Machine Learning Foundations
Week 6

Shubharup G.
BSDS, IIT Madras

March 2024

Contents

1 Introduction

1.1 Recap: Eigendecomposition (aka Diagonalization)

So far we have learnt eigendecomposition, which the standard sense of the term 'diagonalization'.

Eigendecomposition is one of the many known matrix factorization techniques [1].

The two major motivations, by way of utility, which we've seen for diagonalization (eigendecomposition) are:

1. Interpretability: what is happening when we apply the matrix? decomposes what looks like an atomic task into 3 discrete subtasks:

   I change of basis aka rotation

   II disproportional scaling

   III rotation back to original basis

2. Economy of computation: reducing the compute of applying the same matrix repeatedly like $A(A(x)) : A^n(x) = P^{-1}\Lambda^n P$.

   Raising a diagonal matrix to a power n is really easy:
   raise the diagonal entries to the power n.

   e.g. for a numerical example, if $A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$, can immediately read off $A^4 = \begin{bmatrix} 16 & 0 \\ 0 & 81 \end{bmatrix}$

1.2 Uses of Singular Value Decomposition

Interpretability is an overlapping utility for SVD, but it's even better because any real matrix has an SVD: whether square or rectangular, whether with or without an eigenbasis.

SVD factorizes the application of the matrix into three discrete steps or subtasks, much like diagonalization.

SVD might not have use in reducing compute.

But SVD, because of its generality of application to any real matrix, has a prototypical use-case for which it is particularly well-known: loss-minimizing compression. By compression, SVD can reduce the storage requirement for archiving a dataset, while retaining the most important information.

In fact, the Eckhart-Young theorem[2], also simply known as 'the SVD theorem', guarantees that the low-rank approximation gotten using SVD is the best approximation – the approximated matrix has the least loss of all possible choices for approximation, as measured by the Frobenius norm[3] of the difference between the original matrix and the approximated matrix.

[1] Matrix decomposition (Wikipedia)
[2] Low-rank approximation (Wikipedia)
[3] Sum of squares of the entries in the matrix – essentially similar to the Euclidean or L_2 norm for vectors. See Frobenius norm (Wikipedia).

2 Singular Value Decomposition

2.1 Form and Properties of SVD

SVD is a matrix factorization technique [4] canonically written as:

$$\underbrace{A}_{m\times n} = \underbrace{U}_{m\times m}\underbrace{\Sigma}_{m\times n}\underbrace{V}_{n\times n}^{T}$$

- $U == \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_m \\ | & | & & | \end{bmatrix}$.

  The column-vectors of U viz. u_j are known as the ***left singular vectors***.

- $\Sigma = \begin{cases} \left[\begin{array}{cccc|ccc} \sigma_1 & & & & | & | & | \\ & \sigma_2 & & & | & | & | \\ & & \sigma_3 & & \mathbf{0} & \dots & \mathbf{0} \\ & & & \ddots & | & | & | \\ & & & \sigma_\omega & | & | & | \end{array}\right] & \text{if } m < n; \\ \left[\begin{array}{c} \sigma_1 \\ \quad \sigma_2 \\ \quad\quad \sigma_3 \\ \quad\quad\quad \ddots \\ \hline \quad\quad\quad\quad \sigma_\omega \\ \hline \text{———} 0 \text{———} \\ \text{———} \vdots \text{———} \\ \text{———} 0 \text{———} \end{array}\right] & \text{if } m > n; \end{cases}$

  Σ is in general a ***rectangular diagonal matrix***: there is a padding of zero columns if $m < n$, a padding of zero rows if $m > n$. It is a square matrix iff. $m = n$.

  The diagonal entries in the square diagonal partition viz. the σ_j are called the ***singular values***. The count of singular values is $\omega = min(m, n)$.

- $V == \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \dots & v_n \\ | & | & & | \end{bmatrix}$, or $V^T = \begin{bmatrix} — & v_1^T & — \\ — & v_2^T & — \\ & \vdots & \\ — & v_n^T & — \end{bmatrix}$.

  The column vectors of V viz. v_j are known as the ***right singular vectors***.

[4] One among others, to restate.

Properties:

- $\boxed{U \textbf{ and } V \textbf{ are orthogonal matrices.}}$
  - $UU^T = U^TU = I$
  - $VV^T = V^TV = I$

  To recap:

  - Orthogonal matrices have orthonormal columns.

    (Equivalently, they have orthonormal rows.)
  - The inverse of an orthogonal matrix P is its transpose: $P^{-1} = P^T$.
  - An orthogonal matrix P has determinant $det(P) = 1$.

  This is important for relevance here, because **every orthogonal matrix is a change of basis matrix** between some pair of basis frames. This aspect will be revisited when we see SVD of a matrix A as a logical decomposition of the linear transformation represented by A – from what appears to be an atomic task, into the sequential composition of three sub-tasks.

- Similar to eigendecomposition, the order of $\begin{bmatrix} | \\ u_j \\ | \end{bmatrix}$ *and* $\begin{bmatrix} | \\ v_j \\ | \end{bmatrix}$, in U and V respectively, is coupled with the relative order of σ_j in Σ.

  Therefore,

  $\boxed{\begin{array}{l}\text{SVD of a matrix } A \text{ is unique } \textit{up to} \text{ rearrangement} \\ \text{or } \textit{permutation} \text{ of the singular values in } \Sigma.\end{array}}$

2.2 Computing SVD

2.2.1 First method – directly, as defined

- **The left singular vectors viz. the column vectors** $\begin{bmatrix} | \\ u_i \\ | \end{bmatrix}$ **in** U **are the *normalized* eigenvectors of** AA^T.

  – For $shape(A) = m \times n$, $shape(AA^T) = m \times m$.

  – AA^T is a (real) symmetric matrix. By virtue of the Spectral Theorem, this guarantees that it has an orthogonal eigenbasis i.e. *a complete set of m orthogonal eigenvectors*, which can be normalized to be made orthonormal. These go into U as the column vectors.

- **The right singular vectors viz. the column vectors** $\begin{bmatrix} | \\ v_i \\ | \end{bmatrix}$ **in** V **or likewise the row vectors** $\begin{bmatrix} — \, v_i^T \, — \end{bmatrix}$ **in** V^T **are the *normalized* eigenvectors of** A^TA.

  – For $shape(A) = m \times n$, $shape(A^TA) = n \times n$.

  – A^TA is a (real) symmetric matrix. By virtue of the Spectral Theorem, this guarantees that it has an orthogonal eigenbasis i.e. *a complete set of n orthogonal eigenvectors*, which can be normalized to be made orthonormal. These go into V as the column vectors.

- **The singular values** $\sigma_i = \sqrt[+]{\lambda_i}$ **where:**

$$\{\lambda_i\} = \begin{cases} \Lambda(AA^T) & \textbf{if } m < n; \\ \Lambda(A^TA) & \textbf{if } m > n \end{cases}$$

where $\Lambda(M)$ **denotes the multiset[5] of eigenvalues of the matrix** M, aka its *spectrum*.[6]

[5] Extension of the idea of a set to allow for multiplicity of the same element: here, viz. the include repeated eigenvalues. See Multiset (Wikipedia).

[6] Spectrum of a matrix (Wikipedia)

– Can show that for a square matrix M, $\Lambda(M) = \{\hat{v_j}^T M \hat{v_j}\}_j$ for all linearly independent and normalized eigenvectors $\hat{v_j}$ of M.

So also: $\Lambda(M^T M) = \{\hat{w_j}^T (M^T M) \hat{w_j}\}_j$ where $\hat{w_j}$ are the linearly independent and normalized eigenvectors of $M^T M$.

$$\Lambda(M^T M) = \{\hat{w_j}^T (M^T M) \hat{w_j}\}_j$$
$$= \{(M\hat{w_j})^T (M\hat{w_j})\}_j$$
$$= \{y^T y \text{ or } ||y||^2, \text{ where } y = M\hat{w_j}\}_j$$

We know that norm of a vector, by properties, is non-negative. This means the spectrum $\Lambda(M^T M)$ has only non-negative entries[7].

* Can show the same for MM^T also: take $C = M^T$ and so $MM^T = C^T C$ is positive semi-definite.

Therefore, the singular values (positive square roots of non-negative real numbers) are guaranteed to be real.

– The singular values are all positive (by definition), and conventionally placed in descending order of magnitude. The reason for the latter will become apparent in both later sections on SVD for Compression, and SVD for Interpretation.

– It is a known result that **the set of non-zero eigenvalues of AA^T and $A^T A$ are exactly the same:**

$$\underbrace{(\lambda_i \neq 0, r) \in \Lambda(AA^T)}_{\substack{\lambda_i \text{ is a non-zero eigenvalue of } AA^T \\ \text{with multiplicity } r}} \iff \underbrace{(\lambda_i \neq 0, r) \in \Lambda(A^T A)}_{\substack{\lambda_i \text{ is a non-zero eigenvalue of } A^T A \\ \text{with multiplicity } r}}$$

The 'extra' eigenvalues will all be zeroes:

$$\begin{cases} (n-m) \text{ in } \underbrace{\Lambda(\overbrace{A^T A}^{n \times n}) - \Lambda(\overbrace{AA^T}^{m \times m})}_{\Lambda(AA^T) \subset \Lambda(A^T A)} & \text{if } n > m \\ (m-n) \text{ in } \underbrace{\Lambda(\overbrace{AA^T}^{m \times m}) - \Lambda(\overbrace{A^T A}^{n \times n})}_{\Lambda(A^T A) \subset \Lambda(AA^T)} & \text{if } m > n \end{cases}$$

[7] i.e. $M^T M$ **is *positive semi-definite.***

2.2.2 Second method – using inter-relationships (faster)

Let $\omega = min(m, n)$.

Then the left-singular vectors $\begin{bmatrix} | \\ u_j \\ | \end{bmatrix}$

and the right-singular vectors $\begin{bmatrix} | \\ v_j \\ | \end{bmatrix}$

are related as:

$$\left.\begin{aligned} v_j &= \frac{A^T u_j}{\sigma_j} \quad \dots \quad (i) \\ u_j &= \frac{A v_j}{\sigma_j} \quad \dots \quad (ii) \end{aligned}\right\} j \in \{1, 2 \ \dots \ \omega\}$$

The proof is reserved for the appendix.

So then,

$$\begin{cases} \text{If } \underbrace{n > m}_{\text{'broad' matrix}} : & \bullet \text{ Compute the eigenvalues and normalized eigenvectors of } AA^T \ (\equiv m \times m) \\ & \text{– respectively providing the singular values } \{\sigma_j\} \text{ and left singular vectors } \{u_j\}. \\ & \bullet \text{ Use } (i) \text{ to get the corresponding right singular vectors } \{v_j\}. \\ & \bullet \text{ Extend } V \text{ to an an orthonormal basis of } \mathbb{R}^n \text{ by the Gram–Schmidt process.} \\ \text{If } \underbrace{m > n}_{\text{'tall' matrix}} : & \text{Compute the eigenvalues and normalized eigenvectors of } A^T A \ (\equiv n \times n) \\ & \text{– respectively providing the singular values } \{\sigma_j\} \text{ and right singular vectors } \{v_j\}. \\ & \bullet \text{ Use } (ii) \text{ to get the corresponding left singular vectors } \{u_j\}. \\ & \bullet \text{ Extend } U \text{ to an an orthonormal basis of } \mathbb{R}^m \text{ by the Gram–Schmidt process.} \end{cases}$$

3 SVD for Compression

3.1 Dyadic product (aka Outer product)

Thus far, we have learnt about the *inner product* of two vectors in a vector space V.

The inner of product in a vector space V is a bivariate function (i.e. with two arguments) which maps a pair of vectors $a, b \in V$ to a member item of the field underlying V, additionally satisfying some pre-requisite properties.

Prior to Week 5, our vector spaces were always over the field of real numbers, $V \subseteq \mathbb{R}^n$. In Week 5, we expanded the underlying field to the set of the set of complex numbers, $V \subseteq \mathbb{C}^n$ and looked at some interesting generalizations of entities in the real case (e.g. Hermitian matrices for real symmetric matrices, unitary matrices for real orthogonal matrices, etc.).

In Week 6 and onwards, we will again assign ourselves to only working with set of reals $\mathbb{R}$ as the underlying field. Further restatement of this will be excluded for the sake of brevity and avoiding repetition.

So an inner product is a function:

$$\langle , \rangle : V \times V \to \mathbb{R}$$

For example, the *standard inner product*:

$$\langle a, b \rangle = a^T b$$

Importantly, the inner product outputs a scalar.

One among several vector multiplication operations[8] with accepted usage, relevant for us here is the **outer product**:

$$\boxed{u \otimes v = uv^T}$$

- If $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, then the (result of [9]) **outer product is a matrix** $\mathbb{M}_{m \times n}(\mathbb{R})$.
  - Contrast this with the inner product of two vectors, which outputs a scalar. Although, there is the possibly meaningful connection that if both $u, v \in \mathbb{R}^n$, then the inner product is the trace of the outer product: $\langle u, v \rangle = tr(u \otimes v)$.
- Such a matrix is also called a **dyadic matrix** or simply a ***dyad***, following that the outer product is also known as the *dyadic product*.

  For non-zero vectors u, v, the outer product is a dyad of rank = 1.

  To see this:
  - The i^{th} row of uv^T is v^T scaled by the the i^{th} entry of $\vec{u}$.
  - if $\vec{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$, then $uv^T = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} — & v^T & — \end{bmatrix} = \begin{bmatrix} — & (u_1)v^T & — \\ — & (u_2)v^T & — \\ & \vdots & \\ — & (u_n)v^T & — \end{bmatrix}$

[8] Vector multiplication (Wikipedia)

[9] The term 'outer product' can variously refer to the operation ($\otimes$) itself, or what the operation produces. The sense is usually apparent enough from the context.

– So the rows in uv^T are all just scalar multiples of the same vector viz. v^T, so (for $\vec{u}, \vec{v} \neq \vec{0}$) the $rank(uv^T) = 1$.

– Then for a non-zero scalar $c \in \mathbb{R}$, $c(u \otimes v) = c(uv^T)$ is also of rank = 1.

Examples:

1. Recall the projection matrix $\mathbb{P}_a$ for a (non-zero) vector $\vec{a}$, $\mathbb{P}_a = \dfrac{aa^T}{a^T a}$. It is a indeed a dyadic matrix, with $rank(\mathbb{P}_a) = 1$.

3.2 SVD as sum of dyads

You can verify that the SVD factorization of a matrix:

$$\begin{bmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_m \\ | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & \overbrace{| & | & |}^{max(n-m,0)\text{ columns}} \\ & \sigma_2 & & & | & | & | \\ & & \sigma_3 & & \mathbf{0} & \dots & \mathbf{0} \\ & & & \ddots & | & | & | \\ & & & \sigma_\omega & | & | & | \\ \multicolumn{2}{l}{max(m-n,0)\text{ rows}\left\{\right.} & \mathbf{0} & & & & \\ & & \vdots & & & & \\ & & \mathbf{0} & & & & \end{bmatrix} \begin{bmatrix} — & v_1^T & — \\ — & v_2^T & — \\ & \vdots & \\ — & v_n^T & — \end{bmatrix}$$

can be multiplied and produced out as the sum:

$$\boxed{\begin{aligned} A &= U\Sigma V^T \\ &= \sum_{j=1}^{\omega} \sigma_j u_j v_j^T \\ &= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T \dots + \sigma_\omega u_\omega v_\omega^T \end{aligned}}$$

That is, A is *exactly* the sum of a finite number of such dyads, whose count is $\leq \omega$ in general (as it might be that some of the singular values $\sigma_j = 0$).

3.3 Truncated k-sum as the best k-rank approximation

We had said earlier that the singular values σ_j should be listed in descending order of magnitude in Σ.

This is because it directly ties with the following utility, stated here as a fact:

With the singular values in Σ listed as $\sigma_n \geq \sigma_{n+1}$ —

the best[a] k-rank approximation of A is given by:

$$\tilde{A}_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T$$
$$= \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T \ldots \sigma_k u_k v_k^T$$

[a] By the Eckart-Theorem. See Low-rank approximation (Wikipedia)

- A note on the *slight* ambiguity of terminology here.

  Here, $k-$rank does not mean rank *exactly* equal to k , but **rank at most** k i.e. $rank(\tilde{A}_k) \leq k$.

  - A property of $rank(.)$ is that **rank is sub-additive**:

    $rank(A + B) \leq rank(A) + rank(B)$

    and in general: $rank(\sum_{j=1}^{S} M_j) \leq \sum_{j=1}^{S} rank(M_j)$.

    * In particular here,

    $$\begin{aligned} rank(\tilde{A}_k) &= rank(\sum_{j=1}^{k} \sigma_j u_j v_j^T) \\ &\leq \sum_{j=1}^{k} rank(\sigma_j u_j v_j^T) \\ &= \underbrace{1 + 1 + \ldots + 1}_{k \text{ times}} \\ &= k \\ \therefore rank(\tilde{A}_k) &\leq k \end{aligned}$$

  - So if $k = \omega$ i.e. 'no approximation':

    $rank(A) \leq \omega$

    i.e. $rank(A) \leq min(m, n)$

4 Interpreting a Linear Transformation with SVD

To be added.

References

Wikipedia

Wikipedia contributors. *Spectrum of a matrix (Wikipedia)*. URL: https://en.wikipedia.org/w/index.php?title=Spectrum_of_a_matrix.

Wikipedia contributors. *Low-rank approximation (Wikipedia)*. URL: https://en.wikipedia.org/wiki/Low-rank_approximation#Basic_low-rank_approximation_problem.

Wikipedia contributors. *Matrix decomposition (Wikipedia)*. URL: https://en.wikipedia.org/w/index.php?title=Matrix_decomposition.

Wikipedia contributors. *Frobenius norm (Wikipedia)*. URL: https://en.wikipedia.org/wiki/Matrix_norm#Frobenius_norm.

Wikipedia contributors. *Multiset (Wikipedia)*. URL: https://en.wikipedia.org/w/index.php?title=Multiset&oldid=1201923950. [Online; accessed 19-March-2024].

Wikipedia contributors. *Vector multiplication (Wikipedia)*. URL: https://en.wikipedia.org/w/index.php?title=Vector_multiplication.