Introduction to Data Science - 1MS041

Benny Avelin

Department of Mathematics

HT 2023

Recap

1. The volume of the unit ball decreases rapidly as dimension increases
2. The volume is mostly concentrated in a thin shell close to the surface of the ball
3. This is one interpretation why concentration of measure holds
4. The Gaussian annulus theorem says that for the spherical Gaussian most of the volume is on a ball of size $\sqrt{d}$, where d is the dimension.
5. The random projection theorem tells us that length is preserved with high probability if we project onto random Gaussian vectors (follows from the Gaussian annulus theorem / concentration)
6. This allows us to project multiple data points and preserve the pairwise distances with high probability.

1. SVD
2. PCA and SVD
3. Whitening etc.
4. Bla di bla?

Projections

Projection

The projection of a point x onto a line with direction v is

$$(v \cdot x)v$$

PCA the beginning

Let $\{X_1, \ldots, X_n\} \in \mathbb{R}^m$ be IID. Let $v \in \mathbb{R}^m$ be a unit vector and define $Y_i = (X_i \cdot v)$, then consider the minimization problem

$$v_1 := \arg\max_{\|v\|=1} \frac{1}{n} \sum_i (Y_i - \overline{Y}_n)^2.$$

Projections

Projection

The projection of a point x onto a line with direction v is

$$(v \cdot x)v$$

PCA the beginning

Let $\{X_1, \dots, X_n\} \in \mathbb{R}^m$ be IID. Let $v \in \mathbb{R}^m$ be a unit vector and define $Y_i = (X_i \cdot v)$, then consider the minimization problem

$$v_1 := \arg\max_{\|v\|=1} \frac{1}{n} \sum_i (Y_i - \overline{Y}_n)^2.$$

If $\overline{Y}_n = 0$ we get

$$v_1 = \arg\max_{\|v\|=1} \sum_{j=1}^{n} |X_i \cdot v|^2.$$

PCA first steps

Gram matrix

Define the Gram matrix as a matrix A of size $n \times m$ with rows given by X_i.

Then we can rewrite

$$\sum_{j=1}^{n} |X_i \cdot v|^2 = |Av|^2$$

Note

For A being the Gram matrix for Data $\{X_1, \ldots, X_n\}$ the vector v_1 satisfies

$$v_1 = \arg\max_{\|v\|=1} |Av|.$$

PCA first steps

Gram matrix

Define the Gram matrix as a matrix A of size $n \times m$ with rows given by X_i.

Then we can rewrite

$$\sum_{j=1}^{n} |X_i \cdot v|^2 = |Av|^2$$

Note

For A being the Gram matrix for Data $\{X_1, \ldots, X_n\}$ the vector v_1 satisfies

$$v_1 = \arg\max_{\|v\|=1} |Av|.$$

The vector v_1 is called the *first singular vector* of A. The singular value is defined as

$$\sigma_1(A) := |Av_1|.$$

What is the second singular vector

The second singular vector

$$v_2 := \arg \max_{\|v\|=1, v \perp v_1} |Av|.$$

We can continue this process to get more vectors by looking for the next vector orthogonal to the previous 2 etc. There are as many singular vectors as there are dimensions, i.e. m and they form a basis.

What is the second singular vector

The second singular vector

$$v_2 := \arg \max_{\|v\|=1, v \perp v_1} |Av|.$$

We can continue this process to get more vectors by looking for the next vector orthogonal to the previous 2 etc. There are as many singular vectors as there are dimensions, i.e. m and they form a basis.

Lemma

Let A be an $(n \times m)$ matrix, and let v_1 be the first singular vector of A and let $\sigma_1(A)$ be the first singular value (with $\sigma_2(A) < \sigma_1(A)$), then v_1 is an eigenvector of the $m \times m$ matrix $A^T A$, and

$$\max_{\|v\|=1} |Av| = |Av_1| = \sqrt{\lambda_1} = \sigma_1(A)$$

where λ_1 is the first eigenvalue of $A^T A$.

Singular Value Decomposition

We can write each row in A as $X_i = \sum_{j=1}^{m} (X_i \cdot v_j) v_j$ which we can now rewrite as

$$A = \sum_{j=1}^{m} A v_j v_j^T$$

denoting $u_i := \frac{Av_i}{\sigma_i}$ we see that the above expression becomes

$$A = \sum_{j=1}^{m} \sigma_j u_j v_j^T$$

We call the vectors u_j the *left singular vectors* That is

$$A = UDV^T.$$

What is PCA

PCA

Let A be the Gram matrix for $\{X_1, \ldots, X_n\}$ (empirical mean zero), decompose using SVD, $A = UDV^T$, and define the PCA transform as

$$PCA(X_1, \ldots, X_n) = AV = UDV^T V = UD.$$