[ML20] Assignment 9

Danny Radosevich

Due: Apr 13 (b 8:45am)

Let $x \in \mathbb{R}^p$ be a random instance and $\Sigma_x \in \mathbb{R}^{p \times p}$ be its covariance matrix. Let $w_1, w_2 \in \mathbb{R}^p$ be its first two optimal PCA projection vectors; we have studied how to derive them in the lecture.

[1] Please derive the third optimal PCA projection vector w_3 by solving (1), presuming w_1 and w_2 are given. Please answer this question based on the following three steps.

$$\max_{w_3} \; w_3^T \Sigma_x w_3 \quad s.t. \; ||w_3||^2 = 1, \; cov(w_3 x, w_1 x) = 0, \; cov(w_3 x, w_2 x) = 0. \tag{1}$$

(i) Show that $cov(w_3, w_1)$=0 is equivalent to $w_3^T w_1$=0, and $cov(w_3, w_2)$=0 is equivalent to $w_3^T w_2$=0.

$cov(w_3, w_1)$=0 is equivalent to $w_3^T w_1$=0 .
Assume that w_3 and w_1 are statistically uncorrelated, so $cov(w_3, w_1)$=0
$cov(w_3, w_1)$=$E[w_3^T x - E w_3^T x][w_1^T - E w_1^t x]$
treat w_1 and w_3 as unkown constants
$= E[w_3^T (x - Ex)][w_1^T (x - Ex)]$
$= E[w_3^T (x - Ex)][(x - Ex)^t w_1]$
$= w_3^T E(x - Ex)(x - Ex)^t w_1$
$E(x - Ex)(x - Ex)^t = \Sigma_x$
$= w_3^T \Sigma_x w_1$
When solving for w_1 we find that $\Sigma_x w_1 = \lambda w_1$ so:
$= w_3^T \Sigma_x w_1 = \lambda w_3^T w_1 = 0$

$cov(w_3, w_2)$=0 is equivalent to $w_3^T w_2$=0 .
Assume that w_3 and w_2 are statistically uncorrelated, so $cov(w_3, w_2)$=0
$cov(w_3, w_2)$=$E[w_3^T x - E w_3^T x][w_2^T - E w_2^t x]$
treat w_2 and w_3 as unkown constants
$= E[w_3^T (x - Ex)][w_2^T (x - Ex)]$
$= E[w_3^T (x - Ex)][(x - Ex)^t w_2]$
$= w_3^T E(x - Ex)(x - Ex)^t w_2$
$E(x - Ex)(x - Ex)^t = \Sigma_x$
$= w_3^T \Sigma_x w_2$
When solving for w_2 we find that $\Sigma_x w_2 = \lambda w_2$ so:
$= w_3^T \Sigma_x w_2 = \lambda w_3^T w_2 = 0$

(ii) In order to apply Lagrange Multiplier to solve (1), we need to first construct the Lagrange function

$$J = w_3^T \Sigma_x w_3 + \lambda_1 (w_3^T w_3 - 1) + \lambda_2 (w_3^T w_1) + \lambda_3 (w_3^T w_2). \tag{2}$$

Show that $\lambda_2 = 0$ and $\lambda_3 = 0$.

$\boldsymbol{\lambda_2 = 0}$.
$\frac{\partial}{\partial w_2} J' = 2\Sigma_x w_2 - 2\lambda_1 w_2 - \lambda_2 w_1$
$0 = 2\Sigma_x w_3 - 2\lambda_1 w_3 - \lambda_2 w_1$
$w_1^T(0 = 2\Sigma_x w_3 - 2\lambda_1 w_3 - \lambda_2 w_1)$
$0 = w_1^T 2\Sigma_x w_3 - w_1^T 2\lambda_1 w_3 - w_1^T \lambda_2 w_1$
we know:
$w_1^T \Sigma_x w_2 = 0$
$w_1^T w_2 = 0$
$w_1^T w_1 = 1$
Substitute those in to the derivative and we are left with: $\lambda_2 = 0$

$\boldsymbol{\lambda_3 = 0}$.
$\frac{\partial}{\partial w_3} J' = 2\Sigma_x w_3 - 2\lambda_1 w_3 - \lambda_2 w_1 - \lambda_3 w_2$
$0 = 2\Sigma_x w_3 - 2\lambda_1 w_3 - \lambda_2 w_1 - \lambda_3 w_2$
$\lambda_2 = 0$
$0 = 2\Sigma_x w_3 - 2\lambda_1 w_3 - \lambda_3 w_2$
$w_2^t(0 = 2\Sigma_x w_3 - 2\lambda_1 w_3 - \lambda_3 w_2)$
$0 = 2w_2^T \Sigma_x w_3 - 2w_2^T \lambda_1 w_3 - w_2^T \lambda_3 w_2$
$w_2^T \Sigma_x w_2 = 0$
$_2^T w_3 = 0$
$w_2^T w_2 = 1$
So we are left with:
$\lambda_3 = 0$

(iii) Optimize J over w_3 and show the optimal w_3 is an eigenvector of Σ_x. (Need to show derivatives.)
$\frac{\partial}{\partial w_3} J' = 2\Sigma_x w_3 - 2\lambda_1 w_3 = 0$
$\Sigma_x w_3 = \lambda_1 w_3$
This shows that w_3 is an eigenvector of Σ_x with λ being the optimal eigenvalue.

Implement PCA and apply it to learn p projection vectors $w_1, w_2, \ldots, w_{p-1}, w_p$ from a set of instances $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$. Assume these projection vectors are sorted as follows: w_1 is associated with the largest eigenvalue, w_2 with the second largest, ..., and w_p with the smallest eigenvalue.

[2] Plot the distribution of these instances in a 2D feature space in two figures. In Fig 1, the two latent features are generated through w_1 and w_2. In Fig 2, the two are generated through w_{p-1} and w_p.

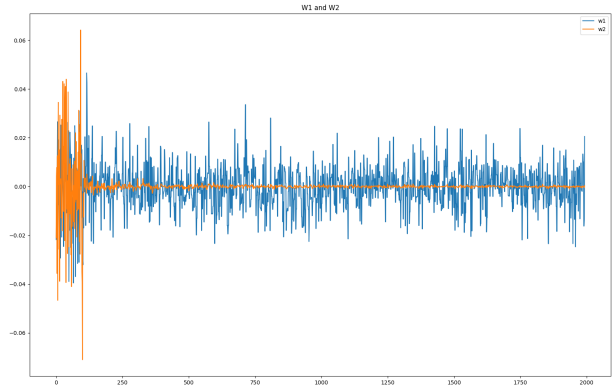

Fig. 1. Data Distribution based on w_1 and w_2.

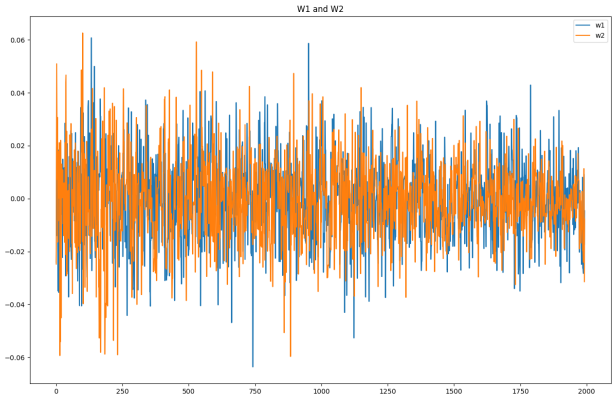

Fig. 2. Data Distribution based on w_{p-1} and w_p.

Continue [2]. Let us train a prediction model in the PCA projected space and evaluate its performance. Draw your results as a curve in Fig 3, where x-axis is k and y-axis is testing mse.

For the prediction task, let $x_1, \ldots, x_n$ be a set of training instances, and $t_1, \ldots, t_m$ be a set of testing instances – keep in mind they are all p-dimensional feature vectors, and their labels are assumed given. Let $\tilde{x} \in \mathbb{R}^k$ be the latent feature vector of an arbitrary instance x (training or testing) generated trough

$$\tilde{x} = \begin{bmatrix} w_1^T x \\ w_2^T x \\ \vdots \\ w_k^T x \end{bmatrix} \in \mathbb{R}^k. \tag{3}$$

For each choice of k, we can follow the following steps to get the testing mse.

(i) Train a set of optimal PCA projection vectors $w_1, w_2, \ldots, w_k$ from the *training set*. (Done in [2].)

(ii) Obtain a set of projected training instances $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n \in \mathbb{R}^k$, where $\tilde{x}_i$ is the latent feature vector of x_i obtained through (3).

(iii) Train a linear regression f model on $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ and their labels. (Note that labels are untouched.)

(iv) Obtain a set of projected testing instances $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_n \in \mathbb{R}^k$, where $\tilde{t}_i$ is the latent feature vector of t_i obtained through (3).

(v) Apply f on $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_n$ to predict their labels and estimate testing mse. (Question: why on the projected testing instances? Can we directly evaluate f on the original testing instances?)

Note: choose 5 values of k yourself and try to make the curve as convergent as possible. (Think about how mse may behave when k is extremely small or extremely large.)

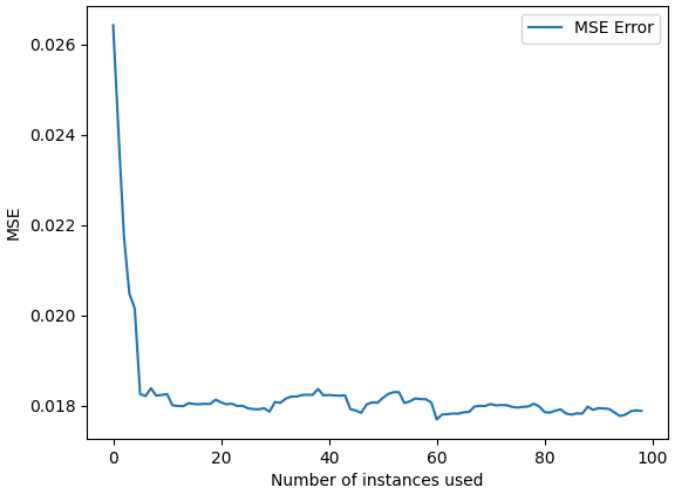

Fig. 3. Data Distribution based on w_{p-1} and w_p.

[4] (Bonus) Kernel PCA is a nonlinear dimensionality reduction technique, which first maps data into a higher dimensional space, and then finds an optimal PCA projection vector in that space. However, since the mapping is implicit, kernel PCA needs to apply the Representer theorem and kernel trick to derive an analytic solution for the optimal projection vector.

Let's verify the Representer theorem. Let $x_1, \ldots, x_n$ be a set of instances and Σ_x be their covariance matrix. Let w be the optimal PCA projection vector learned from this set. Please show that w can be written as

$$w = \sum_{i=1}^{n} \alpha_i \bar{x}_i, \tag{4}$$

where α_i's are unknown coefficients and $\bar{x}_i$ is a 'centered' instance of x_i defined as

$$\bar{x}_i = x_i - \frac{1}{n} \sum_{j=1}^{n} x_j. \tag{5}$$

Tip: recall the optimal PCA project vector satisfies $\Sigma_x w = \lambda w$.

[5] (Bonus) Let us try to mathematically develop a variant of CCA. For a pair of variables x and z, suppose we want to find a pair of projection vectors w and v such that the covariance (not correlation) between these variables in the projected space is maximized. In addition, you are NOT allowed to use the constraints that $var(w^T x) = c_1$ and $var(v^T z) = c_2$ (or their square-roots), where c_1 and c_2 are constants.

Please derive the optimal w and v. If necessary, design your own constraints that do not violate the above requirements. Eventually, you may just show w and v are generalized eigenvectors of some problem.