Please provide complete and well-written solutions to the following exercises.

Due September 10, 9AM PST, to be uploaded as a single PDF document to blackboard (under the Assignments tab).

Homework 1

Exercise 1. Let $x^{(1)}, \dots, x^{(m)}$ be m vectors in $\mathbf{R}^n$ with $\left|\left|x^{(i)}\right|\right| = 1$ for all $1 \leq i \leq m$. Let $\varepsilon > 0$. Assume that $m > (1 + 2/\varepsilon)^n$. Show that there exists $i, j \in \{1, \dots, m\}$ such that $\left|\left|x^{(i)} - x^{(j)}\right|\right| < \varepsilon$.

Consequently, the vectors $x^{(i)}$ and $x^{(j)}$ are highly correlated, so that $\langle x^{(i)}, x^{(j)} \rangle > 1 - \varepsilon^2/2$. That is, if you have enough vectors on a unit sphere, at least two of them will be correlated with each other.

(If you want a hint, read about ε-nets in the notes.)

Exercise 2. Let A be an $m \times n$ real matrix with $m \geq n$. Show that A has rank n if and only if $A^T A$ is positive definite.

(Hint: $A^T A$ is always positive semidefinite.)

Exercise 3. Let $x^{(1)}, \dots, x^{(m)} \in \mathbf{R}^n$. Let $y \in \mathbf{R}^n$. Show that

$$\sum_{j=1}^{m} \left\| x^{(j)} - \frac{1}{m} \sum_{p=1}^{m} x^{(p)} \right\|^2 \leq \sum_{j=1}^{m} \left\| x^{(j)} - y \right\|^2.$$

That is, the barycenter is the point in $\mathbf{R}^n$ that minimizes the sum of squared distances.

Exercise 4. Let $n \geq 2$ be a positive integer. Let $x = (x_1, \dots, x_n) \in \mathbf{R}^n$. For any $x, y \in \mathbf{R}^n$, define $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$ and $||x|| := \langle x, x \rangle^{1/2}$. Let $S^{n-1} := \{x \in \mathbf{R}^n : ||x|| = 1\}$ be the sphere of radius 1 centered at the origin. Let $x \in S^{n-1}$ be fixed. Let v be a random vector that is uniformly distributed in S^{n-1}. Prove:

$$\mathbf{E} \left| \langle x, v \rangle \right| \geq \frac{1}{10\sqrt{n}}.$$

Exercise 5. Run PCA on a "planted" data set on a computer, consisting of 100 samples in $\mathbf{R}^{10}$ of the random variable $(X, Y, Z_3, \dots, Z_{10}) \in \mathbf{R}^{10}$ where X, Y are standard Gaussian random variables, Z_i is a mean i Gaussian random variable with variance 10^{-2}, for all $3 \leq i \leq 10$, and $X, Y, Z_3, \dots, Z_{10}$ are all independent. (You can use your favorite computer program to simulate the random variables.)

Then, run PCA on Airline Safety Information, and try to find out something interesting (this part of the question is intentionally open ended). The data is here, with accompanying article here. (See also here.)

Exercise 6. Run a k-means clustering algorithm (e.g. Lloyd's algorithm) on a "planted" data set in $\mathbf{R}^2$ consisting of 50 samples from (X, Y) and another 50 samples from (Z, W) where X, Y, Z, W are all independent Gaussians with variance 1, X, W have mean zero, Y has mean 1 and Z has mean 2. Try at least the values $k = 2, 3, 4, 5$.

Then, run a k-means clustering algorithm on Airline Safety Information, and try to find out something interesting (this part of the question is intentionally open ended).

Exercise 7. Let n be a positive integer. Let c_n be the number of boolean functions $f\colon \{-1,1\}^n \to \{-1,1\}$ that are linear threshold functions. This quantity is of interest since it roughly quantifies the "expressive power" of linear threshold functions for the supervised learning problem. It is known that

$$c_n = 2^{n^2(1+o(1))}$$

So, the supervised learning problem asks for the linear threshold function that fits the given data among a family of functions of super-exponential size. For another perspective on the "expressive power" of linear threshold functions, we will look into the VC-dimension later in the course.

Using an inductive argument prove the weaker lower bound

$$c_n \geq 2^{n(n-1)/2}.$$

(Hint: induct on n. If $f\colon \{-1,1\}^n \to \{-1,1\}$, consider $\overline{f}\colon \{-1,1\}^{n+1} \to \{-1,1\}$ defined (partially for now) so that $\overline{f}(x_1, \ldots, x_n, -1) := f(x_1, \ldots, x_n)$ for all $(x_1, \ldots, x_n) \in \{-1,1\}^n$. How many ways can we define $\overline{f}$ on the remaining "half" of the hypercube $\{-1,1\}^{n+1}$ such that $\overline{f}$ is a linear threshold function?)

As we will discuss later, it is of interest to state the general learning problem for compositions of linear threshold functions (i.e. neural networks). In this case, asymptotics for the number of such functions were recently found in https://arxiv.org/pdf/1901.00434.pdf.

Exercise 8. Let $a > 0$. Let $X^{(1)}, \ldots, X^{(k)} \in \mathbf{R}^n$ be independent identically distributed samples from a Gaussian random vector with mean $(a, 0, \ldots, 0)$ and identity covariance matrix). Let $X^{(k+1)}, X^{(k+2)}, \ldots, X^{(2k)} \in \mathbf{R}^n$ be independent identically distributed samples from a Gaussian random vector with mean $(-a, 0, \ldots, 0)$, where $a > 0$ is known. As in our analysis of the perceptron algorithm, define

$$\mathcal{B} := \max_{i=1,\ldots,2k} \left|\left|X^{(i)}\right|\right|$$

$$\Theta := \min\left\{ ||w|| : \forall\, 1 \leq i \leq 2k \;\; y_i \langle w, X^{(i)} \rangle \geq 1 \right\}.$$

(If the minimum w does not exist, instead define $\Theta := \infty$.)

Define $y_1 = \cdots = y_k := 1$, and $y_{k+1} = \cdots = y_{2k} := -1$.

Give some reasonable estimates for $\mathbf{E}\mathcal{B}$ and $\mathbf{E}(1/\Theta)$ as a function of a.