Bayesian nonparametric methods are naturally well suited to functional data analysis: The case of functional classification using DPMs

Asma Rabaoui
LAPS-IMS/CNRS, Talence, France
asma.rabaoui@u-bordeaux1.fr

Hachem Kadri
SequeL-INRIA Lille, Villeneuve d'Ascq, France
hachem.kadri@inria.fr

Manuel Davy
LAGIS/CNRS/Vekia SAS, Villeneuve d'Ascq, France
mdavy@vekia.fr

When input data are continuous and we desire to take into account the inherent functional nature of these data, we often need to explore them using functional data analysis (FDA) [1]. In this work, a nonparametric Bayesian approach combining generative models and functional data analysis is presented for classifying functional data which arise naturally in a wide variety of applications. In a recent work [2], the authors have developed a generative classifier for vectorial data based on hierarchical Bayesian models and Markov chain Monte Carlo (MCMC) methods. Here, we show how to build on it a new supervised functional classifier, by merging with the functional approach in [3] based on Dirichlet Process Mixtures of Gaussian Processes (DPMGP). While generalizing learning methods from finite (vectorial) to the infinite-dimensional (functional) case is not always feasible, we show that hierarchical Bayesian classification methodology can be naturally extended from vector to functional settings. We provide theoretical and practical motivations to our approach which relies both on Dirichlet process mixtures and Gaussian processes [4, 5]. By assuming that the model parameter prior distribution is a mixture of Dirichlet processes, our method has the advantage of describing accurately a large class of probability distributions. The specification of the priors on the model parameters is, in our case, guided by mathematical and practical convenience.

Modeling functional data with Gaussian processes. Consider a supervised classification problem of functional data with K classes denoted as $C_1, ..., C_K$, each containing a set of training data (functions) $\mathbf{X}_k = \{\mathbf{x}_{1,k}(\cdot), \ldots, \mathbf{x}_{N_k,k}(\cdot)\}$, $k = 1, \ldots, K$. These functions are assumed to rely on a covariate $t \in \mathbb{R}^d$, while the function itself takes its values in $\mathbb{R}$. A practically important assumption is that a continuous signal $\mathbf{x}_{i,k}(\cdot)$ is known through a set of observed points $\{\mathbf{x}_{i,k}(t_p^{i,k}), p = 1, \cdots, T^{i,k}\} \in \mathbb{R}$ where $t_p^{i,k} \in \mathbb{R}^d$. Note that neither the sampling coordinates $t^{i,k}$'s nor the number of samples $T^{i,k}$'s are assumed to be the same for all the signals. We assume each $\mathbf{x}_{i,k}(\cdot)$, $i = 1, \ldots, N_K$ is a realization of a Gaussian process (GP) and has the following generative model

$$\mathbf{x}_{i,k}(\cdot) \sim \mathcal{GP}(\mathbf{x}_{i,k}(\cdot); \mathbf{m}_{i,k}(\cdot), \mathbf{K}_{i,k}(\cdot,\cdot)) \tag{1}$$

A Gaussian process is characterized by its mean function $\mathbf{m}_{i,k}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ and its covariance function $\mathbf{K}_{i,k}(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and its outcomes are functions. Following the work of Shi [5], we will assume that the functions $\mathbf{x}_{i,k}(\cdot)$ are generated from zero-mean GPs, that is, $\mathbf{m}_{i,k}(\cdot) \equiv 0$, and we consider covariance functions having the form

$$\mathbf{K}_{i,k}(s,t;\theta_{i,k}) = a_0 + a_1 \sum_{q=1}^{d} s_q t_q + a_2 \exp\left(-\frac{1}{2}\sum_{q=1}^{d} b_q |s_q - t_q|^2\right) \tag{2}$$

with $(d+3)$-dimensional parameter $\theta_{i,k} = \{a_0, a_1, a_2, b_1, \cdots, b_d\}$, for $(s,t) \in \mathbb{R}^d \times \mathbb{R}^d$. The parameters $\theta_{i,k}$ belong to a space Θ_k, $i = 1, \ldots, N_k$. In the following, to keep the notations short, we use θ_k to denote the set of parameters $\{\theta_{1,k}, \cdots, \theta_{i,k}, \cdots, \theta_{N_k,k}\}$ for class $\#k$.

Bayesian supervised classification framework. Let $\mathsf{L}_k(\mathbf{x}_k|\theta_k)$ denote the likelihood of observing $\mathbf{x}_k$, assumed to belong to class C_k. If we assume that $\mathbf{x}_k$ is observed through a set of T samples (therefore, a T-dimensional vector), then its likelihood is a T-dimensional multivariate normal distribution $\mathcal{N}_T$ such that $\mathsf{L}(\mathbf{x}_k|\theta_k) = \mathcal{N}_T(\mathbf{x}_k; \theta_k)$, where $|\mathbf{K}(\cdot,\cdot;\theta_k)|$ denotes the determinant of $\mathbf{K}(\cdot,\cdot;\theta_k)$. Now, following the Bayesian methodology, let $p(\theta_k|\mathbf{X}_k)$ be the posterior probability density function (pdf) for θ_k given the training data set $\mathbf{X}_k$ of class C_k. A new observation function $\mathbf{x}$ can be classified thanks to the predictive pdf $p(\mathbf{x}|\mathbf{X}_k) = \int \mathsf{L}_k(\mathbf{x}|\theta_k)p(\theta_k|\mathbf{X}_k)d\theta_k$, $k = 1, \ldots, K$. By assuming the classes have the same prior probabilities of occurrence, the Bayesian maximum *a posteriori* (MAP) classifier assigns $\mathbf{x}$ to the class $\widehat{k} = \arg\max_k p(\mathbf{x}|\mathbf{X}_k)$. To compute $p(\mathbf{x}|\mathbf{X}_k)$ for a given class C_k, the distribution $p(\theta_k|\mathbf{X}_k)$ has to be estimated: this is the aim of the training phase in Bayesian supervised classification. In this work, a nonparametric hierarchical model is used to model each signal parameters $\theta_{i,k}$ pdf, yielding $p(\theta_{i,k}|\boldsymbol{\phi}_k)$ where $\boldsymbol{\phi}_k$ is a set of hyperparameters with prior pdf $p(\boldsymbol{\phi}_k)$ (see, e.g., [2] for details). The posterior pdf of the class parameters θ_k is given by

$$p(\theta_k|\mathbf{X}_k) = \int p(\theta_k, \boldsymbol{\phi}_k|\mathbf{X}_k)d\boldsymbol{\phi}_k = \int p(\theta_k|\boldsymbol{\phi}_k)p(\boldsymbol{\phi}_k|\mathbf{X}_k)d\boldsymbol{\phi}_k \tag{3}$$

To sum up, we use a parametric model for the GP covariance $\mathbf{K}_{i,k}(\cdot,\cdot)$, then we adopt the Bayesian framework and place a prior directly on the parameters of $\mathbf{K}_{i,k}(\cdot,\cdot)$. Dirichlet Process Mixture (DPM) are suitable tools to model prior knowledge over parameters. This allows for a flexible nonparametric modeling framework. Therefore, in the following, $p(\theta_k|\boldsymbol{\phi}_k)$ is assumed to be a DPM.

Hierarchical DPM prior for the model parameters. A DPM model can basically be thought of as a simple mixture model given by the mixed pdf $\theta_{i,k} \sim p(\theta_{i,k}|\boldsymbol{\phi}_{i,k})$ and prior $\boldsymbol{\phi}_{i,k} \sim \mathbb{G}(\boldsymbol{\phi}_{i,k})$ where $\mathbb{G}$ itself is the random outcome of a Dirichlet Process $\mathcal{DP}(\mathbb{G}_0(\varphi_k), \alpha)$ (that is, a probabilistic distribution over probabilistic distributions). In summary, we have :

$$\begin{aligned} \theta_{i,k}|\boldsymbol{\phi}_{i,k} &\sim p(\theta_{i,k}|\boldsymbol{\phi}_{i,k}) \\ \boldsymbol{\phi}_{i,k}|\mathbb{G} &\sim \mathbb{G}(\cdot) \\ \mathbb{G}|\psi_k &\sim \mathcal{DP}(\mathbb{G}_0(\varphi_k), \alpha) \end{aligned}$$

where $\psi_k = \{\alpha, \varphi_k\}$ is the hyperparameter vector, with φ_k a given parameter vector for $\mathbb{G}_0$. The advantage of applying the DP prior [6] to hierarchical models has been addressed extensively in the statistics literature, mostly in recent years, see for example [7]. By integrating over $\mathbb{G}$ through the so called polya urn representation, we see that the joint distribution of $\boldsymbol{\phi}_k = \{\boldsymbol{\phi}_{1,k}, \cdots, \boldsymbol{\phi}_{N_k,k}\}$ may be factored into a product of successive conditional distributions of the following form:

$$\boldsymbol{\phi}_{i,k}|\boldsymbol{\phi}_{-i,k}, \psi_k \sim \frac{1}{\alpha + N_k - 1} \sum_{j=1, j\neq i}^{N_k} \delta_{\boldsymbol{\phi}_{j,k}} + \frac{\alpha}{\alpha + N_k - 1}\mathbb{G}_0 \tag{4}$$

where $\boldsymbol{\phi}_{-i,k}$ denotes $\boldsymbol{\phi}_k \backslash \boldsymbol{\phi}_{i,k}$. This factorization implies that $\boldsymbol{\phi}_{i,k}$ has discrete, though infinite, support.

References

[1] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis, 2nd ed*, Springer Verlag, New York, 2005.

[2] M. Davy and J.Y. Tourneret, "Generative supervised classification using dirichlet process priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1781–1794, 2010.

[3] E. Jackson, M. Davy, A. Doucet, and W. Fitzgerald, "Bayesian unsupervised signal classification by dirichlet process mixtures of gaussian processes," in *ICASSP'07*, April 15-20, 2007, pp. 1077–1080.

[4] C. E. Rasmussen, "Gaussian processes in machine learning," *Advanced Lectures on Machine Learning*, pp. 63–71, 2006.

[5] J. Q. Shi and T. Choi, *Gaussian process regression analysis for functional data*, Chapman & Hall/CRC Press, 2011.

[6] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 249–265, 2000.

[7] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian nonparametric models with applications," in *Bayesian Nonparametrics: Principles and Practice*, N. Hjort, C. Holmes, P. Müller, and S. Walker, Eds. Cambridge University Press, 2010.