Machine Learning: Yoshinari Fujinuma

University of Colorado Boulder

LECTURE 22

Slides adapted from Chenhao Tan, Jordan Boyd-Graber, Chris Ketelsen

Logistics

- In-class quiz on Monday.
- Friday will be a review session.
- HW4 is due today.

Learning objectives

- Learn principal component analysis

Unsupervised learning

- Clustering
  - K-means
- Dimensionality reduction
  - Principal component analysis

Outline

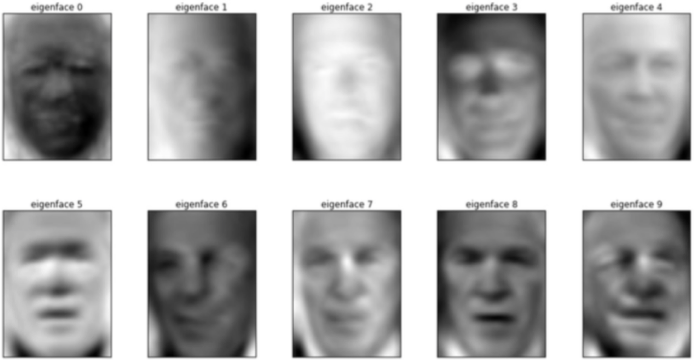
Principal Component Analysis

Example: Eigenfaces / Facial Recognition

- "Labeled Faces in the Wild" dataset
- Roughly 1300 images of 7 different people's faces in various orientation and lighting
- Images are 50x37 grayscale or 1850 features

Example: Eigenfaces / Facial Recognition

Example: Eigenfaces / Facial Recognition



Principal Component Analysis

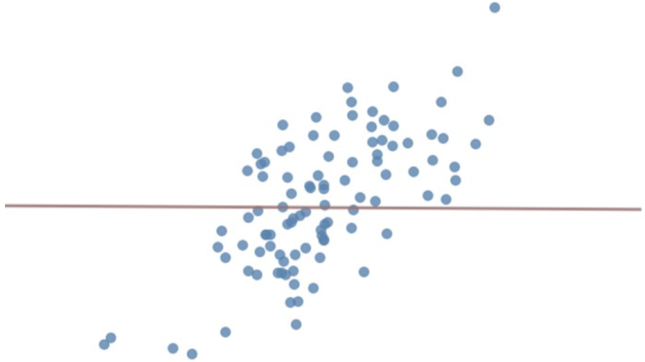
Main idea: The principal components give a new perpendicular coordinate system to view data.

Principal Component Analysis - 1D

How should we reduce this dataset to one dimension?

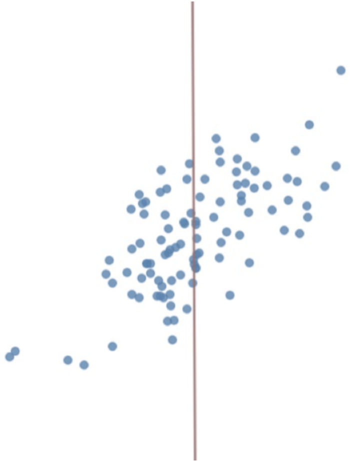
Principal Component Analysis - 1D

How should we reduce this dataset to one dimension?



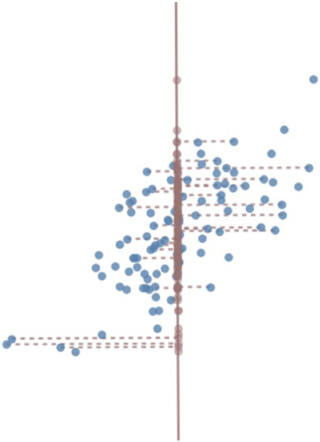
Principal Component Analysis - 1D

How should we reduce this dataset to one dimension?



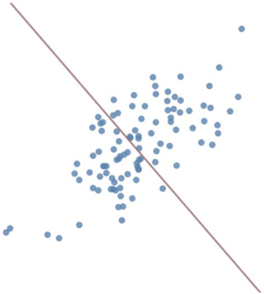
Principal Component Analysis - 1D

How should we reduce this dataset to one dimension?



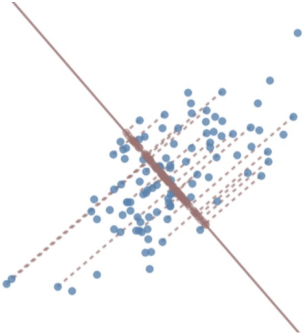
Principal Component Analysis - 1D

How should we reduce this dataset to one dimension?



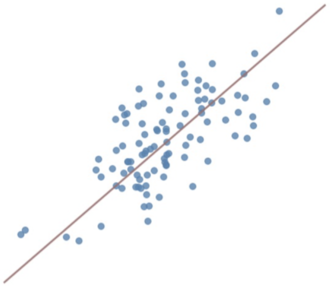
Principal Component Analysis - 1D

How should we reduce this dataset to one dimension?



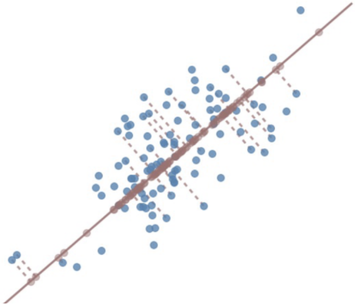
Principal Component Analysis - 1D

How should we reduce this dataset to one dimension?



Principal Component Analysis - 1D

How should we reduce this dataset to one dimension?



Principal Component Analysis

The best vector to project onto is called the **1st principal component.**

Principal Component Analysis

The best vector to project onto is called the **1st principal component.**
What properties should it have?

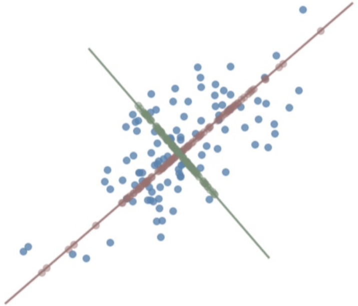
- Capture largest variance in data

After we've found the first, look the second which:

- Captures largest variance
- Should be orthogonal to principal component that came before it

Principal Component Analysis

Principal components of the previous example



Principal Component Analysis: Projection

Let $\mathbf{w_1}$ represent the first principal component (which we know nothing about).
We project ith training example onto the first principal component via dot product.

$$\mathbf{x}_i^T \mathbf{w_1}$$

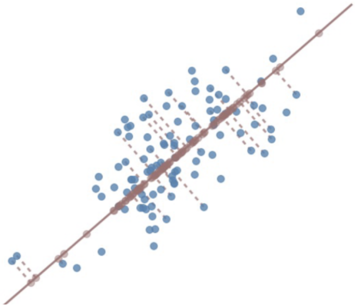
(Note that the above is a scalar)
If we want to project with respect to the original coordinates

$$(\mathbf{x}_i^T \mathbf{w_1})\mathbf{w_1}$$

(Note that the above is a vector)
If we store data in an $m \times d$ matrix X (where $\mathbf{x}_i$ are rows) then, we can project all examples onto the first principal component by

$$X\mathbf{w_1}$$

Principal Component Analysis: Projection



PCA step-by-step

1. **Important:** Center and normalize data before performing PCA.
2. Calculate the covariance matrix X of data points.
3. Calculate eigenvectors and their corresponding eigenvalues.
4. Sort the eigenvectors according to their eigenvalues in decreasing order.
5. Choose first k eigenvectors which satisfies the target explained variance.
6. Transform the original n dimensional data points into k dimensions using the first k eigenvectors.

Principal Component Analysis: Math Background

How do we find the first principle component?
(**Caution**: In practice, just derive eigenvectors, but let's explain why do we use eigenvectors)

Store data in an $m \times d$ matrix X (where $\mathbf{x}_i$ are rows)
Define covariance matrix $C^X = \frac{1}{m-1} X^T X$
Claim: First principle component $\mathbf{w}_1$ is the eigenvector of C^X corresponding to the largest eigenvalue
Recall: $\mathbf{w}$ is an eigenvector of A with associated eigenvalue λ if

$$A\mathbf{w} = \lambda \mathbf{w}$$

Principal Component Analysis: Math Background

How do we find $\mathbf{w_1}$?
As X is already centered, so $\mathrm{mean}(X\mathbf{w}) = 0$. Their variance is

$$\frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{x}_i^T\mathbf{w})^2 = \frac{1}{m-1}(X\mathbf{w})^T(X\mathbf{w}) = \frac{1}{m-1}\mathbf{w}X^TX\mathbf{w}$$

$$\frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{x}_i^T\mathbf{w})^2 = \mathbf{w}^TC^X\mathbf{w} = \sigma_{\mathbf{w}}^2$$

Want to choose $\mathbf{w}$ to have unit length, and make $\sigma_{\mathbf{w}}^2$ as large as possible.

Principal Component Analysis: Math Background

$$
\begin{aligned}
\max_{\mathbf{w}} \quad & \mathbf{w}^T C^X \mathbf{w} \\
\text{s.t.} \quad & \mathbf{w}^T \mathbf{w} = 1
\end{aligned}
$$

Constrained optimization!
Define Lagrangian

$$L(\mathbf{w}, \lambda) = \mathbf{w}^T C^X \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1)$$

$$
\begin{aligned}
\frac{\partial L}{\partial \lambda} &= \mathbf{w}^T \mathbf{w} - 1 &= 0 \\
\nabla_{\mathbf{w}} L &= 2C^X \mathbf{w} - 2\lambda \mathbf{w} &= 0
\end{aligned}
$$

Principal Component Analysis: Math Background

$$\begin{aligned} \frac{\partial L}{\partial \lambda} &= \mathbf{w}^T\mathbf{w} - 1 &= 0 \\ \nabla_{\mathbf{w}} L &= 2C^X\mathbf{w} - 2\lambda\mathbf{w} &= 0 \end{aligned}$$

Solution is $\mathbf{w}$ and λ such that

$$C^X\mathbf{w} = \lambda\mathbf{w} \qquad \text{and} \qquad \mathbf{w}^T\mathbf{w} = 1$$

w is an eigenvector, and max variance is eigenvalue.

$$\sigma^2_{\mathbf{w}} = \mathbf{w}^T C^X \mathbf{w} = \lambda \mathbf{w}^T\mathbf{w} = \lambda$$

So the first principal component is the eigenvector associated with the largest eigenvalue.

Principal Component Analysis

What should k be?
Eigenvalues tell you variance capture

- Make a plot, look for elbows
- Decide based on **explained variance** (EV)

$$EV = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \quad \text{usually choose } k \text{ s.t. } EV > 99\%$$

where λ_i is the ith largest eigenvalue of the covariance matrix C^X

Reposting PCA step-by-step

- Calculate the covariance matrix X of data points.
- Calculate eigenvectors and their corresponding eigenvalues.
- Sort the eigenvectors according to their eigenvalues in decreasing order.
- Choose first k eigenvectors which satisfies the target explained variance.
- Transform the original n dimensional data points into k dimensions using the first k eigenvectors.

(Bonus) Second principal component of Principal Component Analysis

Now what is the second principal component?
Claim: The second principal component is the eigenvector of C^X with second largest eigenvalue.

$$\begin{array}{rl} \max_{\mathbf{w}} & \mathbf{w}^T C^{X'} \mathbf{w} \\ \text{s.t.} & \mathbf{w}^T \mathbf{w} = 1 \end{array},$$

where $X' = X - X\boldsymbol{w}_1\boldsymbol{w}_1^T$, $\boldsymbol{w}_1$ is the eigenvector for λ_1, the largest eigenvalue.