CSC 311: Introduction to Machine Learning

Tutorial - Matrix Decomposition & Probabilistic Models

TA: Vahid Balazadeh

Instructors: Michael Zhang and Chandra Gummaluru

University of Toronto

Based on slides by Haonan Duan

Matrix Decomposition

- We can decompose an integer into its prime factors, e.g., $12 = 2 \times 2 \times 3$.
- Similarly, matrices can be decomposed into product of other matrices.
- Examples are Eigendecomposition, SVD, Schur decomposition, LU decomposition, ...
- Here, we focus on Eigendecomposition and SVD

Eigenvector

- An eigenvector of a square matrix A is a nonzero vector v such that multiplication by A only changes the scale of v:

$$Av = \lambda v$$

- The scalar λ is known as the eigenvalue.
- If v is an eigenvector of A, so is any rescaled vector αv.
- αv has the same eigenvalue as v. Thus, we constrain the eigenvector to be of unit length.

Compute eigenvalues - characteristic polynomial

- Eigenvalue equation of matrix A:

$$\begin{aligned} Av &= \lambda v \\ \lambda v - Av &= 0 \\ (\lambda I - A)v &= 0 \end{aligned}$$

- If nonzero solution for v exists, then it must be the case that:

$$det(\lambda I - A) = 0$$

- Unpacking the determinant as a function of λ , we get a polynomial, called the characteristic polynomial:

$$P_A(\lambda) = det(\lambda I - A) = \lambda^n + c_{n-1}\lambda^{n-1} + \ldots + c_1\lambda + c_0$$

- Compute eigenvalues of $A \rightarrow$ solve $P_A(\lambda) = 0$

Exercise

Consider the matrix:

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

- What are the eigenvalues and eigenvectors of A?

Solution

We first need to calculate the eigenvalues,

$$\det(A - \lambda I) = 0 \implies \det \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = 0$$
$$\implies (2-\lambda)^2 - 1 = 0 \implies \lambda_1 = 3, \lambda_2 = 1$$

Then, we solve $(A - \lambda_i I)v_i = 0$ to find eigenvectors:

$$(A - \lambda_1 I)v_1 = 0 \implies \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} v_1 = 0$$
$$\implies v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \overset{\text{normalize}}{\implies} v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Similarly,

$$(A - \lambda_2 I)v_2 = 0 \implies v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Eigendecomposition

- **Spectral Theorem** - Every symmetric matrix $A \in \mathbb{R}^{n \times n}$ has a set of n orthonormal eigenvectors forming a basis. Furthermore, all eigenvalues are real.
- Therefore, A can be decomposed to the following form

$$A = PDP^{-1}$$

- P is an orthogonal matrix of the eigenvectors of A, and D is a diagonal matrix of eigenvalues.

Eigendecomposition

- **Spectral Theorem** - Every symmetric matrix $A \in \mathbb{R}^{n \times n}$ has a set of n orthonormal eigenvectors forming a basis. Furthermore, all eigenvalues are real.
- Therefore, A can be decomposed to the following form

$$A = PDP^{-1}$$

- P is an orthogonal matrix of the eigenvectors of A, and D is a diagonal matrix of eigenvalues.

$$\begin{aligned} A\underbrace{[v_1, \ldots, v_n]}_{P} &= [Av_1, \ldots, Av_n] \\ &= [\lambda_1 v_1, \ldots, \lambda_n v_n] \\ &= \underbrace{[v_1, \ldots, v_n]}_{P} \underbrace{\begin{bmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{bmatrix}}_{D} \end{aligned}$$

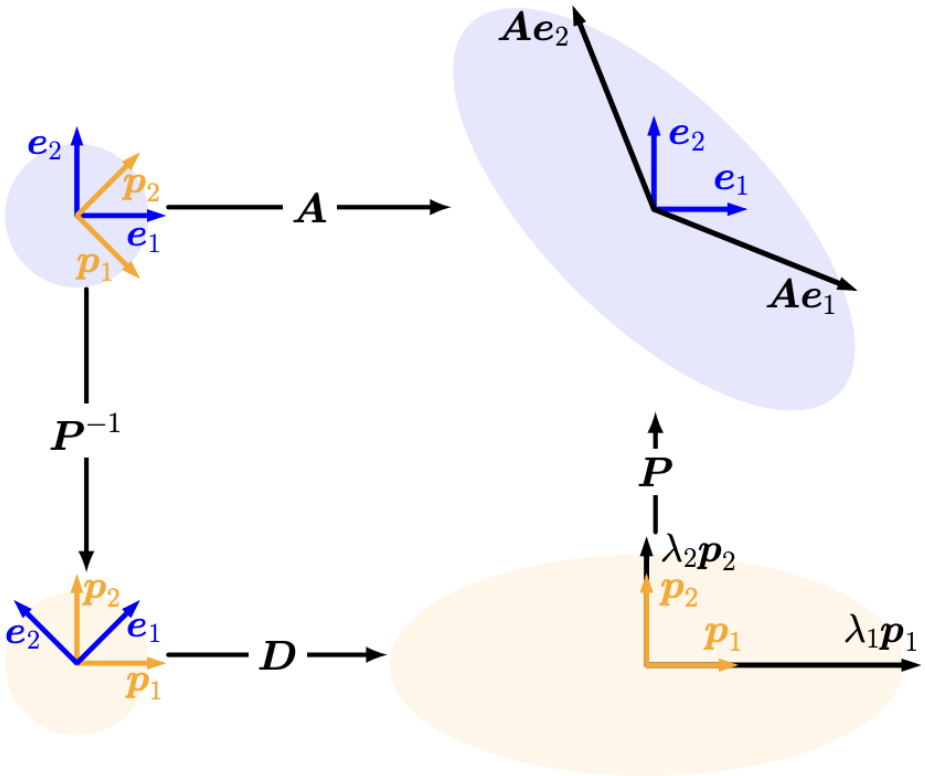
Intuitions of Eigendecomposition

- Diagonal matrix allows fast computations of their determinants, powers and inverses.
- Eigendecomposition transforms a matrix into a diagonal form by changing the basis.

$$\begin{aligned}\det(A) = \det(PDP^{-1}) &= \det(P)\det(D)\det(P)^{-1} \\ &= \det(D) \\ &= \prod_{i=1}^{n} \lambda_i\end{aligned}$$

$$A^{-1} = PD^{-1}P^{-1}$$

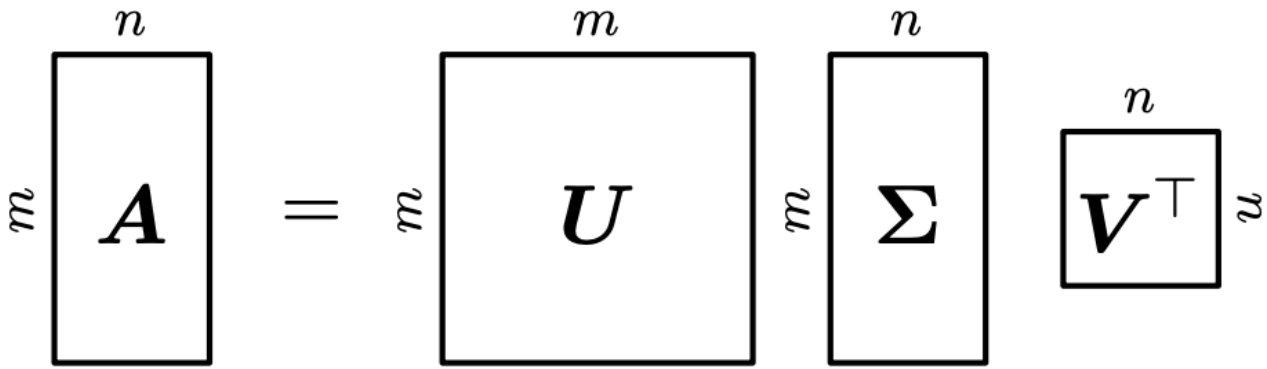
Geometric intuitions of eigendecomposition



- Top-left to bottom-left: P^{-1} performs a basis change.
- Bottom-left to bottom-right: D performs a scaling.
- Bottom-right to top-right: P undoes the basis change.

Singular Value Decomposition (SVD)

- If $A \in \mathbb{R}^{m \times n}$ is not square, eigendecomposition is undefined.
- SVD is a decomposition of the form $A = U\Sigma V^T$.
- SVD is more general than eigendecomposition. **Every** real matrix has a SVD.



SVD - Terminology

- U and V are orthogonal matrices, and Σ is a diagonal matrix (not necessarily square).
- Diagonal entries of Σ are called singular values of A.
- Columns of U are the left singular vectors, and columns of V are the right singular vectors.

SVD and eigendecomposition

- SVD can be interpreted in terms of eigendecomposition.
- Left singular vectors of A are the eigenvectors of AA^T.
- Right singular vectors of A are the eigenvectors of A^TA
- Nonzero singular values of A are square roots of eigenvalues of A^TA and AA^T. A^TA and AA^T are positive semi-definite (PSD), thus their eigenvalues are positive.

Informal Proof

Since $B = AA^\top \in \mathbb{R}^{m \times m}$ is symmetric, eigendecomposition holds

$$B = PDP^{-1}$$

Now, assume SVD exists, i.e., $A = U\Sigma V^\top$. Therefore,

$$B = AA^\top = (U\Sigma V^\top)(V\Sigma^\top U^\top) = U\Sigma\Sigma^\top U^\top$$

Matching those two:

$$PDP^{-1} = U\Sigma\Sigma^\top U^\top$$

Therefore, $U = P$ and $\Sigma \equiv D^{\frac{1}{2}}$ or $\sigma_i = \sqrt{d_i}$.

A similar approach on $C = A^\top A \in \mathbb{R}^{n \times n}$ leads to V.

Exercise

Compute SVD of the matrix:

$$A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$$

Solution

Here, we calculate U and Σ. First, define $B = AA^\top$

$$B = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} = \begin{bmatrix} 17 & 8 \\ 8 & 17 \end{bmatrix}$$

Then, we can calculate the eigenvalues and eigenvectors (using characteristic polynomial): $\lambda_1 = 25, \lambda_2 = 9$ and $v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Therefore, $B = PDP^{-1}$ where

$$P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 25 & 0 \\ 0 & 9 \end{bmatrix}$$

We had $U = P$ and $\Sigma \equiv D^{\frac{1}{2}}$:

$$\Sigma = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}$$

Find V for exercise.

Rank-r approximation

- Given a matrix A, SVD allows us to find its "best" rank-r approximation A_r $(r < n)$.
- Why? store less parameters
- We can write $A = U\Sigma V^\top$ as $A = \sum_{i=1}^{n} \sigma_i u_i v_i^\top$, where σ_i are sorted from the largest to the smallest.

Rank-r approximation

- The rank-r approximation A_r is defined as:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

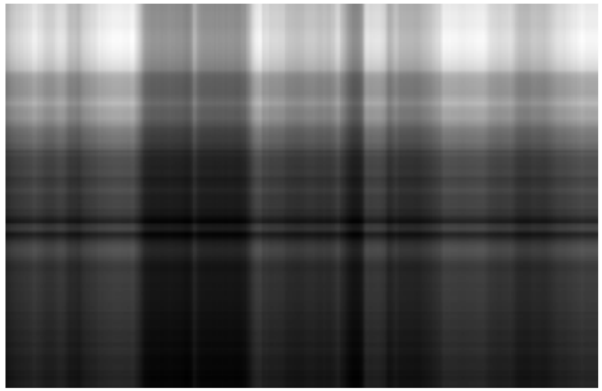
- A_r is the best approximation of rank r by many norms, such as spectral norm.

$$\|A\|_2 := \sup_x \frac{\|Ax\|_2}{\|x\|_2}$$

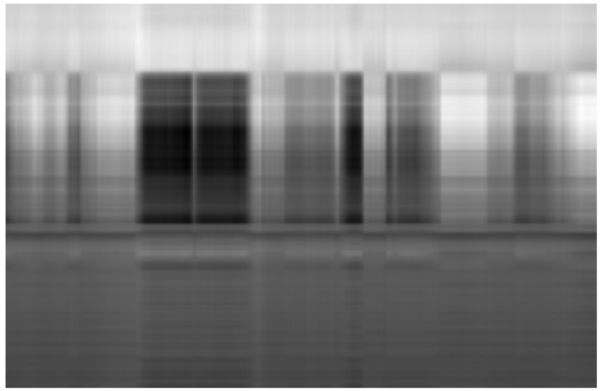
- It means that $\|A - A_r\|_2 \leq \|A - B\|_2$ for any rank r matrix B.

Rank-r approximation

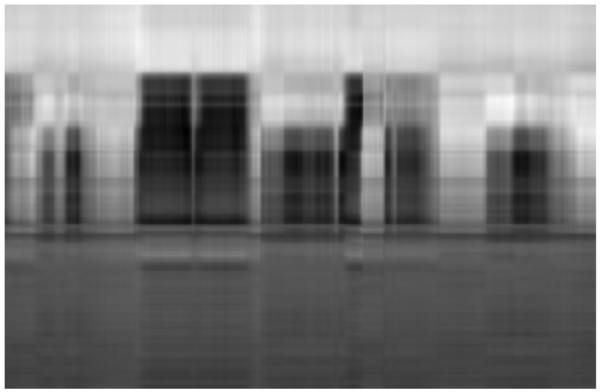
(a) Original image $\boldsymbol{A}$.



(b) Rank-1 approximation $\widehat{\boldsymbol{A}}(1)$.

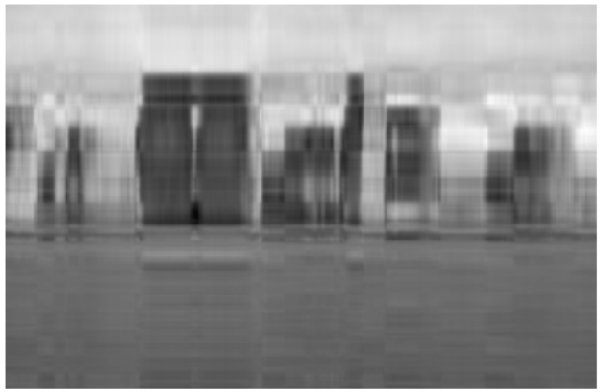


(c) Rank-2 approximation $\widehat{\boldsymbol{A}}(2)$.



(d) Rank-3 approximation $\widehat{\boldsymbol{A}}(3)$.

(e) Rank-4 approximation $\widehat{\boldsymbol{A}}(4)$.



(f) Rank-5 approximation $\widehat{\boldsymbol{A}}(5)$.

Maximum Likelihood Estimation (MLE)

- Goal: estimate parameters θ from observed data $\{x_1, \cdots, x_N\}$
- Main idea: We should choose parameters that assign high probability to the observed data:

$$\hat{\theta} = \text{argmax } L(\theta; x_1, \cdots, x_N)$$

Three steps for computing MLE

1. Write down the likelihood objective:

$$L(\theta; x_1, \cdots, x_N) = \prod_{i=1}^{N} L(\theta; x_i)$$

2. Transform to log likelihood:

$$l(\theta; x_1, \cdots, x_N) = \sum_{i=1}^{N} \log L(\theta; x_i)$$

3. Compute the critical point:

$$\frac{\partial l}{\partial \theta} = 0$$

Example - categorial distribution

$\mathbf{X}$ is a discrete random variable with the following probability mass function ($0 \leq \theta \leq 1$ is an unknown parameter):

$\mathbf{X}$	0	1	2	3
$P(\mathbf{X})$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

- The following 10 independent observations were taken from $\mathbf{X}$: $\{3, 0, 2, 1, 3, 2, 1, 0, 2, 1\}$.
- What is the MLE for θ?

Step 1: Likelihood objective

$$
\begin{aligned}
L(\theta) &= P(X=3)P(X=0)P(X=2)P(X=1)P(X=3) \\
&\quad \times P(X=2)P(X=1)P(X=0)P(X=2)P(X=1) \\
&= (\frac{2\theta}{3})^2(\frac{\theta}{3})^3(\frac{2(1-\theta)}{3})^3(\frac{(1-\theta)}{3})^2
\end{aligned}
$$

Step 2: Log likelihood

$$
\begin{aligned}
l(\theta) &= \log L(\theta) \\
&= 2(\log \frac{2}{3} + \log \theta) + 3(\log \frac{1}{3} + \log \theta) \\
&+ 3(\log \frac{2}{3} + \log(1-\theta)) + 2(\log \frac{2}{3} + \log(1-\theta)) \\
&= C + 5(\log \theta + \log(1-\theta))
\end{aligned}
$$

Step 3: critical points

$$\frac{\partial l}{\partial \theta} = 0$$
$$\rightarrow 5(\frac{1}{\theta} - \frac{1}{1-\theta}) = 0$$
$$\rightarrow \hat{\theta} = 0.5$$

Exercise

Suppose that $X_1, \cdots, X_n$ form a random sample from a uniform distribution on the interval $(0, \theta)$, where of the parameter $\theta > 0$ but is unknown. Find MLE of θ.

Solution

- Calculate the likelihood:

$$L(X_1, \ldots, X_n; \theta) = \prod_i P_\theta(X_i) = \prod_i \frac{\mathbb{I}\left(X_i \in (0, \theta)\right)}{\theta}$$

- Calculate the log-likelihood:

$$l(\theta) = \log \prod_i P_\theta(X_i) = \sum_i \log \frac{\mathbb{I}\left(X_i \in (0, \theta)\right)}{\theta}$$

If $X_i \notin (0, \theta)$, then $\log 0$ will be undefined. Therefore, $\theta \in [\max_i\{X_i\}, \infty)$

- What value of θ maximizes $l(\theta)$? Given that $\theta \in [\max_i\{X_i\}, \infty)$, we have

$$l(\theta) = \sum_i \log \frac{1}{\theta} = -\sum_i \log \theta = -n \log \theta$$

Since log is a monotonic function, increasing θ will increase $\log \theta$ and decrease $l(\theta)$. Therefore, to maximize $l(\theta)$, we choose the smallest feasible value of θ, i.e., $\hat{\theta} = \max_i\{X_i\}$.

Bayesian Inference - Philosophy

- Bayesian interprets probability as degrees of beliefs.
- Bayesian treats parameters as random variables.
- Bayesian learning is updating our beliefs (probability distribution) based on observations.

Bayesian versus Frequentist

- MLE is the standard frequentist inference method.
- Bayesian and frequentist are the two main approaches in statistical machine learning. Some of their ideological differences can be summarized as:

	Frequentist	Bayesian
Probability is	relative frequency	degree of beliefs
Parameter θ is	unknown constant	random variable

Han Liu and Larry Wasserman, Statistical Machine Learning, 2014

The Bayesian approach to machine learning

1. We define a model that expresses qualitative aspects of our knowledge (eg, forms of *distributions*, independence assumptions). The model will have some unknown *parameters*.
2. We specify a *prior* probability distribution for these unknown parameters that expresses our beliefs about which values are more or less likely, before seeing the data.
3. We gather data.
4. We compute the *posterior* probability distribution for the parameters, given the observed data.
5. We use this posterior distribution to draw scientific conclusions and make predictions

---

Radford M. Neal, Bayesian Methods for Machine Learning, NIPS 2004 tutorial

Computing the posterior

- The posterior distribution is computed by the Bayes' rule:

$$P(parameter|data) = \frac{P(parameter)P(data|parameter)}{P(data)}$$

- The denominator is just the required normalizing constant. So as a proportionality, we can write:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Exercise

- Suppose you have a Beta(4, 4) prior distribution on the probability θ that a coin will yield a 'head' when spun in a specified manner.
- The coin is independently spun ten times, and 'heads' appear fewer than 3 times. You are not told how many heads were seen, only that the number is less than 3.
- Calculate your exact posterior density (up to a proportionality constant) for θ and sketch it.

Solution

- Prior:

$$\theta \sim \text{Beta}(4,4) \implies \text{prior}(\theta) = \frac{1}{B(4,4)}\theta^3(1-\theta)^3$$

  where $B(4,4)$ is a normalization constant.

- Number of heads in n trials follows a binomial distribution $Binomial(k; n, \theta)$. For $n = 10$, i.e.,

$$P(\text{Heads} = k; n = 10, \theta) = \binom{10}{k}\theta^k(1-\theta)^{10-k}$$

  Therefore, the Likelihood:

$$\begin{aligned} &L(\text{less than 3 heads out of 10 samples}; \theta) \\ &= \sum_{k=0}^{2} P(\text{Heads} = k; n = 10, \theta) \\ &= (1-\theta)^{10} + 10 \cdot \theta(1-\theta)^9 + 45 \cdot \theta^2(1-\theta)^8 \end{aligned}$$

Solution - Continue

- Posterior:

$$\begin{aligned}
&\text{posterior}(\theta) \propto L(\text{less than 3 heads out of 10 samples}; \theta) \cdot \text{prior}(\theta) \\
&\propto \left[(1-\theta)^{10} + 10 \cdot \theta(1-\theta)^{9} + 45 \cdot \theta^{2}(1-\theta)^{8}\right] \cdot \theta^{3}(1-\theta)^{3} \\
&= \theta^{3}(1-\theta)^{13} + 10 \cdot \theta^{4}(1-\theta)^{12} + 45 \cdot \theta^{5}(1-\theta)^{11}
\end{aligned}$$