67750: Advanced Practical Machine Learning 2017/8

Exercise 1 — Manifold Learning Theoretic Solutions

Dr. Matan Gavish *TA: Daniel Gissin*

1 Theoretical Questions

Note that this section is not graded and you do not have to hand it in.

1.1 PCA

In PCA we diagonalize the empirical covariance matrix of our data, $S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$. For these questions we will assume the data is centered, meaning $\bar{x} = 0$.

1. Show that the data sits on a d-dimensional subspace $V \subset \mathbb{R}^n$ if and only if S is of rank d.
2. Show that the new coordinates are the result of an isometry on the subspace V (distances and norms are preserved).

Solution

1. The empirical covariance matrix is basically an averaging over all of the projection matrices of our centered data points. This should intuitively give us an understanding of why there is such a connection between the rank of S and the dimension of the subspace - if S projects a vector onto a d-dimensional subspace, there most have been d dimensions to our original data points... We assume that our data lies on a d-dimensional subspace V. This means that each data point can be expressed with d orthonormal vectors which we will denote $v_1, v_2, ..., v_d$.

   $$S = \frac{1}{n-1}\sum_{i=1}^{n} x_i {x_i}^T = \frac{1}{n-1}\sum_{i=1}^{n}(\sum_{j=1}^{d} c_{i,j}v_j)(\sum_{j=1}^{d} c_{i,j}v_j)^T = \frac{1}{n-1}\sum_{i=1}^{n}(\sum_{j=1}^{d} c_{i,j}^2 v_j v_j^T) = \frac{1}{n-1}\sum_{j=1}^{d}(\sum_{i=1}^{n} c_{i,j}^2)v_j v_j^T$$

   And we see that S is a summation of d projection matrices onto orthonormal vectors and so is of rank d. We will show the other direction of the proof by noticing that $S = \frac{1}{n-1}X^T X$ (where X's rows are the data points):

   $$S_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k)_i(x_k)_j = \frac{1}{n-1} < X_{:,i}, X_{:,j} >= \frac{1}{n-1} < X^T_{i,:}, X_{:,j} >= \frac{1}{n-1}(X^T X)_{i,j}$$

   So this means that we know that $rank(X^T X) = d$. Now, we will use the SVD decomposition of X to show that $rank(X^T X) = rank(X)$:

   $$X = U\Sigma V^T \rightarrow X^T X = U\Sigma^2 U^T$$

   Since both U and V are unitary (with full rank), then both X and $X^T X$ have the same rank as Σ, which means they have the same rank. Finally, since we just showed that $rank(X) = d$, we can conclude that our data lies on a d-dimensional subspace.

2. In PCA we decompose our symmetric empirical covariance matrix into $S = U\Lambda U^T$, so that U is orthogonal and Λ is diagonal. We then define $y_i = U_d^T x_i$, where U_d is the d vectors corresponding to the highest d eigenvalues of S. Since x_i is in the subspace V which is spanned by the columns of U_d, it can be expressed by a linear combination of the column vectors of U_d:

$$x_i = \sum_{j=1}^{d} c_j u_j \rightarrow ||x_i||^2 = \sum_{j=1}^{d} c_j^2$$

$$||y_i||^2 = ||U_d^T x_i||^2 = \sum_{j=1}^{d} (u_j^T x_i)^2 = \sum_{j=1}^{d} (u_j^T \sum_{k=1}^{d} c_k u_k)^2 = \sum_{j=1}^{d} (\sum_{k=1}^{d} c_k u_j^T u_k)^2 = \sum_{j=1}^{d} c_j^2 = ||x_i||^2$$

Similar arguments show that distances are preserved.

1.2 MDS

1. Show that if the data sits on a d-dimensional subspace of $\mathbb{R}^n$, then the matrix that MDS diagonalizes is of rank d. Assume the data is centered, meaning $\bar{x} = 0$. (Hint: show that we are actually diagonalizing XX^T)

2. Show that the new coordinates are the result of an isometry on the subspace V.

Solution

1. We have seen that if the data sits on a d-dimensional subspace, then $rank(X^T X) = d$. Now, we will show the connection between the matrix we are diagonalizing in MDS and $X^T X$. Our claim will be that if our data is centered:

$$-\frac{1}{2} H \Delta H = XX^T$$

We'll start by with remembering the following equality: $||x_i - x_j||^2 = ||x_i||^2 + ||x_j||^2 - 2x_i x_j = x_i^T x_i + x_j^T x_j - 2x_i x_j$
The matrix XX^T has all the information necessary for the right side of that equality: $\Delta_{i,j} = (XX^T)_{i,i} + (XX^T)_{j,j} - 2(XX^T)_{i,j}$. So if we define $d = diag(XX^T)$, we get:

$$\Delta = d \cdot 1^T + 1 \cdot d^T - 2XX^T$$

And this means:

$$-\frac{1}{2} H \Delta H = -\frac{1}{2} H (d \cdot 1^T + 1 \cdot d^T - 2XX^T) H$$

Now, we will show that $H(d \cdot 1^T)H = H(1 \cdot d^T)H = 0$:

$$H(d \cdot 1^T)H = H(d \cdot 1^T)(I - \frac{1}{n} 11^T) = H(d \cdot 1^T - \frac{1^T 1}{n} 1^T) = H(d \cdot 1^T - d \cdot 1^T) = 0$$

$$H(1 \cdot d^T)H = (H(d \cdot 1^T)H)^T = 0$$

So what we are left with is:

$$-\frac{1}{2} H \Delta H = -\frac{1}{2} H(-2XX^T) H = HXX^T H = XX^T$$

The final equality comes from our assumption that X has mean zero, which means that centering it does not change it. Now finally, since we understand that the matrix we are decomposing in MDS is actually the Gram matrix of our data, we can see the correlation (and equivalence when dealing with the euclidean metric) between MDS and PCA. Since we already saw that $rank(X^TX) = d$, we can safely conclude that $rank(-\frac{1}{2}H\Delta H) = rank(XX^T) = d$ by the same arguments.

2. The proof is identical to the one from the PCA question.

1.3 LLE

In LLE, we obtain an n by n matrix, W, where ω_{ij} is the weight of data point j we use to reconstruct data point i. We also enforce the constraint that each row of W sum to 1. We then minimize

$$\Phi(Y) = \Sigma_{i=1}^{n}(y_i - \Sigma_{j=1}^{n}\omega_{ij}y_j)^2$$

where Y is the n by 1 matrix (we will restrict ourself to the 1 dimensional reduction for simplicity). In class, we showed this to be equivalent to minimizing

$$\Phi(Y) = Y^TMY$$

where

$$M = (I - W)^T(I - W)$$

Also, we saw in class that minimizing RSS_i is equivalent to minimizing $w_i^TGw_i$ (slide 21).

1. Show that $\Phi(Y) = \Phi(Y + c1)$, where c is any constant.
2. Why do we need to impose the constraint that $\frac{1}{n}\Sigma_{i=1}^{n}y_i^2 = 1$ and that $\sum_{i=1}^{n} y_i = 0$?
3. Show that G is invertible iff the k neighbors of x_i are linearly independent.

Solution

1. We will use the fact that $\mathbf{1}$ is an eigenvector of M with eigenvalue 0:

$$\Phi(Y + c1) = (Y + c1)^TM(Y + c1) = Y^TMY + c1^TMY + cY^TM1 + c^21^TM1 =$$

$$= \Phi(Y) + c(M^T1)Y + cY^T0 + c^21^T0 = \Phi(Y) + c(M1)Y = \Phi(Y)$$

2. Our minimization problem is minimized quite easilly by the trivial vector of zeros. Since this solution is not very interesting, we impose the constraint that y be a unit vector ($\frac{1}{n}\Sigma_{i=1}^{n}y_i^2 = 1$). Since Φ is invariant to translation, we can remove that degree of freedom and make sure our solution is centered around 0 by imposing the constraint $\sum_{i=1}^{n} y_i = 0$.

3. G is the gram matrix of the k neighbors of some point, which means $G = XX^T$ where X is a reduced data matrix of only those points. We saw that XX^T is of rank d if and only if the points lie on a d-dimensional subspace, so that means G is of full rank (invertible) if and only if the points are linearly independent (lie completely in $\mathbb{R}^k$).

4. This is simply a matter of differentiating our Lagrangian $L(w, \lambda) = w^T G w - \lambda(1^T w - 1)$:

$$\frac{\partial L}{\partial w} = 2w^T G - \lambda 1^T = 0 \rightarrow w^T G = \frac{\lambda}{2} 1^T \rightarrow w^T = \frac{\lambda}{2} 1^T G^{-1}$$

Differentiating w.r.t. λ will give us back the constraint, which we can use to find the value λ.