ECE M146 Introduction to Machine Learning

Lecture 16 - Spring 2021

Prof. Lara Dolecek

ECE Department, UCLA

Today's Lecture

Recap:

- Unsupervised Learning
- K-means algorithm for clustering

New topic:

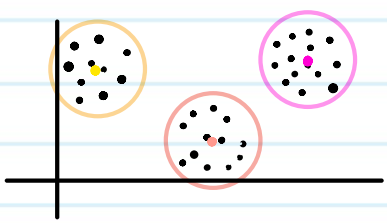
- PCA algorithm for dimensionality reduction
- Linear algebra: SVD, eigen values

Recap: Unsupervised learning

- In the previous lecture, we introduced the notion of unsupervised learning.
- This set-up is described by un-labeled data.
- The goal is to organize or represent this data.

Clustering

- Clustering = group unlabeled data into clusters.
- Use prototypes as cluster representatives.
- This is a form of lossy data compression.



everything in a cluster gets mapped to center point (mean) of cluster

K-means algorithm

- K-means algorithm is an iterative algorithm that iterates between two steps:

1) Cluster Assignment (set indicators given the prototypes)
2) Refitting (set prototypes given the indicators)

suppose we have N data points that we wish to organize into k = 3 clusters

$$r_{nk} = \begin{cases} 1 & \text{if } \operatorname{argmin}_{1 \le j \le k} \|x_n - \mu_j\| = k \\ 0 & \text{else} \end{cases}$$

Today's Lecture

Recap:

- Unsupervised Learning
- K-means algorithm for clustering

New topic:

- PCA algorithm for dimensionality reduction
- Linear algebra: SVD, eigen values

Dimensionality reduction

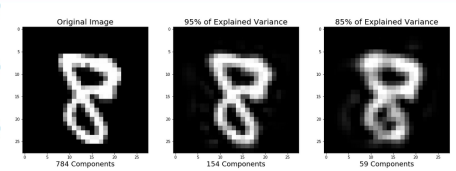
- In clustering, we used prototypes to represent individual clusters (all data points assigned to that cluster).
- Prototypes are of the same dimension as the data points.

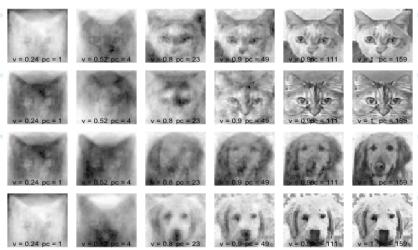
- Another way to represent data efficiently is to **project all data points onto a lower dimensional space.**

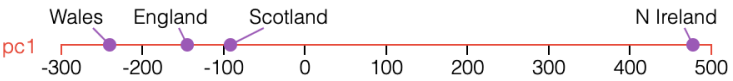
- Applications are also in lossy data compression, visualization and also feature extraction.

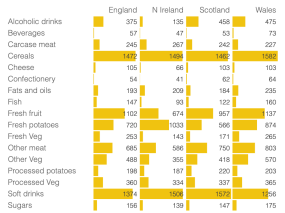
Applications of PCA

- Digit recognition
- Pet identification
- Identified eating habits in the UK









https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

https://setosa.io/ev/principal-component-analysis/

https://bioramble.wordpress.com/2015/09/01/pca-part-5-eigenpets/

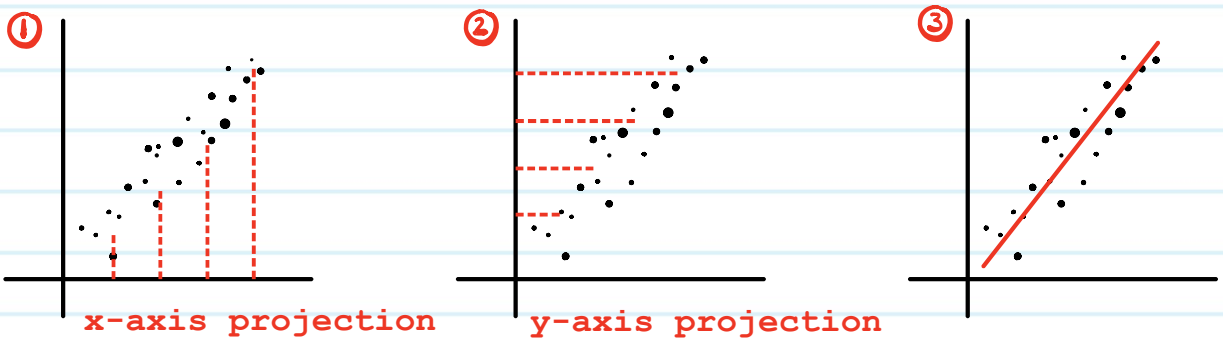
Dimensionality reduction

- Set-up: Suppose data is provided in some D-dimensional space, but it can be well explained in an M-dimensional subspace for M<D.
- Example: D = 2 and M = 1.

- Goal: find this subspace.

How to find the appropriate subspace of dimension M ?

- Some projections:



- We want to look at the projection with the **highest sample variance**, because this will be the most informative choice.

in 1 and 2, the other feature is completely lost

thus, choose 3

First, we need some linear algebra

Notion of a **basis**:

- For a vector space V, a set of vectors $\{v_1, v_2,\ldots,v_D\}$ forms a basis if:
- Vectors $v_1, v_2,\ldots,v_D$ are linearly independent:

$$a_1 v_1 + a_2 v_2 + \ldots + a_D v_D = 0 \quad \Rightarrow \quad a_i = 0 \quad \forall \quad 1 \leq i \leq D$$

- Any vector x in V is uniquely expressed in terms of $\{v_1, v_2,\ldots,v_D\}$ as:

$$x = \sum_{i=1}^{D} b_i v_i$$

- Vector space V must then be of dimension D.

Basis set is not unique

- Here are some examples in 3-D.
- Example 1:
```
(1, 0, 0)
(0, 1, 0)
(0, 0, 1)
```
- Example 2:
```
(1, 0, 0)
(1, 1, 0)
(1, 1, 1)
```
- Example 3:
```
(1, 0, 0)
(0, -1, 0)
(0, 0, 3)
```

Orthonormal basis

- It is also of interest to consider a basis that has vectors that are orthogonal to each other and each has unit norm.

  consider $u_1, u_2, \ldots u_D$

- Mathematically:
- Cond. 1: $u_i^T \cdot u_j = 0 \quad \forall \; i \neq j$
- Cond. 2: $\|u_i\|^2 = 1 \quad \forall \; i$
- This basis is then called **orthonormal basis**.

- Which of the preceding examples constitute orthogonal basis, orthonormal basis, neither ?

condition 1 constitutes orthogonal basis,
condition 1 with condition 2 constitutes orthonormal basis

More on the basis set

- Here are some examples in 3-D.
- Example 1:
  ```
  (1, 0, 0)
  (0, 1, 0)
  (0, 0, 1)
  ```
  orthogonal
  orthonormal
- Example 2:
  ```
  (1, 0, 0)
  (1, 1, 0)
  (1, 1, 1)
  ```
  not orthogonal
  not orthonormal
- Example 3:
  ```
  (1, 0, 0)
  (0, -1, 0)
  (0, 0, 3)
  ```
  orthogonal
  not orthonormal

More on the basis set

even among orthonormal bases, the choice is not unique, e.g.:

$\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\right)$

$\left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0\right)$

$(0, 0, 1)$

Matrix format

- Can organize vectors of an orthonormal basis into a matrix:

suppose we have:

$u_1, u_2, \ldots u_D$

$u_i^T \cdot u_j = 0 \quad \forall \quad i \neq j$

$\|u_i\|^2 = 1 \quad \forall \quad i$

$$U = \left[u_1 \middle| u_2 \middle| \cdots \middle| u_D \right]$$

- Useful property of this matrix:

$$U^T \cdot U = \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_D^T \end{bmatrix} \left[u_1 \middle| u_2 \middle| \cdots \middle| u_D \right]$$

$$= I$$

Back to our set-up: projection with max. sample variance

- Suppose we have N data points, in a D-dimensional space: $x_1, x_2, \ldots, x_N$.
- Suppose we wish to project the data onto the most informative dimension. Let u_1 be a vector in that dimension.
- Then, these are our projections: in that dimension

$$z_1 = u_1^T \cdot x_1$$
$$z_2 = u_1^T \cdot x_2$$
$$\vdots$$
$$z_N = u_1^T \cdot x_N$$

Sample mean and its projection

- Sample mean `before projection`

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- By linearity, projected sample mean is the same as the projection of the sample mean.

$$u_1^T \cdot \bar{x} = u_1^T\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N} \underbrace{u_1^T x_n}_{z_n} = \frac{1}{N}\sum_{n=1}^{N} z_n$$

$$= \bar{z}_n$$

Sample variance in the projection dimension

$$z_n = u_1^T \cdot x_n \qquad \bar{z} = u_1^T \cdot \bar{x}$$

- Consider the following:

$$\frac{1}{N}\sum_{n=1}^{N}(z_n - \bar{z})^2 = \frac{1}{N}\sum_{n=1}^{N}(u_1^T \cdot x_n - u_1^T \cdot \bar{x})^2$$

- Write it in the matrix format:

$$\underset{1\times D}{u_1^T} \cdot \overset{D\times D}{S} \cdot \underset{D\times 1}{u_1} \qquad \text{where} \quad S = \frac{1}{N}\sum_{n=1}^{N}\underset{D\times 1}{(x_n - \bar{x})}\underset{1\times D}{(x_n - \bar{x})^T}$$

Math verification

$$u_1^T \cdot S \cdot u_1 = u_1^T \cdot \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T \cdot u_1$$

$$= \frac{1}{N}\sum_{n=1}^{N}(u_1^T \cdot x_n - u_1^T \cdot \bar{x})(x_n^T \cdot u_1 - \bar{x}^T \cdot u_1)$$

$$= \frac{1}{N}\sum_{n=1}^{N}(u_1^T \cdot x_n - u_1^T \cdot \bar{x})^2$$

Maximization of the sample variance in the projection dimension

- The goal of finding the dimension along which the sample variance is maximized is then equivalent to maximizing $u_1^T \times S \times u_1$?

Maximization of the sample variance in the projection dimension

- The goal of finding the dimension along which the sample variance is maximized is then equivalent to maximizing $u_1^T \times S \times u_1$?

- Almost.
- Note that without further restriction, we could pick an arbitrary dimension and vector u_1 of huge magnitude and artificially maximize this product.
- To make this meaningful, consider only vectors of unit norm:

$$\|u_1\|^2 = 1$$

Mathematical formulation

- What we then have is:

$$\max \; u_1^T \cdot S \cdot u_1 \qquad \text{subject to} \quad \|u_1\|^2 = 1$$

- This is a constrained optimization problem.
- Where did we study constrained optimization problems before ?

SVM

- How did we solve them then ?

Convert to unconstrained optimization and take derivative of the Lagrangian.

Unconstrained optimization

- Mathematical formulation:

$$L = u_1^T \cdot S \cdot u_1 + \lambda_1 (1 - u_1^T \cdot u_1)$$

$$\lambda_1 \neq 0$$

- Set the derivative to zero.

$$\frac{\partial}{\partial u_1} [u_1^T \cdot S \cdot u_1 + \lambda_1 (1 - u_1^T \cdot u_1)] = 0$$

Math, continued

recall:

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x$$

$$\frac{\partial x^T I x}{\partial x} = 2x$$

$$\frac{\partial x^T S x}{\partial x} = (S + S^T)x$$
$$= 2Sx$$

S is symmetric

Eigenvalues and eigenvectors

$$\frac{\partial}{\partial u_1}\left[u_1^T \cdot S \cdot u_1 + \lambda_1 (1 - u_1^T \cdot u_1)\right] = 0$$

$$2Su_1 - \lambda_1 \cdot 2u_1 = 0$$

$$Su_1 = \lambda_1 u_1$$

u_1 is eigenvector of S with eigenvalue λ_1

Upshot

- For any eigen vector, we get a stationary solution.

  (follows from the Lagrangian formulation)

- Matrix S being a covariance matrix is positive semi-definite, so all its eigenvalues are non-negative, and all its eigenvectors are real.

- So we are looking for the eigenvector that corresponds to the largest eigenvalue!

$u_1^T \lambda_1 u_1$

recall: $\max \; u_1^T S u_1$

$= \max \; \lambda_1 \|u_1\|^2$

$= \max \; \lambda_1$

pick largest possible eigenvalue!

Arrived at PCA

- This is the dimension projection.

  `u1 is the unit-1 eigenvector associated with the largest eigenvalue, call it λ1`

- This vector u_1 is called the **first principal component**.

- Hence the name: **principal component analysis.**

General version of PCA

- We can generalize this projection idea to more than one dimension.
- Idea: express S as $S = X^T * X$ (view the matrix X as the centered version of data)

compute mean $\bar{x}$ first, then center the data

$$\tilde{x}_i = x_i - \bar{x}$$

$$X = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_N^T \end{bmatrix}$$

$N \times D$

General version of PCA

- Decompose X as follows: $X = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_N^T \end{bmatrix}$ tall

if $N \gg D$:

$N \times N$ $D \times D$

$X = U \cdot \Sigma \cdot V$

$N \times D$

$U^T U = I_N$

$V^T V = I_D$

$\Sigma = \begin{bmatrix} 0 & & 0 \\ & \ddots & \\ 0 & & 0 \end{bmatrix}$

- This is known as the singular value decomposition (SVD)

Math analysis

- Write S as

$$
\begin{aligned}
S &= X^T X \\
&= (U \Sigma V^T)^T (U \Sigma V^T) \\
&= V \Sigma^T U^T U \Sigma V^T \\
&= V \Sigma^T I_n \Sigma V^T \qquad M = \Sigma^T \Sigma \\
&= V M V^T
\end{aligned}
$$

General PCA:

- Write S as $S = VMV^T$ (M: diagonal)

$$Sv = vM$$

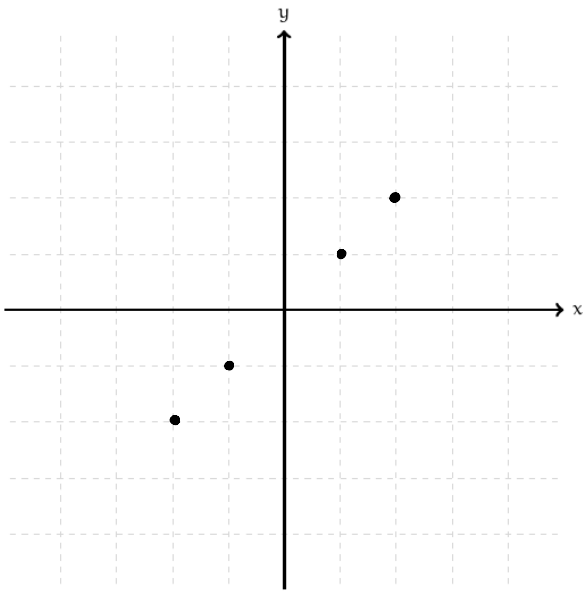
$$S \cdot v_i = v_i m_{ii}$$

$$\vdots$$

- SVD is a powerful matrix decomposition tool. We just showed how to use it in PCA.

also used in GDA to compute inverse of the covariance matrix Σ in $N(\mu, \Sigma)$

Numerical Example



$N = 4$ data pts

note that this data set is already centered at $\bar{x} = [0, 0]$

$$S = \frac{1}{4}\sum_{n=1}^{4} x_n \cdot x_n^T$$

$$= \frac{1}{4}\left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -1 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \end{bmatrix} + \begin{bmatrix} 2 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} -2 & -2 \end{bmatrix} \begin{bmatrix} -2 \\ -2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 2.5 & 2.5 \\ 2.5 & 2.5 \end{bmatrix}$$

Numerical Example

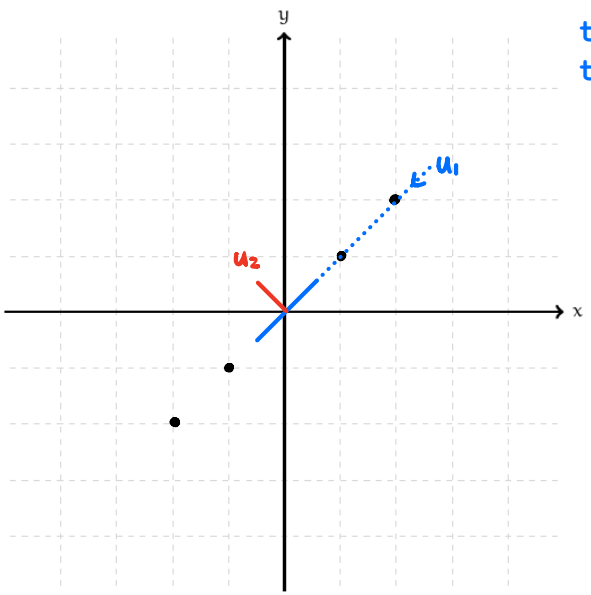
$S \cdot u_1 = \lambda_1 \cdot u_1$

consider unnormalized u normalized

$u_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\lambda_1 = 5$ $\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

$u_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ $\lambda_2 = 0$ $\begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$

Numerical Example



this line is indeed in the direction of u1!

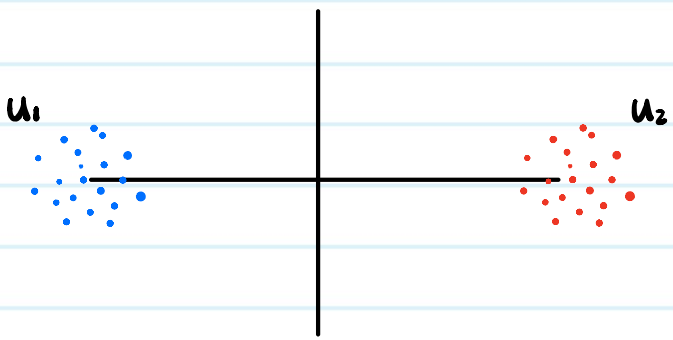
observations:

1. $u_1^T u_2 = 0$
2. $u_2^T x_i = 0$
3. all pts lie on the same dim. as u_1

verify that the variance in the projected dimension 1 is 5

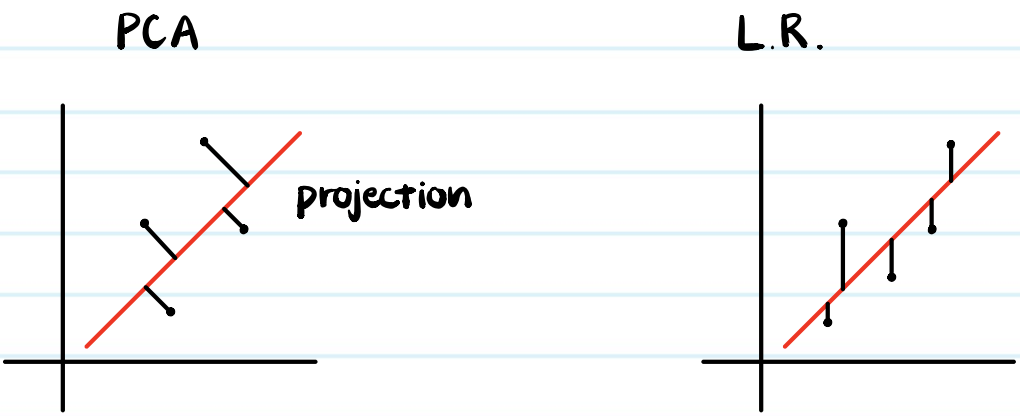
Connections to LDA and to linear regression

recall in LDA we have labeled data:



$$\hat{\Sigma}_i = \frac{1}{N_i} \sum_{n \in \mathcal{N}_i} (x_n - \hat{\mu}_i)(x_n - \hat{\mu}_i)^T$$

Connections to LDA and linear regression



Discussion

- PCA projects onto orthogonal basis functions. It is arguably more practical than just dropping features.
- PCA works well when the data can be well approximated using linear approximation of basis functions but that many not always be the case.
- Sometimes principal components have informative meaning, but sometimes they do not (can lead to incorrect scientific conclusions regarding cause and consequence).