Introduction to Machine Learning

An overview for physicists

Lecture 3
In this lecture

- Principal component analysis
- *(Independent component analysis)*
- Clustering
- K-means

why?
First principal component

Given N M-dimensional data, what is the linear combination maximizing variability?

Given \( \tilde{x}_i = x_i^j \), find \( w_j \), \( |\tilde{w}| = 1 \) to get \( t_i = \sum_j w_j x_i^j \) such that \( \langle t^2 \rangle \) is maximum (we can assume \( \langle \tilde{x} \rangle = 0 \)).

We find a Rayleigh quotient \( \frac{\tilde{w} C \tilde{w}}{\tilde{w} \tilde{w}} \) maximized by the largest eigenvector.
Principal component analysis (PCA)

Given N M-dimensional data, find P linearly independent linear combinations that retain maximum variance. → Diagonalize the covariance matrix.

First form of dimensionality reduction, or feature extraction, or a simple data compression algorithm.

Problems: everything is linear, not scale invariant, what do we really learn?
Independent component analysis

Can we find really independent components (linearly combined)?

Sketch: We assume the sources are non-Gaussian.

Maximize the entropy of the sources, or log-likelihood that they produce the signal:

$$\log |\det W| + \sum_i \log p_i(W_{ij}x_j)$$

where we need a choice for the p’s (nonlinear).
Clustering

- Divide data in groups with certain similarity
- Mostly based on some form of distance
- Division based on centroid, density, connectivity, ...
- Hard or soft clustering, outliers
- What is correct? Is it univocal?
K-means

We want to divide N points in K clusters.

We have a distance $d(x,y)$ (Euclidean or more complex).

Define K initial means $m_i$, then

- Assign each point to the closest mean
- Update each mean to be the mean of the points in that cluster

Iterate until convergence.
K-means: problems

- Clustering depends on starting condition (and distance)
- Fails for weirdly shaped clusters
- All points contribute the same

→ Lots of ad-hoc solutions (soft partition, medoids, Gaussians, ...)

That’s all for now...