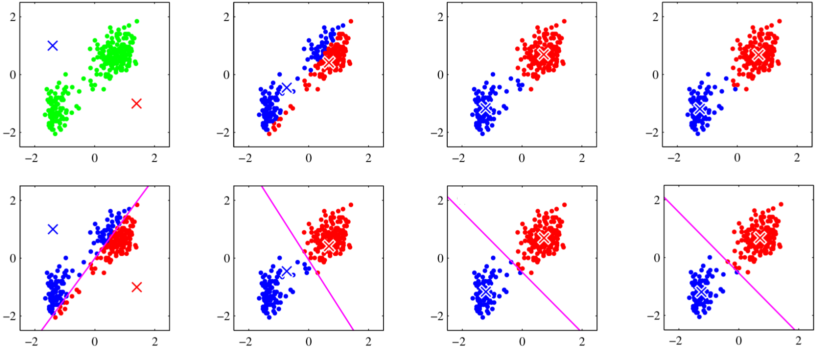
Machine Learning

Unsupervised learning

Maxime Gasse



Introduction

Given a set of observations $\mathbf{v}$, extract useful knowledge.
Highly subjective !

Some supervised learning problems:

- Structure learning: understand relationships between variables
- Outlier Detection: detect novelties / unexpected values
- Clustering: identify groups of similar observations
- Manifold learning: identify an interesting representation space
- Sampling: generate new observations
- ...

In this course we will cover in detail two approaches:

- k-means clustering
- Gaussian mixture models

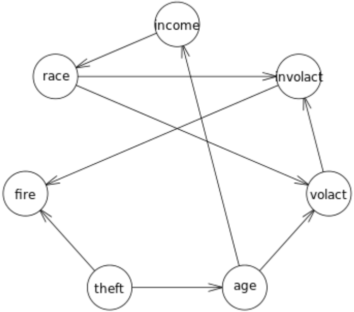
Common supervised learning problems

Structure learning

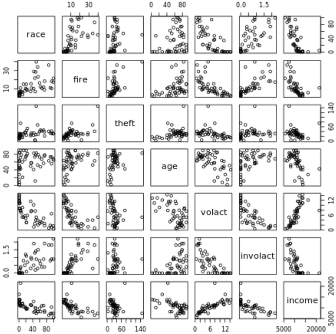
Statistical independence:

$$V_1 \perp\!\!\!\perp V_2 \mid V_3 \iff p(v_1, v_2|v_3) = p(v_1|v_3)p(v_2|v_3), \quad \forall v_1, v_2, v_3$$

Example (Chicago homeowner insurance data)

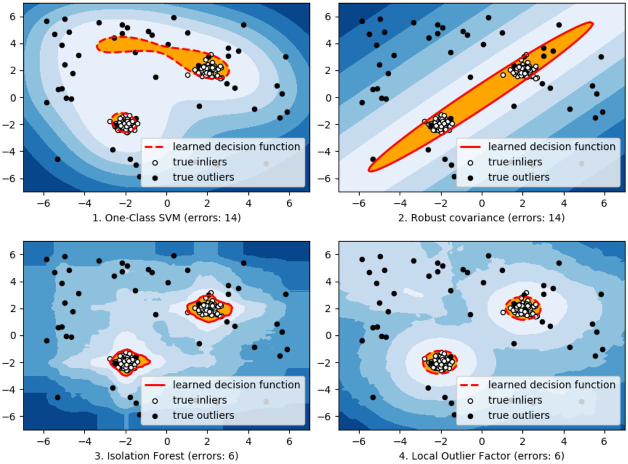


Bayesian network structure



Scatterplot

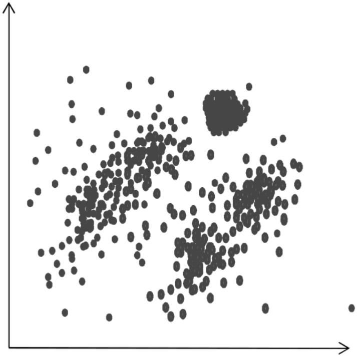
Outlier detection



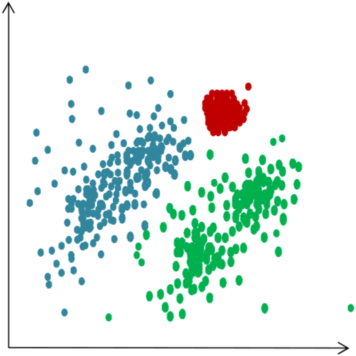
Ill-posed problem. Performance measure ?

Probabilistic interpretation: rare events $\{\mathbf{v} \mid p(\mathbf{v}) < \text{ threshold}\}$.

Clustering



observations

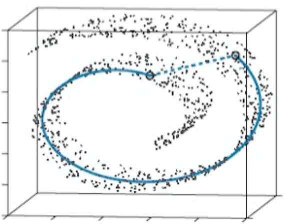


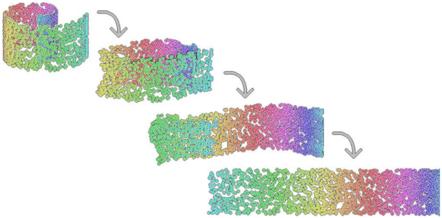
clusters

Ill-posed problem. Performance measure ?
Probabilistic interpretation: hidden variables $p(\mathbf{v}) = \sum_h p(\mathbf{v}|h)p(h)$.

Manifold learning

Find the underlying manifold, where distance is meaningful.



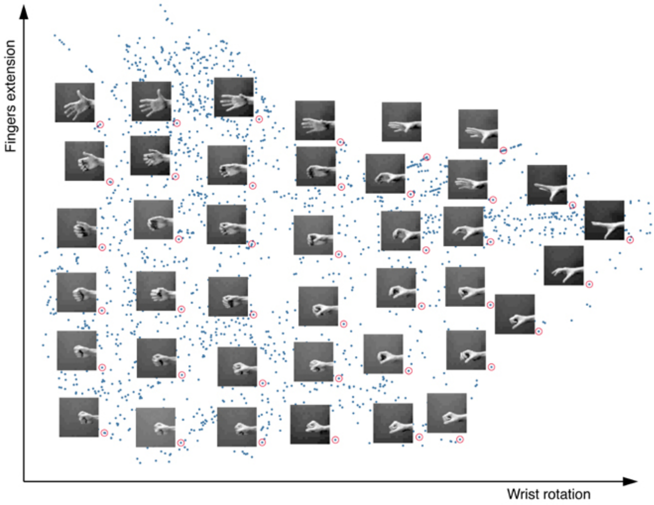


Ill-posed problem. Meaningful distance ?

Probabilistic interpretation: change of variable $\mathbf{z} = \phi(\mathbf{v})$ such that ϕ is reversible and $p(\mathbf{z})$ is simple (uniform, normal. . .).

M. Gashler, D. Ventura, and T. R. Martinez (2011). Manifold Learning by Graduated Optimization.

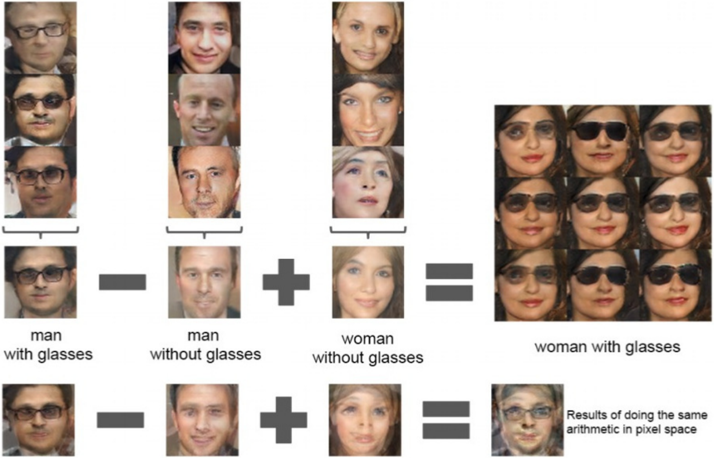
Manifold learning



J. Tenenbaum, V. Silva, and J. Langford (2000). A global geometric framework for nonlinear dimensionality reduction.

Manifold learning

Arithmetic in $\mathcal{Z}$ space.



A. Radford, L. Metz, and S. Chintala (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.

Sampling

Generate new plausible observations.

https://youtu.be/XOxxPcy5Gr4

Probabilistic interpretation: sample $\mathbf{v} \sim p(\mathbf{v})$.

(here: $\mathbf{z} \sim p(\mathbf{z})$ then $\mathbf{v} = \phi^{-1}(\mathbf{z})$).

T. Karras et al. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation.

K-Means Clustering

Main idea

Partition $\mathcal{D}$ into K clusters of similar points, i.e. $\mathbf{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots \mathcal{S}_K\}$.

Minimize the expected within-cluster variance (i.e. barycenter distance).

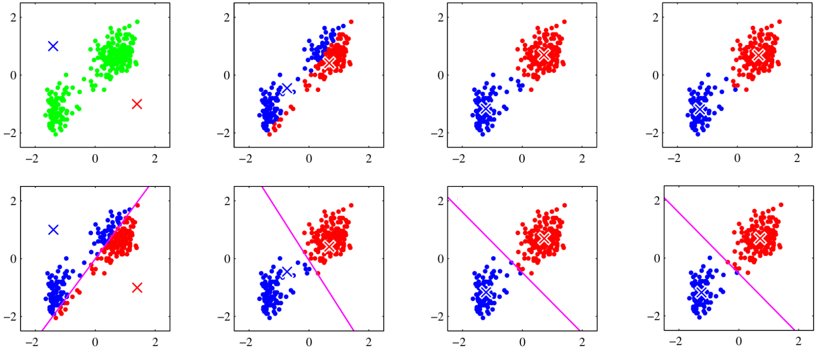
$$\mathbf{S}^\star = \arg\min_{\mathbf{S}} \sum_{i=1}^{K} \sum_{\mathbf{v} \in \mathcal{S}_i} \|\mathbf{v} - \boldsymbol{\mu}_i\|_2^2 \,.$$

NP-hard problem.

Lloyd's algorithm

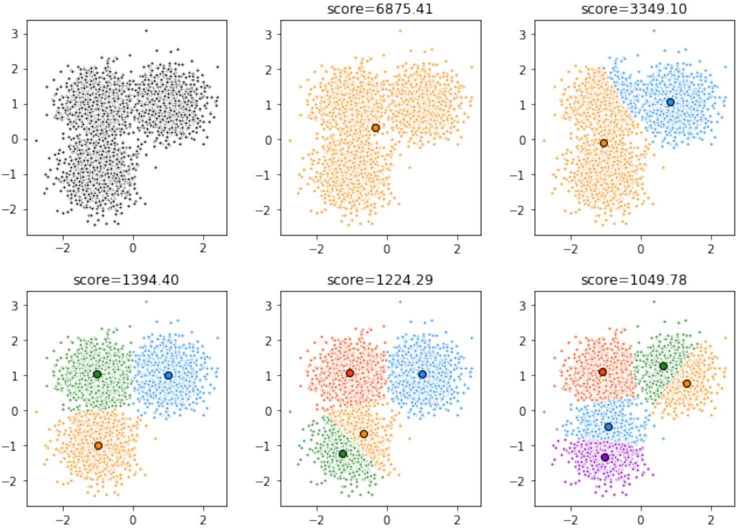
Initialize random barycenters $\{\boldsymbol{\mu}_i\}_{i=1}^{K}$, then repeat until convergence:

1. update boundaries: $\mathcal{S}_i = \{\mathbf{v} \mid \|\mathbf{v} - \boldsymbol{\mu}_i\|_2^2 < \|\mathbf{v} - \boldsymbol{\mu}_j\|_2^2, \forall j \neq i\}$
2. update barycenters: $\boldsymbol{\mu}_i = \frac{1}{|\mathcal{S}_i|} \sum_{\mathbf{v} \in \mathcal{S}_i} \mathbf{v}$



Converges to a local minimum $\implies$ several restarts in practice.

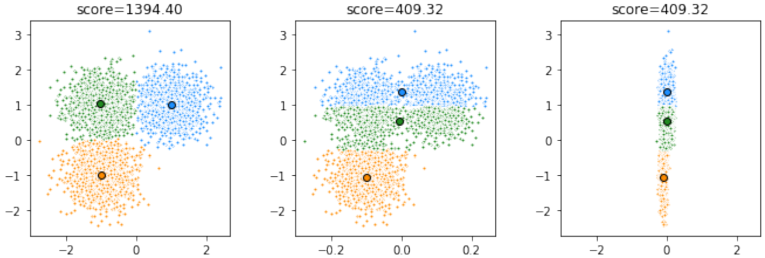
Choice of K ?



$$K \to N \iff \text{score} \to 0$$

Limitations

- piecewise-linear cluster boundaries
- distance metric in high dimensions ?
- $\|\cdot\|_2^2$ in $\mathcal{V}$?



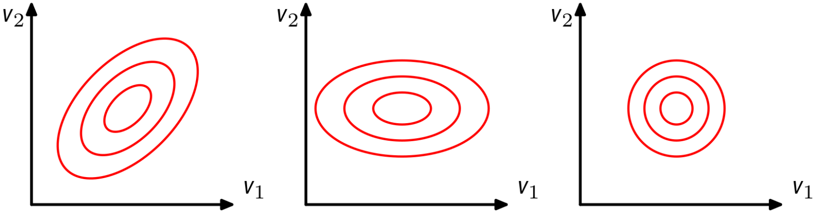
Variant: kernel k-means.

Gaussian mixture model

Single Gaussian

Multivariate normal distribution: $p(\mathbf{v}) = \mathcal{N}(\mathbf{v}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Example ($\mathbf{v} \in \mathbb{R}^2$)



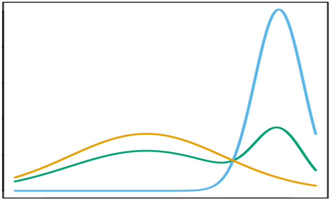
$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix}.$$

Maximum-likelihood parameters:

- $\mu_i = \frac{1}{N} \sum_{\mathbf{v} \in \mathcal{D}} v_i$ (mean);
- $\Sigma_{i,j} = \frac{1}{N-1} \sum_{\mathbf{v} \in \mathcal{D}} (v_i - \mu_i)(v_j - \mu_j)$ (covariance matrix).

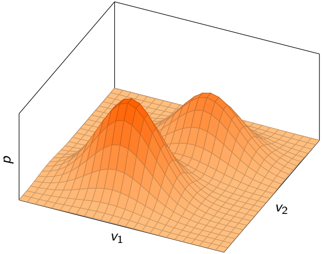
Gaussian mixture

Weighted sum of K normals: $p(\mathbf{v}) = \sum_{k=1}^{K} \pi_k \times \mathcal{N}(\mathbf{v}|\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$.



Let $z \in \{1, \ldots, K\}$ a hidden variable: $p(\mathbf{v}) = \sum_z p(\mathbf{v}, z)$.

$p(\mathbf{v}, z) = p(z)p(\mathbf{v}|z)$
$p(z_k) = \pi_k$
$p(\mathbf{v}|z_k) = \mathcal{N}(\mathbf{v}|\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$



Parameters:

- $\boldsymbol{\pi} \in [0,1]^K$ ($\sum_k \pi_k = 1$)
- $(\boldsymbol{\mu}^{(1)}, \ldots, \boldsymbol{\mu}^{(K)}) \in \mathbb{R}^{D \times K}$
- $(\boldsymbol{\Sigma}^{(1)}, \ldots, \boldsymbol{\Sigma}^{(K)}) \in \mathbb{R}^{D \times D \times K}$

Density estimation

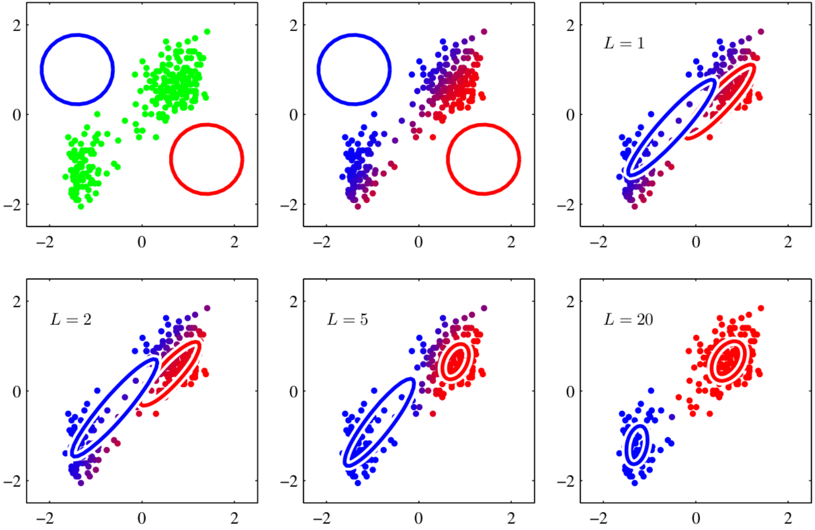
Non-convex problem, no closed-form solution.
$\Longrightarrow$ approximate solution via expectation-maximization (EM).

The EM algorithm: start from random parameters and repeat until convergence:

1. for each data point i and component k, compute $\gamma_k^{(i)} = p(z_k|\mathbf{v}^{(i)})$;
2. for each component k, update its parameters:
   - $\pi_k = \frac{1}{N}\sum_{i=1}^N \gamma_k^{(i)}$
   - $\boldsymbol{\mu}^{(k)} = \frac{1}{N\pi_k}\sum_{i=1}^N \gamma_k^{(i)}\mathbf{v}^{(i)}$
   - $\boldsymbol{\Sigma}^{(k)} = \frac{1}{N\pi_k}\sum_{i=1}^N \gamma_k^{(i)}(\mathbf{v}^{(i)} - \boldsymbol{\mu}^{(k)})(\mathbf{v}^{(i)} - \boldsymbol{\mu}^{(k)})^\intercal$

Converges to a local minimum.

Expectation-maximization illustrated



Gaussian mixture model vs K-means

K-means = GMM + EM under two assumptions:

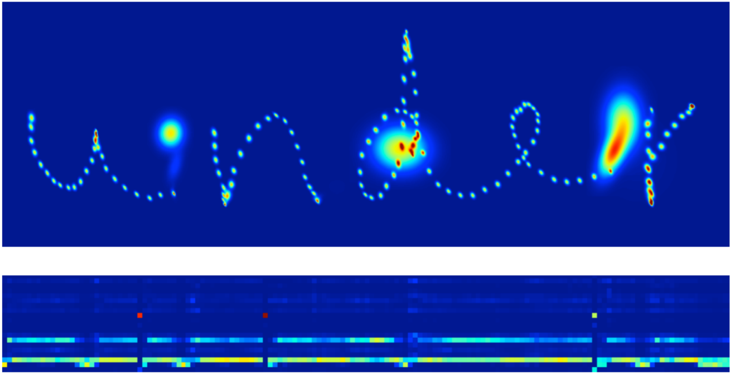
- hard clustering: $\pi_k \in \{0, 1\}$
- isotropic Gaussians: $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$

GMM pros:

- clustering: $\arg\max_z p(z|\mathbf{v})$
- sampling: $z \sim p(z)$, then $\mathbf{v} \sim p(\mathbf{v}|z)$
- outlier/novelty detection: $p(\mathbf{v}) <$ threshold

Conditional GMM in supervised learning

https://www.cs.toronto.edu/~graves/handwriting.html



A. Graves (2013). Generating Sequences With Recurrent Neural Networks.

Course summary

Review of supervised learning tasks

- structure learning
- outlier detection
- clustering
- manifold learning
- sampling

Two methods in detail

- clustering with k-means
- density estimation with Gaussian mixture models