Problem Set 2

Due February 13

Assignment:

1. Undergraduates and graduate students (non-math): Do all four problems.
2. Math graduate students: Do all four problems.

You are encouraged to work other students but turn in your own paper. Please—no electronic copies.

1. Consider sets $A_1, A_2, \ldots, A_n$ and the problem of determining the similarities between sets A_i and A_j for each possible pairing of sets. This is a common data science problem, and, as an example, suppose that each set identifies the shows that may be streamed by Netflix, Amazon, etc. over the course of a particular week.

   (a) We may use Jaccard similarity to compare two sets. Jaccard similarity (Chap 2.7 in *Algorithms for Data Science*) may be defined as follows:

   The Jaccard similarity between sets A and B is the number of elements in both A and B relative to the number of elements in either A and B. Let $|A|$ denote the cardinality of the set A. Then, the Jaccard similarity between A and B is

   $$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{1}$$

   Jaccard similarity possesses several desirable attributes:

   i. If the sets are the same then the Jaccard similarity is 1. Mathematically, if $A = B$, then $A \cap B = A \cup B$ and $J(A,B) = 1$.
   ii. If the sets have no elements in common, then $A \cap B = \emptyset$ and $J(A,B) = 0$.
   iii. $J(A,B)$ is bounded by 0 and 1 because $0 \leq |A \cap B| \leq |A \cup B|$.

   Now, let $\mathbf{x}$ be a vector of length p where p is the number of unique items across all sets. The ith element of $\mathbf{x}$ is

   $$x_j = \begin{cases} 1, \text{ if the } j\text{th item is in set } A, \\ 0, \text{ if the } j\text{th item is not in set } A. \end{cases} \tag{2}$$

   Thus, for each set $\mathbf{A}_i$ there is a vector $\mathbf{x}_i$ that identifies the items in the set. Write the Jaccard similarity between sets A_i and A_j as a function (only) of $\mathbf{a}_i$ and $\mathbf{a}_j$.

(b) An alternative measure of similarity is the empirical conditional probability given by

$$\Pr(A_i|A_j) = \frac{|A_i \cap A_j|}{|A_j|}. \tag{3}$$

Write $\Pr(A_i|A_j)$ as a function of $\mathbf{a}_i$ and $\mathbf{a}_j$. Show that each of the properties listed above for Jaccard similarity do or (do not) hold for $\Pr(A_i|A_j)$. If a matrix of empirical conditional probabilities is constructed, will it be symmetric?

2. A *projection* matrix maps vectors onto a subspace spanned by the columns of a matrix. For instance, the projection matrix associated with the n-length vector $\mathbf{j}$ (consisting of all ones) is $\underset{n\times n}{\mathbf{P}} = \mathbf{j}\mathbf{j}^T/\mathbf{j}^T\mathbf{j}$.

   (a) Suppose that $\mathbf{X}$ is $n \times p$ and full rank. The projection matrix for the space spanned by the columns of $\mathbf{X}$ is $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Show that the projection of $\mathbf{P}$ onto the column space spanned by the columns of $\mathbf{X}$ is $\mathbf{P}$. (In other words, show that $\mathbf{P}^2 = \mathbf{P}$.)

   (b) A matrix that satisfies $\mathbf{P}^2 = \mathbf{P}$ is called *itempotent*. Show that $\mathbf{I} - \mathbf{P}$ is itempotent.

   (c) Prove that if $\mathbf{P}$ has an inverse $\mathbf{P}^{-1}$, then $\mathbf{P}$ is the identity matrix. Comment: any other projection matrix besides the identity is *not* full rank.

   (d) Suppose that $\mathbf{x}$ is a n-vector. Show that $\mathbf{u} = \mathbf{P}\mathbf{x}$ and $\mathbf{v} = (\mathbf{I} - \mathbf{P})\mathbf{x}$ are orthogonal and that $\mathbf{x} = \mathbf{u} + \mathbf{v}$. Comment: this result implies that $\mathbf{P}$ creates two orthogonal subspaces of $\mathbb{R}^n$.

   (e) Show that the vector of fitted values $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ produced by the linear regression prediction function is the projection of $\mathbf{y}$ onto the column space spanned by the columns of $\mathbf{X}$.

3. To reduce the effect of over-fitting in training a prediction function, the objective function is sometime augmented with a *penalty* function. In regression, this process is sometimes called regularization. One regularized objective function is

$$\varepsilon_R(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}, \tag{4}$$

where $\lambda \geq 0$ is a tuning constant usually determined by trial and error.

   (a) Differentiate $\varepsilon_R(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ (compute $\partial\varepsilon_R/\partial\boldsymbol{\beta}$).

   (b) Minimize $\varepsilon_R(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. (Solve $\mathbf{0} = \partial\varepsilon_R/\partial\boldsymbol{\beta}$ for $\boldsymbol{\beta}$). Denote the solution by $\widehat{\boldsymbol{\beta}}_R$.

   (c) Compare $\widehat{\boldsymbol{\beta}}_R$ to $\widehat{\boldsymbol{\beta}}$. In other words, how will they differ?

   (d) Find a small data set, carry out regularized regression over a range of values for λ. (Compute the parameter vectors using matrix operations rather than using an built-in function.) Graph $\varepsilon_R(\widehat{\boldsymbol{\beta}}_R)$. What is a good choice for λ according to your graph (turn in the graph).

4. (From J. Gentle, *Matrix Algebra*). In $\mathbb{R}^2$ with a Cartesian coordinate system, the diagonal directed line segment through the positive quadrant (orthant) makes a 45 angle with each of the positive axes. In 3 dimensions, what is the angle between the diagonal and each of the positive axes? In 10 dimensions? In 100 dimensions? In 1000 dimensions? We see that in higher dimensions any two lines are almost orthogonal. (That is, the angle between them approaches 90.) What are some of the implications of this for data analysis?

5. Consider the problem of assigning group membership to an object with a predictor vector $\mathbf{x}$ given the possible classes A, B, and C. For each class, there is an exemplar vector $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. Normalize the vectors and determine the best prediction of group using the cosine/angle between the vectors

$$\begin{aligned} \mathbf{a} &= [.1 \quad .2 \quad .1] \\ \mathbf{b} &= [0 \quad .6 \quad .6] \\ \mathbf{c} &= [.4 \quad .2 \quad .1] \\ \mathbf{x} &= [2 \quad 3 \quad 0] \end{aligned} \tag{5}$$

   as the measure of similarity to group.