}\) and wish to update
\[
\hat{\beta}_{n+1} = \arg\min_{\beta} \left\{ \|X \beta - Y\|^2 + (x_{n+1}^T \beta - y_{n+1})^2 \right\}.
\]
Give a formula for \(\hat{\beta}_{n+1}\) in terms of \(\hat{\beta}_n, (X^T X)^{-1}, x_{n+1},\) and \(y_{n+1}\).

Assuming you have already computed \(H = (X^T X)^{-1} \in \mathbb{R}^{d \times d}\) and that multiplying \(H\) by a vector, i.e., computing \(Hv\), takes time \(d^2\), roughly how much time does computing your update take? (Note: you can simply say “A few multiples of \(d\),” or “A few multiples of \(d^2\),” or “A few multiples of \(d^3\),” depending on which is true.)
Question 1.3 (The most negatively correlated distribution): A financial analyst tells you that he has a great stock tip that will allow you to short stocks based on others that are doing well. He assures you that he has found a correlation matrix between the \( n \) stocks, \( C \in \mathbb{R}^{n \times n} \), with entries
\[
C_{ij} = \begin{cases} 
1 & \text{if } i = j \\
-1/2 & \text{otherwise}.
\end{cases}
\]
(a) Is his correlation matrix possible? Would you trust him with your graduate stipend?

(b) More generally, consider a correlation matrix \( C_{\rho} \) with entries of the following form:
\[
[C_{\rho}]_{ij} = \begin{cases} 
1 & \text{if } i = j \\
-\rho & \text{otherwise}
\end{cases}
\]
where \( \rho \geq 0 \). How large can \( \rho \) be while \( C_{\rho} \) remains a potentially valid correlation matrix? Hint: A correlation matrix for a random vector \( X \in \mathbb{R}^n \) has entries \( C_{ij} = \text{Cov}(X_i, X_j)/\sqrt{\text{Var}(X_i)\text{Var}(X_j)} \), and so may be written
\[
C = \text{diag}(v)^{-1/2}\text{Cov}(X)\text{diag}(v)^{-1/2},
\]
where \( v \) is a vector with entries \( v_i = \text{Var}(X_i) \) and \( \text{diag}(v) \) is the diagonal matrix with diagonal \( v \).

Question 1.4 (Some basic plotting and data processing): The UCI Machine Learning repository has a collection of useful datasets for experiments and data exploration. This is a question that simply serves as a forcing function for you to pick a computer language, read in data, and plot it appropriately. Using the data in the UCI Wine quality dataset (https://archive.ics.uci.edu/ml/datasets/Wine+Quality, and see the winequality-red.csv file in the data folder there), plot a scatterplot matrix showing the five variables density, alcohol, pH, volatile.acidity, and the target variable quality, which is a measure of wine quality. Such pairwise scatterplots can be a useful tool for data exploration and summarization (see, e.g., Fig. 1.1 of [1]).

In your plots, what do you notice about density, alcohol, and quality? (Just a sentence suffices here.)

Question 1.5 (Predicting high temperatures at SFO): In this question, we use linear regression to predict the high temperature at San Francisco International Airport (SFO). A natural model of the temperature is the following: let \( x \) be the day of the year (that is, \( x = 1 \) corresponds to January 1, while \( x = 365 \) corresponds to December 31, except in leap years); we assume that the temperature
\[
y = \beta_0 + \beta_1 \sin\left(\frac{2\pi}{365}(x - 1)\right) + \beta_2 \cos\left(\frac{2\pi}{365}(x - 1)\right).
\]
(Admittedly this ignores the issue of leap years, but we will punt on that.) Let \( \phi(x) = [1 \sin\left(\frac{2\pi}{365}(x - 1)\right) \cos\left(\frac{2\pi}{365}(x - 1)\right)]^T \) be the vector feature representation above.

The data file simplified-sfo-weather.csv contains weather data for SFO since 1960, including precipitation (in inches), low, and high temperatures (in degrees Fahrenheit). Note that in May 2018, the temperature sensor broke and consequently a few days report NA as the high and low temperatures. You should simply omit those from any averages or model
fitting.\footnote{In R, you may do this automatically in the \texttt{lm} methods using the keyword \texttt{na.action = na.omit}, and in computing a \texttt{mean} using \texttt{na.rm = TRUE}.} Using this data, fit the model (1.1) to predict high temperature (this is column "temphigh" in the file) from the date $x$ of the year for years prior to 1990. For each decade 1961–1970, 1971–1980, 1981–1990, 1991–2000, 2001–2010, and 2011–2020, print the mean of the actual high temperature minus the predicted high temperature for the decade. That is, if $\widehat{\beta} = [\widehat{\beta}_0 \widehat{\beta}_1 \widehat{\beta}_2]^T$ denotes your fit model, compute the average difference
\begin{equation*}
y - \widehat{y} = y - \phi(x)^T \widehat{\beta}
\end{equation*}
over each of those decades. Include your code and a printout of the results.

References