1. Suppose the n-vector w is the word count vector associated with a document and a dictionary of n words. For simplicity assume that all words in the document appear in the dictionary.

   (a) What is $\mathbf{1}'w$?

   (b) What does $w_{282} = 0$ mean?

   (c) Let h be the n-vector that gives the histogram of the word counts, i.e. h_i is the fraction of the words in the document that are word i. Use vector notation to express h in terms of w.

   (d) Execute Listing 1.8 from [SWJ] (Web interface, JSON and string parsing).

   (e) Modify the code so that it find the 20 most popular words in the text.

2. Suppose that a and b are vectors of the same size. The triangle inequality states that,

$$||a + b|| \le ||a|| + ||b||.$$

   Show that we also have,

$$||a + b|| \ge ||a|| - ||b||.$$

3. *Exploring an auto-regressive model*: Suppose that $z_1, z_2, \dots$ is a time series, with the number z_t giving the value at time t. For example z_t could be the gross sales at a particular store value on day t. An *auto-regressive* (AR) model is used to predict z_{t+1} from the previous M values, $z_t, z_{t-1}, \dots, z_{t-M+1}$:

$$\hat{z}_{t+1} = (z_t, z_{t-1}, \dots, z_{t-M+1})'\beta, \qquad t = M, M+1, \dots.$$

   Here $\hat{z}_{t+1}$ denotes the AR model's prediction of z_{t+1}, M is the memory length of the AR model, and the M-vector β is the AR model coefficient vector. For this problem we will assume that the time period is daily, and $M = 10$. Thus, the AR model predicts tomorrow's value, given values over the last 10 days.

   (a) For each of the following cases, give a short interpretation of description of the AR model in English, without referring to mathematical concepts like vectors, inner products and so on. You can use words like 'yesterday' or 'today': (i) $\beta = e_1$, (ii) $\beta = 2e_1 - e_2$, (iii) $\beta = e_6$, (iv) $\beta = \frac{1}{2}e_1 + \frac{1}{2}e_2$.

   (b) Assume that you have a data set of 100 days generated via,

$$z_t = \cos\left(t\frac{2\pi}{7}\right) + 0.2\,\xi_t,$$

   where ξ_t is a standard normal random variable independent of all other variables.

   Generate 1000 independent replications of the vector z and check which of the predictors (i), (ii), (iii), or (iv) works "best" (where you define what that means). Explain your answer and present a clear description of your computation in Julia. Make sure your results are reproducible by fixing a seed.

4. Any real-valued function f that satisfies the four properties given on page 46 of [VMLS] (non-negative homogeneity, triangle inequality, non negativity and definiteness) is called a *vector norm*, and is usually written as $f(x) = ||x||_{\mathrm{mn}}$ where the subscript "mn" is some kind of identifier. The most commonly used norm is the Euclidean norm, which is sometimes written with the subscript 2 as $||x||_2$. Two other common vector norms for n-vectors are the 1-norm $||x||_1$ and the ∞-norm $||x||_\infty$, defined as,

$$||x||_1 = |x_1| + \ldots + |x_n|, \qquad ||x||_\infty = \max\{|x_1|, \dots, |x_n|\}.$$

   Verify (prove) that the 1-norm and the ∞-norm satisfy the four norm properties.

5. Suppose the n-vector c gives the coefficients of a polynomial $p(x) = c_1 + c_2 x + \ldots + c_n x^{n-1}$.

   (a) Let α and β be numbers with $\alpha < \beta$. Find an n-vector a for which,

   $$a^T c = \int_\alpha^\beta p(x)\,dx$$

   always holds. This means that the integral of a polynomial over an interval is a linear function of its coefficients.

   (b) Let α be a number. Find the n-vector b for which,

   $$b^T c = p'(\alpha).$$

   This means that the derivative of the polynomial at a given point is a linear function of its coefficients.

6. The function $\phi : \mathbb{R}^3 \to \mathbb{R}$ satisfies,

   $$\phi(1,1,0) = -1, \qquad \phi(-1,1,1) = 1, \qquad \phi(1,-1,-1) = 1.$$

   (a) Choose one of the following, and justify your choice: (i) ϕ must be linear. (ii) ϕ could be linear. (iii) ϕ cannot be linear.

   (b) Modify the question so as to obtain a different correct scenario amongst (i), (ii) and (iii).

7. Generate 100 independent data points uniformly distributed on the interval $[9, 11]$, denote these via $x = (x_1, \ldots, x_{100})$.

   (a) Prove that the scalar a that minimizes $||x - a\mathbf{1}||$ is $a = \overline{x}$, the sample mean.

   (b) Implement a gradient descent algorithm for obtaining the minimizer.

   (c) What is the range of learning rates for which the algorithm converges to the minimizer from any initial point? (Obtain your answer either analytically or via numerical experimentation).

8. Let n be an even integer and consider the function $f : \mathbb{R}^n \to \mathbb{R}^2$ with,

   $$f(x_1, \ldots, x_n) = \Big(\sum_{i \text{ odd}} x_i^2 + \sum_{i \text{ even}} x_i \;,\; \sum_{i \text{ odd}} x_i + \sum_{i \text{ even}} x_i^2 \Big).$$

   (a) Represent $f(\cdot)$ in terms of norms or inner products (defining auxiliary vectors as needed).

   (b) Determine the Jacobian of f at the point $z \in \mathbb{R}^n$.

   (c) Determine the first order tailor approximation around z, $\hat{f}(x_1, \ldots, x_n)$.

   (d) Consider the approximation at $z = (1, \ldots, 1)$ and take a ball of radius 0.5 around z. Determine (numerically or perhaps analytically) the maximal value of $||f(x) - \hat{f}(x)||$ for x in the ball.

   (e) Repeat for radius 0.25 and radius 1.0.

9. Reproduce the clustering implementation in [SWJ] Listing 9.9, "Manual Implementation of k-means". Then modify the code to handle cases where labels (clusters) for some of the data points are specified ahead of time. Test this on a few examples.

10. Suppose that each of the vectors $b_1, \ldots, b_k$ is a linear combination of the vectors $a_1, \ldots, a_m$, and c is a linear combination of $b_1, \ldots, b_k$. Then c is a linear combination of $a_1, \ldots, a_m$. Show (prove) this first for $m = k = 2$ and then in general.