EE 127 1/23/14
L. El Ghaoui

Homework Assignment #1

Due date: 2/6/14, in class.

Exercise 1 (Rank and nullspace) Consider the image in Figure 1, a gray-scale rendering of a painting by Mondrian (1872-1944). We build a 256×256 matrix A of pixels based on

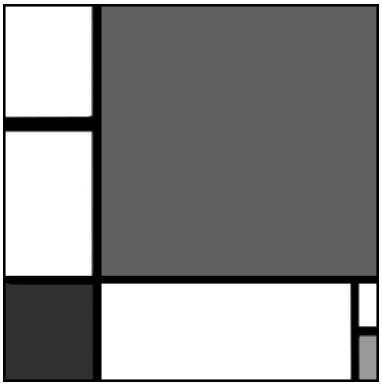

Figure 1: A gray-scale rendering of a painting by Mondrian.

this image by ignoring grey zones, assigning +1 to horizontal or vertical black lines, +2 at the intersections, and zero elsewhere. The horizontal lines occur at row indices $100, 200$ and 230, and the vertical ones, at columns indices $50, 230$.

1. What is nullspace of the matrix?
2. What is its rank?

Exercise 2 (Interpretation of covariance matrix) We are given m data points $x^{(1)}, \ldots, x^{(m)}$ in $\mathbb{R}^n$, and denote by Σ the sample covariance matrix:

$$\Sigma \doteq \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \hat{x})(x^{(i)} - \hat{x})^\top,$$

where $\hat{x} \in \mathbb{R}^n$ is the sample average of the points:

$$\hat{x} \doteq \frac{1}{m} \sum_{i=1}^{m} x^{(i)}.$$

We assume that the average and variance of the data projected along a given direction does not change with the direction. In this exercise we will show that the sample covariance matrix is then proportional to the identity.

We formalize this as follows. To a given normalized direction $w \in \mathbb{R}^n$, $\|w\|_2 = 1$, we associate the line with direction w passing through the origin, $\mathcal{L}(w) = \{tw : t \in \mathbb{R}\}$. We then consider the projection of the points $x^{(i)}$, $i = 1, \ldots, m$, on the line $\mathcal{L}(w)$, and look at the associated coordinates of the points on the line. These *projected values* are given by

$$t_i(w) \doteq \arg\min_t \ \|tw - x^{(i)}\|_2, \quad i = 1, \ldots, m.$$

We assume that for any w, the sample average $\hat{t}(w)$ of the projected values $t_i(w)$, $i = 1, \ldots, m$, and their sample variance $\sigma^2(w)$, are both constant, independent of the direction w. Denote by $\hat{t}$ and σ^2 the (constant) sample average and variance. Justify your answer to the following questions as carefully as you can.

1. Show that $t_i(w) = w^\top x^{(i)}$, $i = 1, \ldots, m$.
2. Show that the sample average $\hat{x}$ of the data points is zero.
3. Show that the sample covariance matrix Σ of the data points is of the form $\sigma^2 I_n$. *Hint:* the largest eigenvalue $\lambda_{\max}$ of the matrix Σ can be written as: $\lambda_{\max} = \max_w \{w^\top \Sigma w : w^\top w = 1\}$, and a similar expression holds for the smallest eigenvalue.

Exercise 3 (Latent semantic indexing) Latent semantic indexing is an SVD-based technique that can be used to discover text documents similar to each other. Assume that we are given a set of m documents $D_1, \ldots, D_m$. Using a "bag-of-words" technique described in Section 2.1 of the hyper-textbook, we can represent each document D_j is described by an n-vector d_j, where n is the total number of distinct words appearing in the whole corpus. In this exercise, we assume that the vectors d_j are constructed as follows: $d_j(i) = 1$ if word i appears in document D_j, and 0 otherwise. We refer to the $n \times m$ matrix $M = [d_1, \ldots, d_m]$ as the "raw" term-by-document matrix. We will also use a normalized version of that matrix: $\tilde{M} = [\tilde{d}_1, \ldots, \tilde{d}_m]$, where $\tilde{d}_j = d_j / \|d_j\|_2$, $j = 1, \ldots, m$. (In practice, other numerical representation of text documents can be used. For example we may use the *relative frequencies* of words in each document, instead of the l_2-norm normalization employed here.)

Assume we are given another document, referred to as the "query document," which is not part of the collection. We describe that query document as a n-dimensional vector q, with zeros everywhere, except a 1 at indices corresponding to the terms that appear in the query. We seek to retrieve documents that are "most similar" to the query, in some sense. We denote by $\tilde{q}$ the normalized vector $\tilde{q} = q/\|q\|_2$.

1. A first approach is to select the documents that contain the largest number of terms in common with the query document. Explain how to implement this approach, based on a certain matrix-vector product, which you will determine.

2. Another approach is to find the closest document by selecting the index j such that $\|q - d_j\|_2$ is the smallest. This approach can introduce some biases, if for example the query document is much shorter than the other documents. Hence a measure of similarity based on the normalized vectors, $\|\tilde{q} - \tilde{d}_j\|_2$, has been proposed, under the name of "cosine similarity". Justify the use of this name for that method, and provide a formulation based on a certain matrix-vector product, which you will determine.

3. Assume that the normalized matrix $\tilde{M}$ has an SVD $\tilde{M} = U\Sigma V^\top$, with Σ a $n \times m$ matrix containing the singular values, and the unitary matrices $U = [u_1, \ldots, u_n]$, $V = [v_1, \ldots, v_m]$ of size $n \times n$, $m \times m$ respectively. What could be an interpretation of the vectors u_l, v_l, $l = 1, \ldots, r$? *Hint:* discuss the case when r is very small, and the vectors u_l, v_l, $l = 1, \ldots, r$, are sparse.

4. With real-life text collections, it is often observed that M is effectively close to a low-rank matrix. Assume that a optimal rank-k approximation ($k \ll \min(n, m)$) of $\tilde{M}$, $\tilde{M}_k$, is known. In the Latent Semantic Indexing approach[1] to document similarity, the idea is to first project the documents and the query onto the sub-space generated by the singular vectors $u_1, \ldots, u_k$, and then apply cosine similarity approach to the projected vectors. Find an expression for the measure of similarity.

Exercise 4 (Projections and PCA Computation) The dataset for this problem consists of the votes of $n = 100$ Senators in the 2004-2006 US Senate for a total of $m = 542$ bills. Yay (Yes) votes are represented as 1's, Nay (No) as -1's, and the other votes are recorded as 0.

The file `senate.mat` contains the data with the matrix of votes $M \in \mathbb{R}^{m \times n}$.

1. Perform an SVD on the data (make sure to center the data first).

   (a) Plot the singular values and comment.

   (b) Plot the explained variance and comment.

2. In machine learning and statistics, PCA can also be used to reduce the dimensionality of problem while still retaining most of the "information". For n observations and p covariates, this can be a useful speed-up when $n << p$ and the complexity of training a model increases primarily with p.

   In the following exercises, we look at how reductions to dimensions of size 1 and 2 can also be helpful in visualizing data and giving an intuition about its structure.

   (a) Project the data onto the line with maximal variance. Describe the procedure to do this, plot the results and discuss.

[1] In practice, it is often observed that this method produces better results than cosine similarity in the original space, as in part 2.

(b) Project the data onto the two-dimensional plane with a view that has the largest variance by mapping points $x \to \Pi x$ where $\Pi = [u_1\ u_2]^T$ is a matrix that contains the singular vectors corresponding to the first two singular values. Describe the procedure to do this, plot the results and discuss.