Homework 1 - Stat 37710: Machine Learning

Cooper Nederhood

2018.04.05

1. **Question 1: Let A be a symmetric $d \times d$ matrix**

   (a) ANSWER:
   Because A is symmetric,
   $$\langle v, Av' \rangle = \langle Av, v' \rangle$$
   And each is an eigenvector so,
   $$\langle v, \lambda' v' \rangle = \langle \lambda v, v' \rangle$$
   $$(\lambda' - \lambda)\langle v, v' \rangle = 0$$
   But $\lambda' \neq \lambda$ so $v \perp v'$

   (b) ANSWER:
   By assumption, S spans V of dimension k so there exist k linearly independent vectors in S. By applying Gram-Schmidt we can construct an orthonormal basis $v^1, ..., v^k$ from this linearly independent set such that $span\{v^1, ..., v^k\} = V$. Further, each v^i will be a linear combination of eigenvectors corresponding to λ. Thus, each v^i is in the λ eigenspace and thus is an eigenvector with eigenvalue λ.
   W.L.O.G let our spanning linearly independent eigenvectors be $\{w_i, ..., w_k\}$. Then performing Gram-Schmidt we have:
   $$v_1 = \frac{w_1}{||w_1||}$$
   $$v_2 = \frac{w_2 - \langle w_2, v_1 \rangle v_1}{||w_2 - \langle w_2, v_1 \rangle v_1||}$$
   $$v_i = \frac{w_i - \langle w_i, v_1 \rangle v_1 - ... - \langle w_i, v_{i-1} \rangle v_{i-1}}{||w_i - \langle w_i, v_1 \rangle v_1 - ... - \langle w_i, v_{i-1} \rangle v_{i-1}||}$$

   (c) ANSWER:
   Taken together statements (a) and (b) imply that a $d \times d$ symmetric matrix A has d linearly independent eigenvectors (even if eigenvalues are repeated) and there exist eigenvectors $v_i, ..., v_d$ such that $||v_i|| = 1$. This further implies that A is orthogonally diagonalizable and therefore we can write A as:
   $$A = \begin{bmatrix} v_1 & \dots & v_d \end{bmatrix} \begin{bmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_d \end{bmatrix} \begin{bmatrix} v_1 & \dots & v_d \end{bmatrix}^T$$
   $$A = \begin{bmatrix} \lambda_1 v_1 & \dots & \lambda_d v_d \end{bmatrix} \begin{bmatrix} v_1 & \dots & v_d \end{bmatrix}^T$$
   $$A = \sum_{i=1}^{d} \lambda_i v_i v_i^T$$
   NOTE: $P = \begin{bmatrix} v_1 & \dots & v_d \end{bmatrix}$ has orthonormal columns so $P^{-1} = P^T$

(d) ANSWER:
Below we find constrained extremum of the desired function.

$$f(x) = \frac{w^T Aw}{||w||^2} = \frac{w^T}{||w||} A \frac{w}{||w||} = u^T Au$$

Where $||u|| = 1$. So, do Langrange maximization of $f(u)$ such that $u^T u = 1$

$$L(u, \lambda) = u^T Au - \lambda(u^T u - 1)$$

$$\frac{\partial L}{\partial u} = 2Au - 2\lambda u = 0$$

$$\frac{\partial L}{\partial \lambda} = u^T u - 1 = 0$$

Together these equations show we have the following extremum:

$$Au = \lambda u$$

$$||u|| = 1$$

So extremum at eigenvalues with size 1. And because of the ordering of each λ_i we can see that the maximum occurs at v_d and the minimum occurs at v_1.

2. **Question 2: Orthogonal projections minimize distance to a subspace**

(a) ANSWER:
To find the point in $V \in R^k$ closest to x we can take each dimension $1, ..., k$ independently. So, WLOG fix some dimension k.
Let $x_{V_k} = (x \cdot p_k)p_k$ so x_{V_k} is the orthog projection of x to p_k. Let y be some point in the p_k space.

$$||x - x_{V_k}||^2 \leq ||x - x_{V_k}||^2 + ||x_{V_k} - y||^2$$
$$= ||x - x_{V_k} + x_{V_k} - y||^2 = ||x - y||^2$$

Thus, to minimize $||x - y||$ we must set $y = x_{V_k}$. Now, when projecting to the space of $p_1, ..., p_k$ we simply repeat this for k dimensions and we have our result.

(b) ANSWER:
For notations sake, we show the result for a single observation x_i and then summing over all x_i the result can be repeated.

$$||x_i - \sum_j \langle v_j, x_i \rangle v_j||^2 = \langle x_i - \sum_j \langle v_j, x_i \rangle v_j, x_i - \sum_j \langle v_j, x_i \rangle v_j \rangle$$

$$= \langle x_i, x_i \rangle - 2\langle x, \sum_j \langle v_j, x_i \rangle v_j \rangle + \langle \sum_j \langle v_j, x_i \rangle v_j, \sum_j \langle v_j, x_i \rangle v_j \rangle$$

$$= \langle x_i, x_i \rangle - 2[\sum_j \langle x, v_j \rangle \langle x, v_j \rangle] + \sum_j \langle x, v_j \rangle \langle x, v_j \rangle + \sum_{i,j: i \neq j} \langle x, v_j \rangle \langle x, v_j \rangle \langle v_i, v_j \rangle$$

$$= ||x||^2 - \sum_j \langle x, v_j \rangle^2$$

The first term is cleary not affected by v_j and the left term, being negative, implies that our original minimization problem is equivalent to *maximizing* $\sum_j \langle x, v_j \rangle^2$ which, per slide 12 of 02DimensionalityReduction reduces to the Rayleigh quotient problem $v^T \Sigma v$. And we know we maximize this by choosing the largest so desired Eigenvectors from the var-cov matrix Σ
NOTE: slide 12 has the proof of this final result, so rather than regurgitate it I am simply appealing to the result.

3. **Question 3: Gram matrix questions**

   (a) ANSWER:
   Let $A = [x_1, x_2, ..., x_n]$ where $x_i \in R^d$ be the centered data matrix which therefore has dimension $d \times n$.
   Further, because $n \geq d$, the matrix A has max rank of d.
   The gram matrix, G, can be defined as $G = A^T A$.
   We will show that $rank(A^T A) = rank(A)$ and thus $rank(G) \leq d$
   To show that $rank(A^T A) = rank(A)$ we show that the dimension of each null space is equal thus implying the respective ranks are equal. Let $N(A)$ denote the null space of A.
   Let $x \in N(A)$. Thus:

$$\Rightarrow Ax = 0$$
$$\Rightarrow A^T Ax = 0$$
$$\Rightarrow x \in N(A^T A)$$

   Similarly, let $x \in N(A^T A)$. Thus:

$$\Rightarrow A^T Ax = 0$$
$$\Rightarrow x^T A^T Ax = 0$$
$$\Rightarrow (Ax)^T Ax = 0$$
$$\Rightarrow Ax = 0$$
$$\Rightarrow x \in N(A)$$

   Putting this together, the null spaces are equal which by rank-nullity implies their ranks are equal

   (b) ANSWER:
   The matrix $K \in R^{n \times n}$ is PSD which implies we can compute the Cholesky decomposition such that $K = R^T R$ where R is an upper-triangle matrix. Further, K is clearly the Gram matrix corresponding to R and because K has rank r, we know R also has rank r. Let us extract the n columns from R. We know there are n columns and we know there are r linearly independent columns but each column is in some unknown space $R^?$ which may have dimension less than d. Thus, if needed we can append zeros to each column such that each column, which we can call $x_i \in R^d$, and the resulting Gram matrix of this data is K.

4. **Question 4: Centering matrix P questions**

   (a) ANSWER:
   As defined we have P is symmetric and thus $P^T = P$ so $P^2 = PP$. Below I show that $PP = P$

$$PP = \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \end{bmatrix} \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \end{bmatrix}$$

$$PP_{i=i} = (1 - \frac{1}{n})^2 + (n-1)\frac{1}{n^2} = 1 - \frac{1}{n} = P_{i=i}$$

$$PP_{i \neq j} = 2[-\frac{1}{n}(1 - \frac{1}{n})] + (n-2)\frac{1}{n^2} = -\frac{1}{n} = P_{i \neq j}$$

   Thus, $P^2 = P$

(b) ANSWER:
First, we show going the "$\Rightarrow$" direction:
So, assume $Pv = 0$

$$Pv = \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = 0$$

This results in the n system of equations. NOTE: the $\langle v_i \rangle$ denotes the ith element is excluded from the list:

$$(1-\frac{1}{n})v_1 - \frac{1}{n}(v_2 + ... + v_n) = 0 \qquad \text{(eq. 1)}$$

$$(1-\frac{1}{n})v_i - \frac{1}{n}(v_1 + ...\langle v_i \rangle ... + v_n) = 0 \qquad \text{(eq. i)}$$

Solving this system yields $v_1 = v_2 = ... = v_n$. To illustrate, we can solve for $n = 2$. If $n = 2$, then we have:

$$v_1 - \frac{v_1}{n} - \frac{v_2}{n} = 0$$
$$-\frac{v_1}{n} + v_2 - \frac{v_2}{n} = 0$$

Combining, we have $v_1 - v_2 = 0 \Rightarrow v_1 = v_2$.
So v is the vector of ones times some constant (which could be zero) and we have our result.
Next, we go the "$\Leftarrow$" direction:
If $v = 0$ the result is obvious. If $v = [1]\lambda$ we can simply do the following:

$$Pv = \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \lambda$$

Each n row has equation of the form $1 - \frac{1}{n} - \frac{n-1}{n}$ which clearly equals zero so we have our result.

5. **QUESTION 5: Local linear embedding eigenvector problem derivation**
ANSWER: Define $\Psi(y_1, ..., y_n) = \sum_i^n ||y_i - \sum_j w_{i,j} y_j||^2$. Without loss of generality, to simplify notation let $y_i \in R^1$
Thus, the objective function is

$$\Psi(y_1, ..., y_n) = \sum_i^n (y_i - \sum_j w_{i,j} y_j)^2$$

$$= \sum_i^n [y_i^2 - y_i(\sum_j w_{i,j} y_j) - (\sum_j w_{i,j} y_j) y_i + (\sum_j w_{i,j} y_j)^2]$$
$$= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T(\mathbf{WY}) - (\mathbf{WY})^T\mathbf{Y} + (\mathbf{WY})^T(\mathbf{WY})$$
$$= ((\mathbf{I}-\mathbf{W})\mathbf{Y})^T((\mathbf{I}-\mathbf{W})\mathbf{Y})$$
$$= \mathbf{Y}^T(\mathbf{I}-\mathbf{W})^T(\mathbf{I}-\mathbf{W})\mathbf{Y}$$

Let $\mathbf{M} = \mathbf{I} - \mathbf{W}$. Then $\Psi = \mathbf{Y}^T\mathbf{M}\mathbf{Y}$. Note, $\mathbf{M}$ is the Gram matrix.
Because we assume $y_i \in R^1$ the var-cov I constraint becomes $\mathbf{Y}^T\mathbf{Y} = 1$. Constructing the Lagrangian we have:

$$\mathscr{L}(\mathbf{Y}, \lambda) = \mathbf{Y}^T\mathbf{M}\mathbf{Y} - \lambda(\mathbf{Y}^T\mathbf{Y} - 1)$$

$$\frac{\partial \mathscr{L}}{\partial \mathbf{Y}} = 2\mathbf{MY} - 2\lambda\mathbf{Y} = 0$$
$$\mathbf{MY} = \lambda\mathbf{Y}$$

Just as with PCA, $\mathbf{M}$ is a symmetric matrix and thus has n orthonormal eigenvectors. Thus we can maximize and minimize (in this case we want to minimize) by selecting the eigenvectors corresponding to the smallest desired eigenvalues.