INFR10086 Machine Learning (MLG) Semester 2, 2025/6

Tutorial 4: Singular Value Decomposition

In this tutorial, we will review singular value decomposition (SVD), its applications to principal component analysis (PCA), and their geometric interpretation.

1 Singular value decomposition

Singular value decomposition is a factorization of a matrix. For a matrix A, the singular value decomposition is

$$A = \sum_{i=1}^{d} \sigma_i u_i v_i^\top \tag{1}$$

where $\sigma_1 \geq \cdots \geq \sigma_d$ are called singular values. The vectors $u_1, \dots, u_d$ satisfies $u_i^\top u_j = 0$ for any $i \neq j$ and $u_i^\top u_i = 1$ for $i = 1, \dots, d$, and similarly the vectors $v_1, \dots, v_d$ satisfies $v_i^\top v_j = 0$ for any $i \neq j$ and $v_i^\top v_i = 1$ for $i = 1, \dots, d$.

Discussion. Show that in the definition above if $U = \begin{bmatrix} u_1 & \dots & u_d \end{bmatrix}$ and $V = \begin{bmatrix} v_1 & \dots & v_d \end{bmatrix}$, then $U^\top U = I$ and $V^\top V = I$.

Discussion. Show that $A = \sum_{i=1}^{d} \sigma_i u_i v_i^\top$ can be succinctly written as $A = U \Sigma V^\top$, where

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_d \end{bmatrix}. \tag{2}$$

It is quite amazing such a complicated factorization exists for any matrix. The above is merely the definition, and we haven't said why such a factorization exists and how to compute it. They are both fascinating topics, but in this tutorial, we will take advantage of `numpy.linalg.svd`.

2 Principal component analysis

Recall that the principal component analysis (PCA) defined in class is a series of optimization problem finding the maximum variance. For a data matrix X, the PCA of X is defined as

$$
\begin{aligned}
v_1 &= \arg\max_v v^\top X^\top X v \quad \text{s.t. } \|v\|_2^2 = 1 && (3) \\
v_2 &= \arg\max_v v^\top X^\top X v \quad \text{s.t. } \|v\|_2^2 = 1 \text{ and } v^\top v_1 = 0 && (4) \\
&\vdots && (5) \\
v_d &= \arg\max_v v^\top X^\top X v \quad \text{s.t. } \|v\|_2^2 = 1 \text{ and } v^\top v_i = 0 \text{ for } i = 1, \dots, d-1 && (6)
\end{aligned}
$$

Discussion. The above definition of PCA is a bit mouthful. Show that the above can be compactly written as

$$X^\top X = \sum_{i=1}^{d} \lambda_i v_i v_i^\top \tag{7}$$

where $\lambda_1 \geq \cdots \geq \lambda_d$ are the corresponding eigenvalues for $v_1, \dots, v_d$. In addition, $v_i^\top v_j = 0$ for any $i \neq j$ and $v_i^\top v_i = 1$.

Discussion. Similar to SVD, show that if $V = \begin{bmatrix} v_1 & \dots & v_d \end{bmatrix}$, then $V^\top V = I$, and $X^\top X = V \Lambda V^\top$ where

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{bmatrix}. \tag{8}$$

We can clearly see some similarity between SVD and PCA.

Discussion. Show that if we write the SVD of X as $X = U \Sigma V^\top$ then

$$X^\top X = V \Sigma^2 V^\top. \tag{9}$$

We can conclude that the eigenvalues of $X^\top X$ are the squares of singular values, i.e., $\lambda_i = \sigma_i^2$ for $i = 1, \dots, d$. This also provides us a simple approach to compute PCA. We compute the SVD of X, avoiding the computation of $X^\top X$.

```
U, S, V = numpy.linalg.svd(X)
```

The matrix `V` gives the principal components, and `S * S` gives the corresponding eigenvalues.

3 Geometric interpretation of SVD

There is a geometric interpretation to SVD and PCA, and it has something to do with ellipsoids. An ellipsoid is a generalization of ellipse in high dimensions. An ellipsoid is defined as the set of points $\{x : (x - v)^\top A(x - v) = 1\}$ for any shift vector v and any positive-definite matrix A.[1]

Implementation. Suppose

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix} \tag{10}$$

Run `plot-ellipse.py` and convince yourself that $x^\top Ax = 1$ is an ellipse in 2D.

Discussion. Run `plot-ellipse.py` with $A = \begin{bmatrix} 0.2 & 0 \\ 0 & 2 \end{bmatrix}$ and see what happens. Run `plot-ellipse.py` again with $A = \begin{bmatrix} 0.5 & 0 \\ 0 & 5 \end{bmatrix}$ and see what happens. What do you think a_{11} and a_{22} are controlling?

Note that the ellipses in the above example is axis aligned, but a rotated ellipse should still be an ellipse. Recall that a rotation matrix in 2D is defined as

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \tag{11}$$

where θ is the angle that we rotate. To rotate the ellipse, we apply R to the points in $\{x : (x - v)^\top A(x - v) = 1\}$. If $y = Rx$, then $x = R^{-1}y = R^\top y$. In other words, the points after rotation should satisfy $\{x : (x - v)^\top RAR^\top (x - v) = 1\}$.

Discussion. A rotation of 45° amounts to multiplying the matrix

$$R = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}. \tag{12}$$

Going back to our example ellipse, if $A = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix}$, show that $RAR^\top = \begin{bmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{bmatrix}$.

Discussion. Run `plot-ellipse.py` with $A = \begin{bmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{bmatrix}$. What do you see?

The form $(x - v)^\top A(x - u)$ reminds us of the Gaussian distribution. Recall that the probability

[1] A positive-definite matrix A satisfies $z^\top Az > 0$ for any $z \neq 0$.

density function of a multivariate Gaussian distribution is defined as

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right), \tag{13}$$

where μ is the mean vector, Σ is the covariance matrix, and $|\Sigma|$ is the determinant of Σ.

Discussion. Run `plot-gaussian.py`. Note that the inverse of the covariance matrix $\Sigma^{-1} = A = \begin{bmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{bmatrix}$. Compare the result you get from `plot-gaussian.py` and `plot-ellipse.py`. What similarities do they share?

Discussion. Use `numpy.linalg.svd` to do SVD on the inverse of the covariance Σ. What are the singular values? Since $\begin{bmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{bmatrix}$ is a rotation of $A = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix}$, what do the singular values mean?

Imagine that we draw n samples from a multivariate Gaussian to form the data matrix X. The data should look like an ellipsoid. The matrix $\frac{1}{\sqrt{n}}X^\top X$ is an estimate of the covariance matrix Σ. The PCA of X is about finding the eigenvectors of $X^\top X$ that correspond to the directions of maximal variance.

Discussion. Run `plot-gaussian-samples.py`. Why are the eigenvectors pointing along the axes of the ellipsoid?

Discussion. What is the geometric interpretation of SVD?