CS760 Machine Learning

Principal Component Analysis

Lecturer: Jerry Zhu *jerryzhu@cs.wisc.edu*

1 Basic Linear Algebra Review

Scalar (1×1), vector (default column vector, $n \times 1$), matrix ($n \times m$). Matrix transpose $\left(A^\top\right)_{ij} = A_{ji}$.

A $n \times m$ matrix A times a $m \times p$ matrix B is a $n \times p$ matrix C, with $C_{ij} = \sum_{k=1}^{m} A_{ik}B_{kj}$. Check dimensions.

$(AB)C = A(BC), A(B+C) = AB + AC, (A+B)C = AC + BC, (A+B)^\top = A^\top + B^\top, (AB)^\top = B^\top A^\top$. Note in general $AB \neq BA$.

The following is specific to square matrices.

Diagonal matrix: $A_{ij} = 0, \forall i \neq j$. Identity matrix I is diagonal with $I_{ii} = 1, \forall i$. $AI = IA = A$ for all square A.

Some square matrices have inverses: $AA^{-1} = A^{-1}A = I$. $(AB)^{-1} = B^{-1}A^{-1}$. $(A^\top)^{-1} = (A^{-1})^\top$.

The trace is the sum of diagonal elements (or eigenvalues) $\mathrm{Tr}(A) = \sum_i A_{ii}$.

The determinant $|A|$ is the product of eigenvalues. $|AB| = |A||B|$, $|a| = a$, $|aA| = a^n|A|$, $|A^{-1}| = 1/|A|$. A matrix A is invertible iff $|A| \neq 0$.

If $|A| = 0$ for a $n \times n$ square matrix A, A is said to be singular. This means at least one column is linearly dependent on (i.e., a linear combination of) other columns (same for rows). Once all such linearly dependent columns and rows are removed, A is reduced to a smaller $r \times r$ matrix, and r is called the rank of A.

A $m \times m$ matrix A has m eigenvalues λ_i and eigenvectors (up to scaling) u_i s.t. $Au_i = \lambda_i u_i$. In general λ's are complex numbers. If A is real and symmetric, λ's are real numbers, and u's are orthogonal. The u's can be scaled to orthonormal, i.e., length one, so that $u_i^\top u_j = I_{ij}$. The spectral decomposition is $A = \sum_i \lambda_i u_i u_i^\top$. For invertible A, $A^{-1} = \sum_i \frac{1}{\lambda_i} u_i u_i^\top$. This shows why the determinant must be non-zero.

A real symmetric matrix A is positive semi-definite, if its eigenvalues $\lambda_i \geq 0$, $\forall i$. Equivalently, $\forall x \in \mathbb{R}^n, x^\top Ax \geq 0$. It is (strictly) positive definite if $\lambda_i > 0$, $\forall i$.

A positive semi-definite matrix has rank r equal to the number of positive eigenvalues. The remaining $n - r$ eigenvalues are zero.

For vector $x \in \mathbb{R}^n$, we have
0-norm: $\|x\|_0 =$ count of nonzero elements
1-norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
2-norm (the Euclidean norm, or just 'the norm', length: $\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$

∞-norm: $\|x\|_\infty = \max_{i=1}^n |x_i|$

2 Principal Component Analysis (PCA)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^D$. It is convenient to assume that the points are centered $\sum_i \mathbf{x}_i = 0$. One can always centered the data by subtracting the sample mean:

$$\mu = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i.$$

$$\mathbf{x}_i := \mathbf{x}_i - \mu.$$

We want to represent these points in some lower dimensional space $\mathbb{R}^d$ where typically $d \ll D$. We form the sample covariance matrix

$$S = \frac{1}{n-1} \sum_i \mathbf{x}_i \mathbf{x}_i^\top. \tag{1}$$

We then perform an eigen decomposition

$$S = U \Lambda U^\top, \tag{2}$$

where the columns of U are the eigenvectors $u_1, \ldots, u_D$, the diagonal elements of Λ are the eigenvalues $\lambda_1, \ldots, \lambda_D$. Assuming the eigenvalues are sorted from large to small: $\lambda_1 \geq \ldots \geq \lambda_D$. Take the first d eigenvectors $u_1, \ldots, u_d$. The new representation of any $\mathbf{x}_i$ is

$$(u_1^\top \mathbf{x}_i, \ldots, u_d^\top \mathbf{x}_i)^\top.$$

2.1 The Variance Preservation View

PCA can be justified in several ways. Let's consider a projection onto a line going through the origin. Such a line can be specified by a vector $\mathbf{w} \in \mathbb{R}^D$. The projection of $\mathbf{x}$ is

$$\frac{\mathbf{w}^\top \mathbf{x}}{\|\mathbf{w}\|}. \tag{3}$$

For simplicity, let us consider $\mathbf{w}$ with unit length. The variance of the projected dataset is

$$\frac{1}{n-1} \sum_{i=1}^n (\mathbf{w}^\top \mathbf{x}_i)^2 = \mathbf{w}^\top S \mathbf{w}, \tag{4}$$

where

$$S = \frac{1}{n-1} \sum_i \mathbf{x}_i \mathbf{x}_i^\top \tag{5}$$

is the sample covariance matrix since we assume the dataset is centered. The goal of PCA (in this 1D case) is to find the $\mathbf{w}$ that maximizes the variance, in the hope that it maximally preserves the distinction among points. This leads to the following optimization problem

$$\max_{\mathbf{w}} \quad \mathbf{w}^\top S \mathbf{w} \tag{6}$$

$$\text{s.t.} \quad \|\mathbf{w}\| = 1. \tag{7}$$

Let's solve it by forming the Lagrangian

$$\mathbf{w}^\top S \mathbf{w} + \lambda(1 - \mathbf{w}^\top \mathbf{w}). \tag{8}$$

The gradient w.r.t. $\mathbf{w}$ is

$$\nabla = 2S\mathbf{w} - 2\lambda\mathbf{w}. \tag{9}$$

Setting to zero, we find that

$$S\mathbf{w} = \lambda\mathbf{w}, \tag{10}$$

i.e., the desired direction $\mathbf{w}$ is an eigenvector of S! But which one? Recall the projected variance is

$$\mathbf{w}^\top S\mathbf{w} = \mathbf{w}^\top \lambda\mathbf{w} = \lambda, \tag{11}$$

we see that we want λ to be the largest eigenvalue of S and $\mathbf{w}$ the corresponding eigenvector. In other words, let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of S in non-increasing order, and $u_1, \ldots, u_n$ be the corresponding eigenvectors. Then u_1 is the maximum variance preserving direction, and the resulting variance is simply λ_1. This is PCA with $d = 1$: a D-dimensional point $\mathbf{x}$ is projected to a scalar $u_1^\top \mathbf{x}$. Note that when S's top eigenvalue has multiplicity larger than one, e.g., $\lambda_1 = \lambda_2$, then PCA is not unique: any unit vector in $\text{span}(u_1, u_2)$ can be the PCA direction.

If we want $d > 1$, it can be shown that we want to project $\mathbf{x}$ onto the first d eigenvectors

$$\mathbf{x} \to (u_1^\top \mathbf{x}, \ldots, u_d^\top \mathbf{x})^\top. \tag{12}$$

Recall that one can view $u_1, \ldots, u_D$ as the D major-to-minor axes of an ellipsoid represented by the sample covariance matrix (NB this does not assume that the underlying distribution is Gaussian). Clearly, if $d = D$ then $u_1 \ldots u_D$ is a basis for $\mathbb{R}^D$, and this PCA projection amounts to a rotation of the coordinate system (align them with the eigenvectors) without any loss of information.

2.2 The Minimum Reconstruction Error View

Using *any* orthonormal basis $\mathbf{u}_1 \ldots \mathbf{u}_D$, a training point $\mathbf{x}_i$ (recall it has been centered) can be written as

$$\mathbf{x}_i = \sum_{j=1}^{D} \alpha_{ij}\mathbf{u}_j \tag{13}$$

where

$$\alpha_{ij} = \mathbf{u}_j^\top \mathbf{x}_i. \tag{14}$$

Consider the d-term approximation to $\mathbf{x}_i$:

$$\hat{\mathbf{x}}_i = \sum_{j=1}^{d} \alpha_{ij}\mathbf{u}_j. \tag{15}$$

We want the approximation error to be small for all training points:

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} \| \sum_{j=d+1}^{D} \alpha_{ij}\mathbf{u}_j\|^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=d+1}^{D} \alpha_{ij}^2 \tag{16}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=d+1}^{D} \mathbf{u}_j^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{u}_j = \sum_{j=d+1}^{D} \mathbf{u}_j^\top S\mathbf{u}_j. \tag{17}$$

If $d = D-1$, i.e., we need to remove a single dimension, it is easy to see that $\mathbf{u}_D = u_D$ because $u_D^\top S u_D = \lambda_D$ is the smallest among all unit vectors. Similarly, the other dimensions to remove are subsequently the eigenvectors corresponding to the least eigenvalues.

2.3 The Singular Value Decomposition View

Recap: We could have formed a $n \times D$ matrix X with the centered points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Then our sample covariance matrix is

$$S = \frac{1}{n-1} X^\top X,$$

and our eigen decomposition is

$$S = U \Lambda U^\top.$$

If we form the $D \times d$ matrix $U_d = [u_1 | \ldots | u_d]$, The PCA projection of X is

$$X U_d.$$

This is a $n \times d$ matrix where the ith row is the new representation of $\mathbf{x}_i$.

But we will now perform singular value decomposition (SVD) on X directly, without forming S at all. SVD performs

$$X_{n \times D} = L_{n \times m} \Sigma_{m \times m} V^\top_{m \times D}$$

where $m = \min(n, D)$, L contains orthonormal columns, so does V, and Σ is a diagonal matrix with singular values $\sigma_1, \ldots, \sigma_m$ on the diagonal. If we were to write the sample covariance matrix using SVD of X, we get

$$S = \frac{1}{n-1} X^\top X = \frac{1}{n-1} V \Sigma L^\top L \Sigma V^\top = \frac{1}{n-1} V \Sigma^2 V^\top.$$

Equating this with

$$S = U \Lambda U^\top, \tag{18}$$

we see that

$$\lambda_i = \frac{\sigma_i^2}{n-1}, \;\; V = U. \tag{19}$$

So the PCA projection of X can be performed via SVD as

$$X V_d$$

as well.