S&DS 365 / 565

Data Mining and Machine Learning

shift from
supervised learning to

Unsupervised Learning

Yale

Loose idea

Given unlabeled data, find structure.

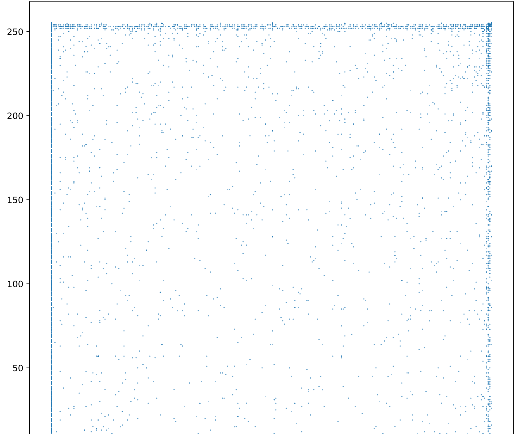
Methods
- clustering
- t-SNE
- Bayesian

Topic Modelling of docs, e-mails

Given unlabeled data, find structure.

2-D
(x_i, y_i)

Suppose I have a dataset that is 12593×784. How can I visualize it? Maybe plot some coordinates against eachother?

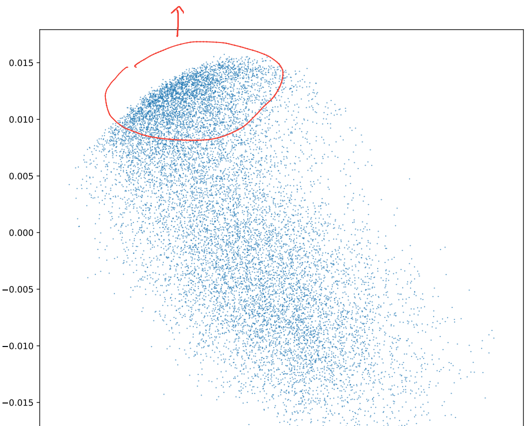


Given unlabeled data, find structure.

baseline

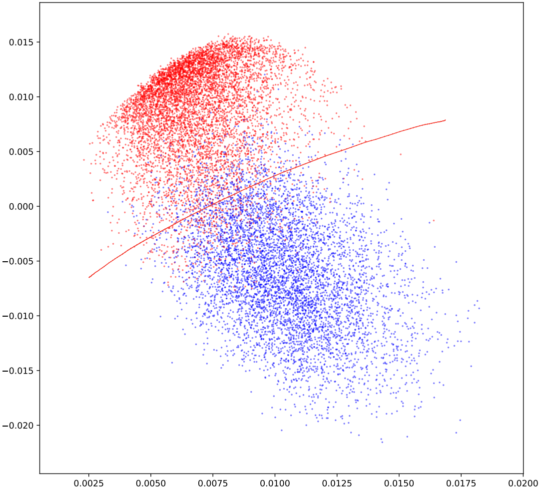
PCA is a workhorse method for computing interesting directions of the data. Here I use PCA to plot “interesting” direction of the data. We will discuss what “interesting” means later.

concentration



Let’s add some color.

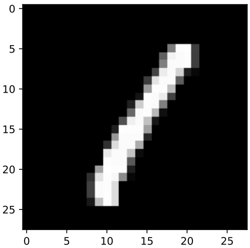
based on the label

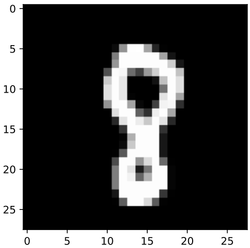


Actually, I generated the plots from data with labels. Images to be exact, the pattern will become very clear to you.

Actually, I generated the plots from data with labels. Images to be exact, the pattern will become very clear to you.

2 labels





Now: clustering

Aim: Find good representatives

Example: Identify distinct $\underbrace{\text{communities}}_{\text{label}}$ of butterflies based on wing size, mass, color

features

Example: Digits images, find representatives (obviously, but since we have labels we can check our clustering)

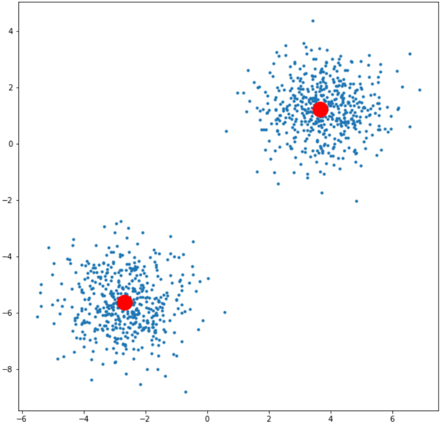
Clustering is a loose concept

Start with **k-means** clustering (explored in previous HW) 1

Dataset: $x_i \in \mathbb{R}^p$ $i \in [n]$

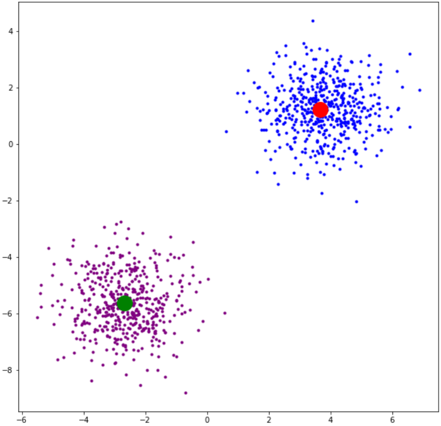
Goal: find vectors (centers) μ_j $j \in [K]$ such that they represent the data well.

k=2 plot in 2-D



We can assign points to each center.

mapping $\pi(i) \in [2]$



We assign purple points to `center green` and blue points to `center red`

This provides a mapping π of points to centers. So $\pi(i) \in [K]$. There are K different centers, so that explains the K part of K-means.

To understand the means part we can now present how the K-means algorithm decides if centers represent the data "well."

This provides a mapping π of points to centers. So $\pi(i) \in [K]$. There are K different centers, so that explains the K part of K-means.

To understand the means part we can now present how the K-means algorithm decides if centers represent the data "well."

representative

Goal: Find centers μ_j such that

$$\mu_j = \arg\min_{\mu_{\pi(i)}} \qquad \sum_{i=1}^{n} \|x_i - \mu_{\pi(i)}\|_2^2$$

is minimized. Here, $\pi(i)$ is actually a function that maps example i to centers j. So we have to find both μ and π.

In general the computation is intractable. However, if we are given π finding μ_j is easy. Homework assignment to check to see that

take mean

$$\widehat{\mu}_j = \frac{1}{n_j} \sum_{i|\pi(i)=j} x_i$$

$j \in [k]$

where $n_j = \sum_{i|\pi(i)=j} 1$. This explains the **means** part of K-means!

number of constituents in group j

map example i to the closest representative μ_j

Similarly, given μ_j finding π is also very easy.

$$\pi(i) = \arg\min_j \|x_i - \mu_j\|_2^2$$

These observations yield a natural **alternating minimization** algorithm.

k-means clustering algorithm

random

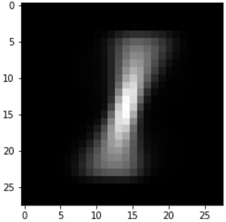
Starting with initial guess for μ_j

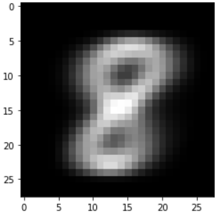
- Compute π
- Then compute μ_j
- repeat until convergence

Let's try it out on the two digits with $K = 2$.

fuzzy images coz people write digits "1" "8" differently
The images capture all possible 笔迹 then average them

output 2 clusters :





Let's try it out on the two digits with $K = 3$.