Learning the topology of a labeled data set

Gaillard Pierre*, Aupetit Michaël*, Gérard Govaert**

*CEA - DASE ; **Compiegne University of Technology

1 Introduction

In statistical supervised learning, it is assumed that natural data are generated by structured probabilistic systems with eventually much fewer degrees of freedom than the ambient space, such that the observed data are supposed to come from a set of labeled principal manifold. Then, basic supervised problems involve a given set of M observed labeled training data $\mathcal{X} = \{x_i, c_i | i = 1, ..., M\}$, where $x_i \in \mathbb{R}^D$ is a feature vector and $c_i \in \{1, ..., K\}$ is its associated class label. The ultimate goal of classification problems is to design a classifier which predicts the class of new feature vectors with a minimum error rate. However, prediction is only the last step of the learning process, which can be enriched by *the study of the topological properties* (*e.g.* intrinsic dimension, connectedness) of the labeled principal manifolds. For example, in the context of exploratory analysis of labeled data, the connectedness of the labeled manifolds allows evaluating the complexity of the classification problem [2].

2 A generative model for learning classes' topology

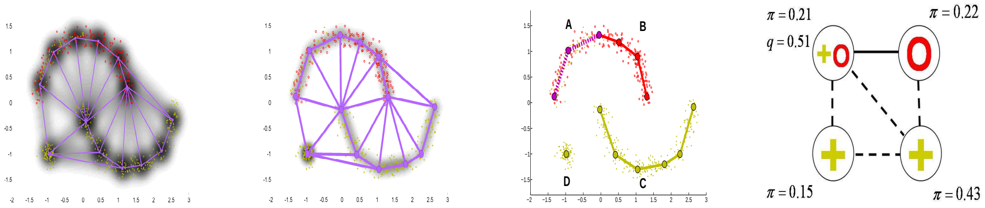
In the context of unsupervised learning, Aupetit [1] proposed the *Generative Gaussian Graph* (*GGG*) to learn the connectedness of the principal manifolds. In this work, we extend the GGG to supervised learning. In the same way as the GGG, we make the assumptions that the principal manifolds are close to a subgraph of the Delaunay graph (DG) of some prototypes $\underline{w}$ and that the data have been corrupted with an additive spherical Gaussian noise with mean 0 and variance σ^2. The topology of the graph is assumed to stand for the one of the principal manifolds. Moreover in order to model a possible superposition (full overlapping) of manifolds of different classes, we use the path proposed by Miller and Uyar [3] that extend "classical" Gaussian mixtures (*GM*) to supervised learning. We assume that the j^{th} component of the model can generate data from K different classes c with respective probabilities β_{cj}. The supervised model is therefore defined by the weighted sum of the *labeled generative vertices* and *generative edges* of the DG which spans $\underline{w}$:

$$p(x, c; \underline{\pi}, \underline{\beta}, \underline{w}, \sigma, DG) = \sum_{j \in DG} \pi_j \beta_{cj} g_j(x; \sigma) \tag{1}$$

such that $\beta_{cj} \geq 0$, $\sum_{c=1}^{K} \beta_{cj} = 1 \; \forall j, \forall d$ and $\pi_j \geq 0$ and $\sum_{j=1} \pi_j = 1$ and where :

$$g_j(x; \sigma) = \left\{ \begin{array}{ll} (2\pi\sigma^2)^{(-D/2)} exp(-\frac{(x-w_j)^2}{2\sigma^2}) & \text{if } j \text{ is the generative vertex } w_j \\ \frac{1}{||w_j - w_{j'}||} \int_{w_j}^{w_{j'}} g^0(x|w; \sigma) dw & \text{if } j \text{ is the generative edge } [w_j; w'_j] \end{array} \right. \tag{2}$$

The learning objective was chosen to be the joint likelihood over the observed labeled data. In order to maximize the likelihood of this model, we use the *EM* algorithm. To get the topology representing graph *wrt* the classes from the generative model, the core idea is to prune from the initial DG, the edges for which there is no chance they generated the



(a) (b) (c) (d)

Fig. 1: **Principle of the Supervised Generative Gaussian Graph**

labeled data. The complete algorithm is the following: **(1)** Initialize the location of the N prototypes $\underline{w}$ using an isotropic *GM*. Since the complexity of our model is closely related to the number N of prototypes, we choose N according to the best *GM* in the sense of the Bayesian Information Criterion to build the initial graph ; **(2)** Construct the DG of the prototypes ($O(DN^3)$); **(3)** Set the weights $\underline{\pi}$ giving equiprobability to each edge and vertex, the conditional class probabilies $\underline{\beta}$ to $1/K$ and the variance σ^2 to the one found by the *GM* at step (1); **(4)** Maximize the likelihood of the supervised generative graph with the EM algorithm ; **(5)** Prune the edges having an edge prior $\pi_j \in \underline{\pi}^*$ less than a threshold ϵ. The remaining edges define the supervised topology representing graph and model the connectedness of the density of all the classes. **(6)** Based on the previous work [2], we build a class graph which summarize the connectedness of the classes.

3 A toy example (figure 1)

We draw 600 2-D data samples from a set of labeled manifolds (figure 1.c) : two quarter-circles (A,B), one half-circle (C) and one point (D). (A) is mixed '+' and 'o', (B) the right one is 'o' , (C) and (D) are '+'. We locate 13 prototypes over the data distribution using a *GM* and we build their DG (a). We use the EM-algorithm to tune the model parameters in order to increase its likelihood (b). We prune the irrevelant edges associated to a null prior (c). The topology of the classes emerges from the resulting graph from which we can learn about the *intra and inter* class connectedness summarized in the class-graph (d): the class 'o' is in one connected component while the class '+' is in 3 components. (A) and (B) are density-connected (connected in the final graph) and it is represented with the bold-link in the class-graph. The others components are density-disconnected but some are adjacent in the initial DG and it is represented with the dotted links. Moreover, thanks to the parameter $\underline{\beta}$, we can also characterize the degree of overlapping (value q) of the classes '+' and 'o'.

References

[1] M. Aupetit, Learning topology with the generative gaussian graph and the em algorithm, NIPS, 2006, pp. 83–90.

[2] M. Aupetit, T. Catz, High-dimensional labeled data analysis with topology representing graphs, Neurocomputing, Elsevier 63 (2005) 139–169.

[3] D. Miller, S. Uyar, A mixture of experts classifier with learning based on both labelled and unlabelled data, NIPS, vol. 9, 1997, pp. 571–577.