CS5339 Machine Learning

Linear Algebra Background

Lee Wee Sun
School of Computing
National University of Singapore
leews@comp.nus.edu.sg

Semester 2, 2019/20

Linear Algebra

We will go through linear algebra background that is useful for the course.

Outline

Basic Definitions

- The inner product of two d dimensional vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ is

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{d} u_i v_i.$$

- We also denote the inner product as $\mathbf{u}^T \mathbf{v}$ or $\mathbf{u} \cdot \mathbf{v}$.
- We also have $\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos\theta$ where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$.

- Norms
  - The Euclidean norm (ℓ_2 norm) is $\|\mathbf{u}\|_2 = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$.
  - The ℓ_1 norm is $\|\mathbf{u}\|_1 = \sum_{i=1}^{d} |u_i|$.
  - The ℓ_∞ norm is $\|\mathbf{u}\|_\infty = \max_i |u_i|$.
  - When the subscript is not used, we refer to the Euclidean norm.

- A subspace of $\mathbb{R}^d$ is a subset of $\mathbb{R}^d$ that is closed under addition and scalar multiplication.
- The span of a set of vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ is a subspace containing all vectors $\sum_{i=1}^{k} \alpha_i \mathbf{u}_i$, where $\alpha_i \in \mathbb{R}$.

- A set of vectors $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is linearly independent if for every i, $\mathbf{u}_i$ is not in the span of $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, \ldots, \mathbf{u}_k$.
- If U is linearly independent, then $\sum_i a_i \mathbf{u}_i = 0$ implies $a_i = 0$ for $i = 1, \ldots, k$.
- We say U is a basis of a subspace V if it is linearly independent and spans V.
- The dimension of V is the size of a basis of V.
- We call U orthogonal if $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ for $i \neq j$.
- In addition, U is orthonormal if $\|\mathbf{u}_i\| = 1$ for all i.

- If the columns of matrix U are orthonormal:
  - $U^T U = I$ where I is the identity matrix, diagonal with 1's on the diagonal: this is because $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ if $i \neq j$ and equals 1 otherwise.
  - If U is square, then $U^T = U^{-1}$.
  - Multiplication by U preserves distances: $\|U\mathbf{x}\|^2 = \langle U\mathbf{x}, U\mathbf{x} \rangle = \mathbf{x}^T U^T U\mathbf{x} = \|\mathbf{x}\|^2$

- For a matrix $A \in \mathbb{R}^{m,d}$
  - The range is the span of its column, $y = A\mathbf{u}$.
  - The null space is all the vectors $\mathbf{v}$ satisfying $A\mathbf{v} = 0$.
  - The rank of A is the dimension of its range.
- The transpose A^T is the matrix whose entry (i, j) equals the entry (j, i) of A. A is symmetric if $A^T = A$.

Outline

Eigenvalues and Eigenvectors

- For a matrix $A \in \mathbb{R}^{d,d}$, a non-zero vector $\mathbf{u}$ is an eigenvector with eigenvalue λ if $A\mathbf{u} = \lambda\mathbf{u}$.
- **Example [1]:** Consider the matrix

$$S = \begin{pmatrix} 30 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

  - The matrix has rank 3. The 3 non-zero eigenvalues are $\lambda_1 = 30$, $\lambda_2 = 20$, and $\lambda_3 = 10$. The corresponding eigenvectors are

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{x}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \ \mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

- Consider a vector $\mathbf{v} = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$. We can always write $\mathbf{v}$ as a linear combination of the eigenvectors of S.

$$\mathbf{v} = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix} = 2\mathbf{x}_1 + 4\mathbf{x}_2 + 6\mathbf{x}_3.$$

- If we multiply $\mathbf{v}$ by S,

$$\begin{aligned} S\mathbf{v} &= S(2\mathbf{x}_1 + 4\mathbf{x}_2 + 6\mathbf{x}_3) \\ &= 2S\mathbf{x}_1 + 4S\mathbf{x}_2 + 6S\mathbf{x}_3 \\ &= 2\lambda_1\mathbf{x}_1 + 4\lambda_2\mathbf{x}_2 + 6\lambda_3\mathbf{x}_3 \\ &= 60\mathbf{x}_1 + 80\mathbf{x}_2 + 6\mathbf{x}_3. \end{aligned}$$

- Some observations:
  - Since x_i is an eigenvector, the output direction after multiplying by S is unchange: $S\mathbf{x}_i = \lambda_i \mathbf{x}_i$.
  - But the magnitude is changed, by the factor λ_i.
  - If λ_i is small, the output in the direction $\mathbf{x}_i$ is small. If we set it to zero, the resulting vector would still be a good approximation to the original vector.

- **Theorem 1:** (SSBD Theorem C.1 (Spectral Decomposition)) If $A \in \mathbb{R}^{d,d}$ is a symmetric matrix of rank k, then there exists an orthonormal basis of $\mathbb{R}^d$, $\mathbf{u}_1, \ldots, \mathbf{u}_d$, such that each $\mathbf{u}_i$ is an eigenvector of A.
  - A can be written as $A = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$, where each λ_i is the eigenvalue corresponding to the eigenvector $\mathbf{u}_i$.
  - This can be written equivalently as $A = U \Sigma U^T$, where the columns of U are the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_d$, and Σ is a diagonal matrix with $\Sigma_{i,i} = \lambda_i$.

- Continued ...
  - The number of λ_i which are nonzero is the rank of the matrix.
    - The eigenvectors which correspond to the nonzero eigenvalues span the range of A.
    - The eigenvectors which correspond to zero eigenvalues span the null space of A.

Outline

Positive Definite Matrices

- A symmetric matrix $A \in \mathbb{R}^{d,d}$ is positive definite if all its eigenvalues are positive. It is positive semidefinite if all its eigenvalues are non-negative.
- **Theorem 2:** (SSBD Theorem C.2) Let $A \in \mathbb{R}^{d,d}$ be a symmetric matrix. Then, the following are equivalent definitions of positive semidefiniteness of A:
  - All the eigenvalues of A are nonnegative.
  - For every vector $\mathbf{u}$, $\langle \mathbf{u}, A\mathbf{u} \rangle \geq 0$.
  - There exists a matrix B such that $A = BB^T$.

Outline

Singular Value Decomposition

- Let $A \in \mathbb{R}^{m,d}$ be a matrix of rank r.
- Unit vectors $\mathbf{v} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathbb{R}^m$ are called the right and left singular vectors of A with corresponding singular value $\sigma > 0$ if

$$A\mathbf{v} = \sigma\mathbf{u} \text{ and } A^T\mathbf{u} = \sigma\mathbf{v}.$$

- First we show that if we can find r orthonormal singular vectors with positive singular values, then we can decompose $A = U\Sigma V^T$ with U containing the left singular vectors, V the right singular vectors, and Σ is a $r \times r$ diagonal matrix with the singular values in the diagonal.

Lemma 3: (SSBD Lemma C.3) Let $A \in \mathbb{R}^{m,d}$ be a matrix of rank r. Assume that $\mathbf{v}_1, \ldots, \mathbf{v}_r$ is an orthonormal set of right singular vectors of A, $\mathbf{u}_1, \ldots, \mathbf{u}_r$ is an orthonormal set of corresponding left singular vectors of A, and $\sigma_1, \ldots, \sigma_r$ are the corresponding singular values. Then,

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T .$$

It follows that if U is a matrix whose columns are the $\mathbf{u}_i$'s, V is a matrix whose columns are the $\mathbf{v}_i$'s, and Σ is a diagonal matrix with $\Sigma_{i,i} = \sigma_i$, then

$$A = U \Sigma V^T .$$

Proof:

- From $A^T\mathbf{u} = \sigma\mathbf{v}$, any right singular vector of A must be in the range of A^T.
- Therefore $\mathbf{v}_1, \ldots, \mathbf{v}_r$ is an orthonormal basis for the range of A^T.
- Complete it into an orthonormal basis of $\mathbb{R}^d$ by adding the vectors $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_d$.
  - Note that these v's are orthogonal to the columns of A^T (rows of A).

- Let $B = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. Since $\mathbf{v}_i$ forms an orthonormal basis, showing $A\mathbf{v}_j = B\mathbf{v}_j$ for all j is sufficient to show $A = B$.
- If $j > r$, $A\mathbf{v}_j = 0$ as $\mathbf{v}_j$ is orthogonal to the rows of A. Similarly $B\mathbf{v}_j = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_j = 0$ due to the orthogonality of $\mathbf{v}_i$ with $\mathbf{v}_j$.
- For $j \leq r$

$$B\mathbf{v}_j = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_j = \sigma_j \mathbf{u}_j = A\mathbf{v}_j$$

from the definition of left and right singular vectors. □

We relate the singular values of A to the eigenvalues of A^TA and AA^T.

Lemma 4: (SSBD Lemma C.4) $\mathbf{v}, \mathbf{u}$ are right and left singular vectors of A with singular value σ iff $\mathbf{v}$ is an eigenvector of A^TA with corresponding eigenvalue σ^2 and $\mathbf{u} = \sigma^{-1}A\mathbf{v}$ is an eigenvector of AA^T with corresponding eigenvalue σ^2

Proof:

- Suppose σ is a singular value of A with right and left singular values $\mathbf{v}$ and $\mathbf{u}$. Then

$$A^T A\mathbf{v} = \sigma A^T \mathbf{u} = \sigma^2 \mathbf{v},$$

and

$$AA^T \mathbf{u} = \sigma A\mathbf{v} = \sigma^2 \mathbf{u}.$$

- For the other direction, if $\lambda \neq 0$ is an eigenvalue of $A^T A$ with corresponding eigenvector $\mathbf{v}$ then $\lambda > 0$ because $A^T A$ is positive semidefinite.
- Let $\sigma = \sqrt{\lambda}$. We first show that $\mathbf{u} = \sigma^{-1} A\mathbf{v}$ is the corresponding eigenvector for AA^T.

$$AA^T\mathbf{u} = AA^T A\mathbf{v}/\sigma = A\sigma^2\mathbf{v}/\sigma = \sigma^2\mathbf{u}.$$

- Hence

$$A\mathbf{v} = \sigma\mathbf{u},$$

and

$$A^T\mathbf{u} = \frac{1}{\sigma}A^T A\mathbf{v} = \frac{\sigma^2}{\sigma}\mathbf{v} = \sigma\mathbf{v}.$$

□

We can now use SSBD Lemma C.4 to prove that a matrix A always has decomposition $A = U\Sigma V^T$.

Corollary 5: (SSBD Corollary C.6 (The SVD theorem))
Let $A \in \mathbb{R}^{m,d}$ with rank r. Then $A = U\Sigma V^T$ where Σ is an $r \times r$ matrix with nonzero singular values of A and the columns of U, V are orthonormal left and right singular vectors of A. Furthermore, for all i, $\Sigma_{i,i}^2$ is an eigenvalue of A^TA, the i-th column of V is the corresponding eigenvector of A^TA and the i-th column of U is the corresponding eigenvector of AA^T.

Proof:

- We will argue that the rank of $A^T A$ is r when the rank of A is r later.
- Given that $A^T A$ is a symmetric positive semidefinite matrix with rank r, SSBD Theorem C.2 states that there is a decomposition $A^T A = V \Sigma V^T$ with r nonzero (positive) eigenvalues of $A^T A$ corresponding to r eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ of $A^T A$.
- Using SSBD Lemma C.4, we find that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are left and right singular vectors corresponding to singular values $\sigma_1, \ldots, \sigma_r$.
- Applying SSBD Lemma C.3 completes the proof.

- We now argue that the rank of $A^T A$ is the same as the rank of A.
  - Let $\mathbf{x}$ be in the null space of A. This means that $A\mathbf{x} = 0$ implying that $A^T A\mathbf{x} = 0$. So the null space of A is a subset of the null space of $A^T A$.
  - Now, let $\mathbf{x}$ be in the null space of $A^T A$. This means that $A^T A\mathbf{x} = 0$ which implies that $\mathbf{x}^T A^T A\mathbf{x} = 0$ or $\|A\mathbf{x}\|^2 = 0$. So $A\mathbf{x}$ must equal 0, implying that the null space of $A^T A$ is a subset of the null space of A.

□

We now consider approximating the matrix.
The Frobenius norm of a matrix A is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{d} a_{ij}^2}.$$

Lemma 6: Let $A \in \mathbb{R}^{m,d}$ with rank r. If $\mathbf{v}_1, \ldots, \mathbf{v}_r$ is an orthonormal set of right singular vector of A corresponding to $\sigma_1 \geq \sigma_2 \geq \ldots, \sigma_r$, then

$$\mathbf{v}_1 = \operatorname{argmax}_{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\| = 1} \|A\mathbf{v}\|$$

$$\mathbf{v}_2 = \operatorname{argmax}_{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\| = 1, \langle \mathbf{v}, \mathbf{v}_1 \rangle = 0} \|A\mathbf{v}\|$$

$$\vdots$$

$$\mathbf{v}_r = \operatorname{argmax}_{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v}\| = 1, \forall i < r, \langle \mathbf{v}, \mathbf{v}_i \rangle = 0} \|A\mathbf{v}\|.$$

Proof:

- Let $V \in \mathbb{R}^{d,r}$ be the matrix of right singular vectors corresponding to $\sigma_1 \geq \sigma_2 \geq \ldots, \sigma_r$.
- It suffices to consider vectors $\mathbf{v}$ in the span of the columns of V: any component v' orthogonal to V has $Av' = U\Sigma V^T v' = 0$. We can hence write $\mathbf{v} = V\mathbf{x}$.
- From the orthonormality of U and V

$$\|A\mathbf{v}\|^2 = \|AV\mathbf{x}\|^2 = \|U\Sigma V^T V\mathbf{x}\|^2 = \|U\Sigma\mathbf{x}\|^2$$
$$= \|\Sigma\mathbf{x}\|^2 = \sum_{i=1}^{n} \sigma_i^2 x_i^2.$$

Note that $\|\mathbf{x}\| = 1$ as $\|\mathbf{v}\| = 1$. Furthermore, $\langle \mathbf{v}_i, \mathbf{v} \rangle = 0$ corresponds to $x_i = 0$. When $\|\mathbf{v}\| = 1$, maximizing $\|A\mathbf{v}\|^2$ subject to $\langle \mathbf{v}_j, \mathbf{v} \rangle = 0 \ \forall j < i$ gives $\mathbf{x} = e_i$, the unit vector with 1 in position i and zero elsewhere. This implies that the solution is the i-th singular vector of A. □

Lemma 7: Given a rank r matrix A and an orthonormal set of r vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ that span the rows of A, the Frobenius norm of A satisfies $\|A\|_F^2 = \sum_{j=1}^{r} \|A\mathbf{v}_i\|^2$. Furthermore, if $A = A_1 + A_2$ with $\mathbf{v}_1, \ldots, \mathbf{v}_k$ spanning the rows of A_1 and $\mathbf{v}_{k+1}, \ldots, \mathbf{v}_r$ spanning the rows of A_2, for $k < r$, then $\|A_1 + A_2\|_F^2 = \|A_1\|_F^2 + \|A_2\|_F^2$. If the orthonormal set is the set of singular vectors, then $\|A\mathbf{v}_j\|^2 = \sigma_j^2$, where $\sigma_1, \ldots, \sigma_r$ are the singular values.

Proof:

- Assume that the matrix $A \in \mathbb{R}^{m,d}$ represents m data points in d dimension where each point i, $\mathbf{a}_i$ is represented as a row i in the matrix.
- $\langle \mathbf{a}_i, \mathbf{v}_j \rangle$ is then the coefficient of projection of $\mathbf{a}_i$ in the direction of $\mathbf{v}_j$.
- Similarly, $A\mathbf{v}_j$ is a vector of the projection coefficients of all the data points in the direction of $\mathbf{v}_j$.

- Because $\mathbf{v}_1, \ldots, \mathbf{v}_r$ is orthonormal and spans all the rows of A, we can write $\mathbf{a}_i = V\mathbf{b}$. As V is orthonormal, $\|\mathbf{a}_i\| = \|\mathbf{b}\|$.
- Furthermore, the j-th coefficient of $\mathbf{b}$ is $\langle \mathbf{a}_i, \mathbf{v}_j \rangle$ giving $\|\mathbf{a}_i\|^2 = \sum_j \langle \mathbf{a}_i, \mathbf{v}_j \rangle^2$.
- To compute the square of the Frobenius norm, we can sum the coefficients over all the rows

$$\|A\|_F^2 = \sum_{i=1}^{m} \|\mathbf{a}_i\|^2 = \sum_{i=1}^{m} \sum_{j=1}^{r} \langle \mathbf{a}_i, \mathbf{v}_j \rangle^2 = \sum_{j=1}^{r} \sum_{i=1}^{m} \langle \mathbf{a}_i, \mathbf{v}_j \rangle^2$$
$$= \sum_{j=1}^{r} \|A\mathbf{v}_j\|^2.$$

- If $A = A_1 + A_2$ as described, the squared norm of each row can also be decomposed into two orthogonal components $\|\mathbf{a}_{1i} + \mathbf{a}_{2i}\|^2 = \sum_{j=1}^{k} \langle \mathbf{a}_{1i}, \mathbf{v}_j \rangle^2 + \sum_{j=k+1}^{r} \langle \mathbf{a}_{2i}, \mathbf{v}_j \rangle^2$ giving

$$\|A_1 + A_2\|_F^2 = \sum_{j=1}^{k} \|A_1 \mathbf{v}_j\|^2 + \sum_{j=k+1}^{r} \|A_2 \mathbf{v}_j\|^2.$$

- Consequently,

$$\|A_1 + A_2\|_F^2 = \|A_1\|_F^2 + \|A_2\|_F^2.$$

- From the properties of the singular vectors, $A\mathbf{v}_j = \sigma_j \mathbf{u}_j$. So $\|A\mathbf{v}_j\|^2 = \sigma_j^2 \|\mathbf{u}_j\|^2 = \sigma_j^2$. □

- From the properties of singular vectors, $A\mathbf{v}_j = \sigma_j \mathbf{u}_j$ gives the vector of projection coefficients. Hence $\sigma_j \mathbf{u}_j \mathbf{v}_j^T$ gives the projection of the matrix in the direction $\mathbf{v}_j$, i.e. row i of $\sigma_j \mathbf{u}_j \mathbf{v}_j^T$ is the projection of data point i onto the vector $\mathbf{v}_j$.
- Summing up over the first k singular vectors

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

we get the projection of A onto the space spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

Theorem 8: A_k gives the best rank k approximation to A in the Frobenius norm with $\|A - A_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$.

Proof:

From Lemma 7, $\|A - A_k\|_F^2 = \|A\|_F^2 - \|A\mathbf{w}_1\|^2 - \cdots - \|A\mathbf{w}_k\|^2$ for an orthonormal basis $\mathbf{w}_1, \ldots, \mathbf{w}_k$.

Hence, we seek an orthonormal basis that maximizes $\|A\mathbf{w}_1\|^2 + \cdots + \|A\mathbf{w}_k\|^2$.

We show that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ maximizes the expression by induction:

- The base case of rank 1 is true by Lemma 6.
- Assume that this is true for the first k basis and consider the best $k+1$ bases. Let W be the $k+1$ dimensional subspace that gives the best approximation. Assume bases $(\mathbf{w}_1, \ldots, \mathbf{w}_{k+1})$ for W.

- $\mathbf{w}_{k+1}$ can be chosen to be orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_k$ since W is of higher dimension than k: choose a vector that is orthogonal to the projections of $\mathbf{v}_1, \ldots, \mathbf{v}_k$ onto W.
- By Lemma 6, $\mathbf{v}_{k+1}$ maximizes $\|A\mathbf{v}\|^2$ among all unit vectors orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Hence $\|A\mathbf{w}_{k+1}\|^2 \leq \|A\mathbf{v}_{k+1}\|^2$. By induction hypothesis, we also have

$$\|A\mathbf{w}_1\|^2 + \cdots + \|A\mathbf{w}_k\|^2 \leq \|A\mathbf{v}_1\|^2 + \cdots + \|A\mathbf{v}_k\|^2,$$

  giving

$$\|A\mathbf{w}_1\|^2 + \cdots + \|A\mathbf{w}_{k+1}\|^2 \leq \|A\mathbf{v}_1\|^2 + \cdots + \|A\mathbf{v}_{k+1}\|^2.$$

- From Lemma 7, $\|A - A_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$. □

Another matrix norm of interest is the 2-norm or spectral norm:

$$\|A\|_2 = \max_{\mathbf{x}:\|\mathbf{x}\|\leq 1} \|A\mathbf{x}\|.$$

From Lemma 6, the 2-norm of a matrix is equal to its first singular value σ_1.

References

1. The material is mostly taken from Appendix C of SSBD.

References I

[1] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*. Vol. 1. 1. Cambridge university press Cambridge, 2008.