COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2024.

Lecture 15

Logistics

- We released Problem Set 3 on Friday. It is due Friday 11/8 at 11:59pm.
- Midterm grade have been posted. The average was good – around an 80%.
- Reach out to me if you are concerned about your midterm grade or your grade in the class overall.

Quiz Questions/Concerns

- Linear algebra review.
- Concerns about grades.
- Participation grade.

Summary

Last Class:

- No-distortion embeddings for data lying in a k-dimensional subspace via an orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$ for that subspace.
- View as low-rank matrix factorization. Introduce concept of low-rank approximation.
- Idea of approximating a data matrix $\mathbf{X}$ with $\mathbf{X}\mathbf{V}\mathbf{V}^T$ when the data points lie close to the subspace spanned by $\mathbf{V}$'s columns.

This Class:

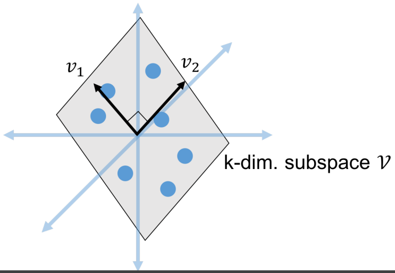
- Finding an optimal orthogonal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$ to minimize $\|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2$ when out data does not exactly lie in a low-dimensional subspace, via eigendecomposition.
- Measuring the error of the optimal low-rank approximation via covariance matrix eigenvalues.

Low-Rank Factorization

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie in a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be written as

$$\mathbf{X} = \mathbf{X}\mathbf{V}\mathbf{V}^T \text{ (implies } \mathbf{rank}(\mathbf{X}) \leq k)$$

- $\mathbf{V}\mathbf{V}^T$ is a projection matrix, which projects the rows of $\mathbf{X}$ (the data points $\vec{x}_1, \ldots, \vec{x}_n$ onto the subspace $\mathcal{V}$.

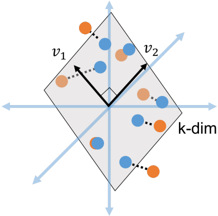


$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Approximation

Claim: If $\vec{x}_1, \ldots, \vec{x}_n$ lie close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T$$



Note: $\mathbf{X}\mathbf{V}\mathbf{V}^T$ has rank k. It is a low-rank approximation of $\mathbf{X}$.

$$\mathbf{X}\mathbf{V}\mathbf{V}^T = \underset{\mathbf{B} \text{ with rows in } \mathcal{V}}{\arg\min} \|\mathbf{X} - \mathbf{B}\|_F^2 = \sum_{i,j} (\mathbf{X}_{i,j} - \mathbf{B}_{i,j})^2.$$

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Low-Rank Approximation

So Far: If $\vec{x}_1, \ldots, \vec{x}_n$ lie close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as:

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}\mathbf{V}^T.$$

This is the closest approximation to $\mathbf{X}$ with rows in $\mathcal{V}$ (i.e., in the column span of $\mathbf{V}$).

- Letting $\mathbf{V}\mathbf{V}^T\vec{x}_i$, $\mathbf{V}\mathbf{V}^T\vec{x}_j$ be the i^{th} and j^{th} projected data points,

$$\|\mathbf{V}\mathbf{V}^T\vec{x}_i - \mathbf{V}\mathbf{V}^T\vec{x}_j\|_2 = \|\mathbf{V}^T\mathbf{V}\mathbf{V}^T\vec{x}_i - \mathbf{V}^T\mathbf{V}\mathbf{V}^T\vec{x}_j\|_2. = \|\mathbf{V}^T\vec{x}_i - \mathbf{V}^T\vec{x}_j\|_2.$$

- I.e., we can use the rows of $\mathbf{X}\mathbf{V} \in \mathbb{R}^{n \times k}$ as a compressed approximate data set.

Key question is how to find the subspace $\mathcal{V}$ and correspondingly $\mathbf{V}$.

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Properties of Projection Matrices

Quick Exercise 1 : Show that $\mathbf{V}\mathbf{V}^T$ is idempotent. I.e., $(\mathbf{V}\mathbf{V}^T)(\mathbf{V}\mathbf{V}^T)\vec{y} = (\mathbf{V}\mathbf{V}^T)\vec{y}$ for any $\vec{y} \in \mathbb{R}^d$.

Quick Exercise 2: Show that $\mathbf{V}\mathbf{V}^T(\mathbf{I} - \mathbf{V}\mathbf{V}^T) = \mathbf{0}$ (the projection is orthogonal to its complement).

Pythagorean Theorem

Pythagorean Theorem: For any orthonormal $\mathbf{V} \in \mathbb{R}^{d \times k}$ and any $\vec{y} \in \mathbb{R}^d$,

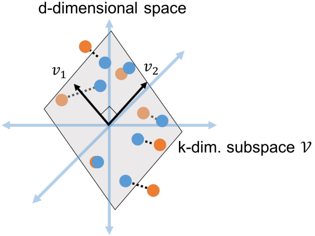
$$\|\vec{y}\|_2^2 = \|(\mathbf{V}\mathbf{V}^T)\vec{y}\|_2^2 + \|\vec{y} - (\mathbf{V}\mathbf{V}^T)\vec{y}\|_2^2.$$

Best Fit Subspace

If $\vec{x}_1, \ldots, \vec{x}_n$ are close to a k-dimensional subspace $\mathcal{V}$ with orthonormal basis $\mathbf{V} \in \mathbb{R}^{d \times k}$, the data matrix can be approximated as $\mathbf{X}\mathbf{V}\mathbf{V}^T$. $\mathbf{X}\mathbf{V}$ gives optimal embedding of $\mathbf{X}$ in $\mathcal{V}$.

How do we find $\mathcal{V}$ (equivilantly $\mathbf{V}$)?

$$\underset{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}}{\arg\min} \|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 = \sum_{i,j} (\mathbf{X}_{i,j} - (\mathbf{X}\mathbf{V}\mathbf{V}^T)_{i,j})^2 = \sum_{i=1}^{n} \|\vec{x}_i - \mathbf{V}\mathbf{V}^T\vec{x}_i\|_2^2 \quad \underset{\text{orthonor}}{\text{ar}}$$



$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Solution via Eigendecomposition

$\mathbf{V}$ minimizing $\|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2$ is given by:

$$\underset{\text{orthonormal } \mathbf{V} \in \mathbb{R}^{d \times k}}{\arg\max} \|\mathbf{X}\mathbf{V}\|_F^2 = \sum_{i=1}^{n} \|\mathbf{V}^T \vec{x}_i\|_2^2 = \sum_{j=1}^{k} \|\mathbf{X}\vec{v}_j\|_2^2$$

Surprisingly, can find the columns of $\mathbf{V}$, $\vec{v}_1, \ldots, \vec{v}_k$ greedily.

$$\vec{v}_1 = \underset{\vec{v} \text{ with } \|v\|_2 = 1}{\arg\max} \|\mathbf{X}\vec{v}\|_2^2$$

$$\vec{v}_2 = \underset{\vec{v} \text{ with } \|v\|_2 = 1,\ \langle \vec{v}, \vec{v}_1 \rangle = 0}{\arg\max} \|\mathbf{X}\vec{v}\|_2^2$$

$$\cdots$$

$$\vec{v}_k = \underset{\vec{v} \text{ with } \|v\|_2 = 1,\ \langle \vec{v}, \vec{v}_j \rangle = 0\ \forall j < k}{\arg\max} \|\mathbf{X}\vec{v}\|_2^2.$$

$\vec{v}_1, \ldots, \vec{v}_k$ are the top k eigenvectors of $\mathbf{X}^T\mathbf{X}$ by the *Courant-Fischer Principle.*

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: orthogonal basis for subspace $\mathcal{V}$. $\mathbf{V} \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.

Review of Eigenvectors and Eigendecomposition

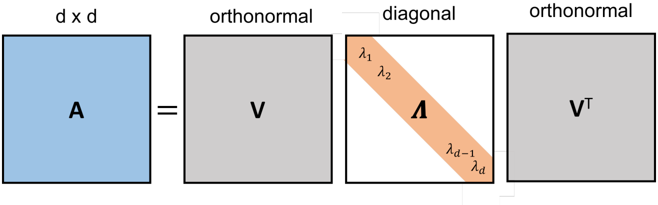
Eigenvector: $\vec{x} \in \mathbb{R}^d$ is an eigenvector of a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ if $\mathbf{A}\vec{x} = \lambda \vec{x}$ for some scalar λ (the eigenvalue corresponding to $\vec{x}$).

- That is, $\mathbf{A}$ just 'stretches' x.
- If $\mathbf{A}$ is symmetric, can find d orthonormal eigenvectors $\vec{v}_1, \ldots, \vec{v}_d$. Let $\mathbf{V} \in \mathbb{R}^{d \times d}$ have these vectors as columns.

$$\mathbf{AV} = \begin{bmatrix} | & | & & | \\ \mathbf{A}\vec{v}_1 & \mathbf{A}\vec{v}_2 & \cdots & \mathbf{A}\vec{v}_d \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \lambda_1\vec{v}_1 & \lambda_2\vec{v}_2 & \cdots & \lambda\vec{v}_d \\ | & | & & | \end{bmatrix} = \mathbf{V}\boldsymbol{\Lambda}$$

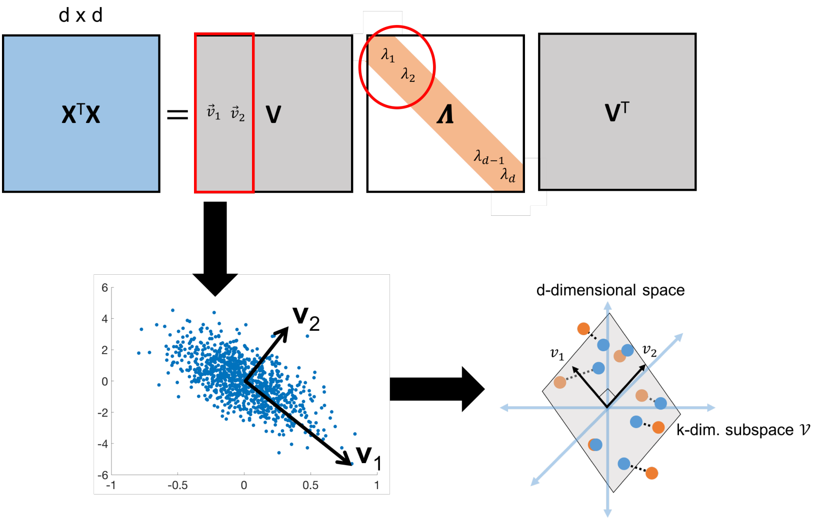
Yields eigendecomposition: $\mathbf{A}\mathbf{V}\mathbf{V}^T = \mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$.

Review of Eigenvectors and Eigendecomposition



Typically order the eigenvectors in decreasing order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$.

Low-Rank Approximation via Eigendecomposition



Low-Rank Approximation via Eigendecomposition

Upshot: Letting $\mathbf{V}_k$ have columns $\vec{v}_1, \ldots, \vec{v}_k$ corresponding to the top k eigenvectors of the covariance matrix $\mathbf{X}^T\mathbf{X}$, $\mathbf{V}_k$ is the orthogonal basis minimizing

$$\|\mathbf{X} - \mathbf{X}\mathbf{V}_k\mathbf{V}_k^T\|_F^2.$$

This is principal component analysis (PCA).

How accurate is this low-rank approximation? Can understand using eigenvalues of $\mathbf{X}^T\mathbf{X}$.

$\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$: data points, $\mathbf{X} \in \mathbb{R}^{n \times d}$: data matrix, $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^d$: top eigenvectors of $\mathbf{X}^T\mathbf{X}$, $\mathbf{V}_k \in \mathbb{R}^{d \times k}$: matrix with columns $\vec{v}_1, \ldots, \vec{v}_k$.