Fundamentals of Probabilistic Data Mining
Chapter I - Introduction

Jean-Baptiste Durand

Ensimag
What is probabilistic data mining?

- Probabilistic does not just mean there is probability somewhere.
- Probabilistic means we want / need to model our data / descriptors \( x \) using probabilities.
- For example in classification, neural networks (and other classifiers) provide \( P(c|x) \) for every possible class \( c \).
- Probabilistic generative models:

\[
 p(x, c) = p(x|c)p(c)
\]

and then

\[
 p(c|x) = \frac{p(x|c)p(c)}{\sum_k p(x|k)p(k)}
\]

(Bayes rule).

- \( p(x|c) \) denotes either \( P(X = x|C = c) \), or if \( X \) is continuous, the probability density function (pdf) of the conditional distribution \( P(X|C = c) \).
Why probabilistic data mining?

- Infer hidden variables / exploit partly missing data
- Example: clustering, image segmentation
- Incorporate particular requirements in clustering
- Infer underlying probabilistic dependencies
- Model complex data (on grids, graphs, ...)
- Simulate phenomena (speech synthesis), make predictions (regime switching in time series)

Segmentation of time series with respect to the variance
Clustering

- Data: points \((x_j)_{j=1,...,n}\) in \(\mathbb{R}^d\).
- Aim: find (maybe predict?) clusters.
- Model-based approach: let \(z_j\) be the cluster of \(x_j\). If \(z_i = z_j = k\) then \(x_i\) and \(z_j\) should have the same (conditional) distribution \(p_k\).

\(Z\) is an unknown / latent variable, useful for clustering.

(e.g. \(z = 0 \rightarrow \text{blue}\))
Why model-based clustering?

- We may want clusters with same means but different variances or covariances.
- Thus we may choose \( p(x|z = k) = \mathcal{N}(0, \sigma_k I) \), or \( p(x|z = k) = \mathcal{N}(0, \sigma_k \Sigma) \), etc.
Probabilistic PCA

- Principal Component Analysis (PCA): represent data in $\mathbb{R}^d$ in lower-dimensional spaces, exploiting redundancy (correlation) between variables.
- Visualization of high-dimensional points in 2D or 3D.
- Visualization of correlations between variables.
- What if for most or even every point $x_j$, some coordinates $x_{i_1,j}, \ldots, x_{i_1,i_j}$ are missing?
- Probabilistic (i.e., model-based) PCA relies on a generative model to exploit partially observed / unknown data.

$$X_i = W\gamma_i + \mu + \varepsilon_i$$ with $\gamma_i \sim \mathcal{N}(0, I_M)$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_D)$ $\perp \perp \gamma_i$ and $W \in \mathbb{R}^{D \times M}$ $M \ll D$
Probabilistic graphical models

- Infer conditional dependence/independence relations between variables.
- Relations summarized by a graph.
- Put weights on edges if possible? (quantitative variables).
- Example: Number of bees, wasps, mint plant, mass of pesticides per km².
- How to deduce Qpesticide from Nbees = 150,000 and Nwasps = 3,000? Qpesticides
Hidden Markov models

- Special case of clustering with temporal dependencies
- Also a graphical model
- Piecewise statistically invariant features with Markovian jumps
Exploit unlabelled data in classification

See full example in Bishop and Lasserre (2007).

- Labelling data (images) is often costly.
- Just collecting data is cheaper.
- How to use a probabilistic model to use both labelled and unlabelled data (semi-supervised learning)?

![Semi-supervised learning: data set](image)
Use just labelled data?

Boundary learned on labelled data  Induced boundary on all data

- Assuming (wrongly...) \( p(x_i | c_i, \theta) \sim \mathcal{N}(\mu_{c_i}, \Sigma_{c_i}) \)
- Estimation of \( p(c_i | x_i) \) is robust to model specification.
- Small amount of labelled data insufficient to learn boundary.
- Labelled data have to be used somehow.
Assuming that points in the same cluster $c$ have distribution $\mathcal{N}(\mu_c, \Sigma_c)$.

Due to model misspecification, inferred boundary is very far from truth.
Combine approaches

Assumptions:
- \( p(x, c, \theta, \tilde{\theta}) = p(\theta, \tilde{\theta})p(c|x, \theta)p(x|\tilde{\theta}) \)
- \( \theta \) parameter for labelled data, \( \tilde{\theta} \) for unlabelled data
- Inference on \( \theta \) benefits from robustness in learning \( p(c|x, \theta) \)
- Inference on \( \tilde{\theta} \) benefits from a larger data set
- Information on \( \tilde{\theta} \) and \( \theta \) is combined in \( p(\theta, \tilde{\theta}) \) (Bayesian approach)

Boundary learned on unlabelled data
PDM in gesture recognition

The Graffiti alphabet

Sequence of images capturing a gesture (from Martin and Durand, 2000)

This problem belongs to the class of ... problems.
What kind of approach to use?
Modelling Graffiti: 1

- Sequential data: hidden Markov models (HMMs)
- Find what is invariant (up to some noise)
- Image differences
- High-dimensional data (more than 1,000, more than sequence length)
- Information: essentially 2D (PCA)
- We cannot be sure the same projection is relevant for every (part of) a letter

Sequence of images capturing a gesture
Modelling Graffiti: 2

- Phase in HMM: phase in a gesture (↗, ↘, ...)
- Generative models → sequence of Gaussian PCAs where the projection planes, means and variances depend on letter and phase.
- Train one HMM per (class of) letter.
- Given some new sequence, computes its probability under each trained HMM (letter) and choose that with highest probability.

```
Mixture of probabilistic PCA  Induced states and clustering
```

![Diagram 1](image1.png)

![Diagram 2](image2.png)
Pre-requisites

- Probability theory: pdf, cdf, expectation, random vectors, correlation, conditional distributions, ...
- Statistics: maximum likelihood estimator and its usual properties (asymptotic normality, ...)
- Constrained optimisation, Lagrange multipliers
- Bonus: geometric approach of PCA, multiple linear regression
**Rules**

- **Evaluation:** labs coef. 0.5, final exam coef. 0.5
- **Labs** (teamwork, 3–4 members). Grading system:
  - < 10 (unfinished);
  - > 18 (optional research-like questions);
  - [10; 18] (other cases)
  - → with support from Xavier Alameda-Pineda.
- **Final exam:** authorized documents are 6 pages from https://v1.overleaf.com/20329178tpnwnrgpnhyy (collaborative writing), nothing else.
Resources

- The slides of the courses, future data sets and scripts are on http://chamilo.grenoble-inp.fr/courses/ENSIMAGWMM9AM21/
- Also contains reminders on the pre-requisites, and some booklet (same contents as slides).
Modelling project and seminar (MSIAM)

- Tracking multiple speakers with auditory, visual or audio-visual data.
- Proposed by Xavier Alameda-Pineda.
- Involves hidden Markov models with estimation by variational approximation.

Tracking Results
Silent:red
Speaking:green

Audio-Contribution
Vision contribution
Dynamic contribution

tracking from audio-visual data
Next course...

... will be in IM2AG F316!
References

Bishop, C. and Lassere, J. Generative or Discriminative? Getting the Best of Both Worlds.  
*Bayesian Statistics* 8, pp. 3-24 (2007)

J. Martin et J.-B. Durand.  
Grenoble.