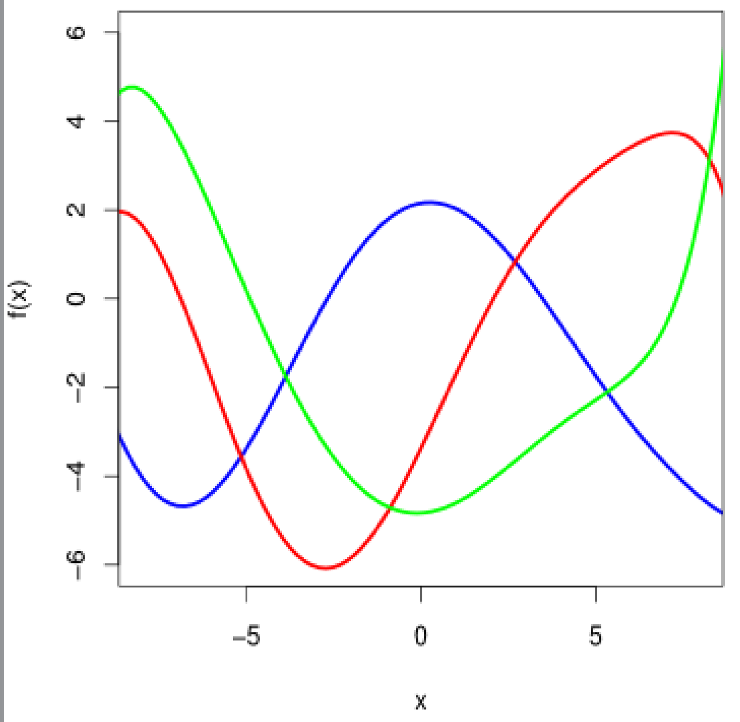
Robust Multi-Class Gaussian Process Classification

Daniel Hernández-Lobato(1), José Miguel Hernández-Lobato(2) and Pierre Dupont(1)

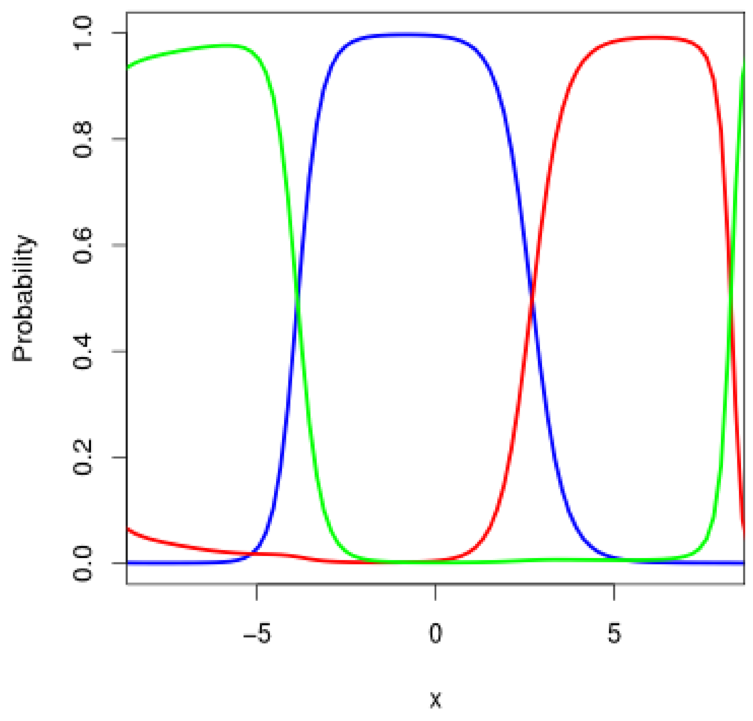
(1) - Machine Learning Group, ICTEAM Institute, Université catholique de Louvain
(2) - Computational and Biological Learning Group, Department of Engineering, University of Cambridge

Neural Information Processing Systems Foundation

Introduction to Multi-Class Gaussian Process Classifiers



Association Rule



- **Bayes' theorem** is used to infer each latent function from the data.
- A **Gaussian process** prior is assumed for each latent function.
- Typically, the rule **only considers** at most errors in the labels of the data **near the decision boundaries**, which can produce **over-fitting**.
- Labeling errors can also be accounted for by considering **additive Gaussian noise** around each latent function. However, this leads to the **same problem**.
- We propose a Robust Multi-class Gaussian process classifier (RMGPC) which is **robust** to errors located **far away** from the decision boundaries.

Robust Multi-Class Gaussian Process Classification

Consider n training instances in the form of a collection of feature vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with associated labels $\mathbf{y} = \{y_1, \ldots, y_n\}$, where $y_i \in \mathcal{C} = \{1, \ldots, l\}$. When there is **no noise**, we assume:

$$y_i = \arg\max_k \quad f_k(\mathbf{x})\,.$$

We introduce a set of binary latent variables $\mathbf{z} = \{z_1, \ldots, z_n\}$ to indicate when this rule is satisfied ($z_i = 0$) in practice or not ($z_i = 1$). In the latter case, we consider that $(\mathbf{x}_i, y_i)$ is an **outlier** and that $\mathbf{x}_i$ has been assigned a class **sampled uniformly** from $\mathcal{C}$. The likelihood for the latent functions $\mathbf{f} = \{f_1, \ldots, f_l\}$ is:

$$\mathcal{P}(\mathbf{y}|\mathbf{X}, \mathbf{z}, \mathbf{f}) = \prod_{i=1}^{n} \left[\prod_{k \neq y_i} \Theta(f_{y_i}(\mathbf{x}_i) - f_k(\mathbf{x})) \right]^{1-z_i} \left[\frac{1}{l}\right]^{z_i},$$

where $\Theta(\cdot)$ is a step function. The likelihood **only depends** on the **number of prediction errors made** and not on their location.

The prior for $\mathbf{z}$ is a multivariate Bernoulli distribution with parameter ρ, and the prior for ρ is a conjugate beta distribution. The prior for the latent functions is a product of l **independent Gaussian processes**.

We are interested in computing:

$$\mathcal{P}(\rho, \mathbf{z}, \mathbf{f}|\mathbf{y}, \mathbf{X}) = \frac{\mathcal{P}(\mathbf{y}|\mathbf{X}, \mathbf{z}, \mathbf{f})\mathcal{P}(\mathbf{z}|\rho)\mathcal{P}(\rho)\mathcal{P}(\mathbf{f})}{\mathcal{P}(\mathbf{y}|\mathbf{X})}\,.$$

for making **predictions** and for estimating the **probability** that a given instance is an **outlier**. Namely:

$$\mathcal{P}(z_i = 1|\mathbf{y}, \mathbf{X}) = \sum_{\mathbf{z}_j} \int \mathcal{P}(\rho, \mathbf{z}, \mathbf{f}|\mathbf{y}, \mathbf{X}) d\rho d\mathbf{f}\,, \quad \text{with } j \neq i.$$

The model evidence, $\mathcal{P}(\mathbf{y}|\mathbf{X})$, is useful for **hyper-parameter optimization**.

Expectation Propagation

Approximates the exact posterior using a parametric distribution:

$$\mathcal{Q}(\rho, \mathbf{z}, \mathbf{f}) = \prod_{k=1}^{l} \mathcal{N}(\mathbf{f}_k|\mu_k, \mathbf{\Sigma}_k)\mathrm{Bern}(\mathbf{z}|\mathbf{p})\mathrm{beta}(\rho|a, b)\,,$$

where $\mathcal{N}(\cdot|\mu_k, \mathbf{\Sigma}_k)$ denotes a multivariate Gaussian with mean μ_k and covariance matrix $\mathbf{\Sigma}_k$, $\mathrm{Bern}(\cdot|\mathbf{p})$ denotes a multi-variate Bernoulli with parameter vector $\mathbf{p}$ and $\mathrm{beta}(\cdot|a, b)$ denotes a beta distribution with parameters a and b.

The parameters of $\mathcal{Q}$ are determined by approximately **minimizing**:

$$\text{Kullback-Liebler}\left(\mathcal{P}(\rho, \mathbf{z}, \mathbf{f}|\mathbf{y}, \mathbf{X}) || \mathcal{Q}(\rho, \mathbf{z}, \mathbf{f})\right)\,.$$

Expectation propagation also **approximates** the model evidence $\mathcal{P}(\mathbf{y}|\mathbf{X})$. Furthermore, it is possible to evaluate the **gradient** of such approximation with respect to the parameters of the prior. This is very useful, for example, to **find** the parameters of the **covariance matrices** of $\mathcal{P}(\mathbf{f})$.

The **total cost** of expectation propagation is $\mathcal{O}(ln^3)$ since we assume a factorized approximation.

Data-sets and Balanced Class Rate

Dataset	# Instances	# Attributes	# Classes	# Source
New-thyroid	215	5	3	UCI
Wine	178	13	3	UCI
Glass	214	9	6	UCI
SVMguide2	319	20	3	LIBSVM

BCR: **Average of the l accuracies** computed on the data instances of each class.

Experimental Results: BCR as a Function of the Noise Level η

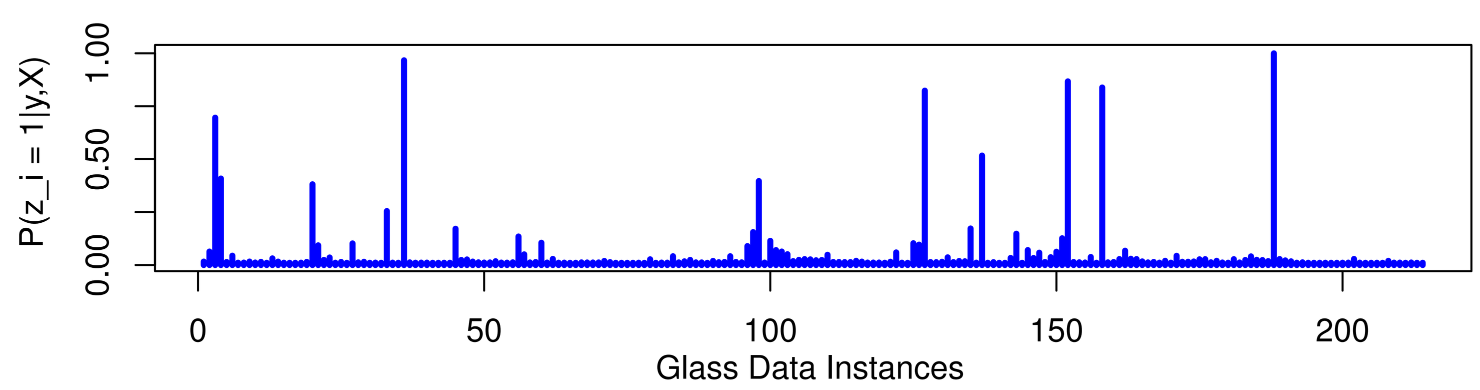
Average BCR in % of each method for each problem, as a function of η.

Dataset	RMGPC	SMGPC	HTPC	RMGPC	SMGPC	HTPC
		$\eta = 0\%$			$\eta = 5\%$	
New-thyroid	**94.2±4.5**	**93.9±4.4**	**90.0±5.5** ◁	**92.7±4.9**	**90.7±5.8** ◁	**89.7±6.1** ◁
Wine	**98.0±1.6**	**98.0±1.6**	**97.3±2.0** ◁	**97.5±1.7**	**97.3±2.0**	**96.6±2.2** ◁
Glass	**65.2±7.7**	**60.6±8.6** ◁	**59.5±8.0** ◁	**63.5±8.0**	**58.9±8.0** ◁	**57.9±7.5** ◁
SVMguide2	**76.3±4.1**	**74.6±4.2** ◁	**72.8±4.1** ◁	**75.6±4.3**	**73.8±4.4** ◁	**71.9±4.5** ◁
		$\eta = 10\%$			$\eta = 20\%$	
New-thyroid	**92.3±5.4**	**89.0±5.5** ◁	**88.3±6.6** ◁	**89.5±6.0**	**85.9±7.4** ◁	**85.7±7.7** ◁
Wine	**97.0±2.2**	**96.4±2.6**	**95.6±4.6** ◁	**96.6±2.7**	**95.5±2.6** ◁	**95.1±3.0** ◁
Glass	**63.9±7.9**	**58.0±7.4** ◁	**55.7±7.7** ◁	**59.7±8.3**	**55.5±7.3** ◁	**52.8±7.8** ◁
SVMguide2	**74.9±4.4**	**72.8±4.7** ◁	**71.5±4.7** ◁	**72.8±5.1**	**71.4±5.0** ◁	**67.5±5.6** ◁

RMGPC: Robust Multi-class Gaussian Process Classifier.
SMGPC: Standard Multi-class Gaussian Process Classifier.
HTPC: Heavy-tailed Process Classifier.

When the performance of a method is significantly different from the performance of RMGPC, as estimated by a Wilcoxon rank test (p-value $< 1\%$), the corresponding BCR is marked with the symbol ◁.

Outlier Identification: Glass Data-set



Posterior probability that each data instance form the *Glass* dataset is an outlier.

Average test error in % of each method on each data instance more likely to be an outlier.

		Glass Data Instances						
		3-rd	36-th	127-th	137-th	152-th	158-th	188-th
Test Error	RMGPC	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
	SMGPC	100.0±0.0	92.0±5.5	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
	HTPC	100.0±0.0	84.0±7.5	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
$\mathcal{P}(z_i = 1\|\mathbf{y}, \mathbf{X})$		0.69	0.96	0.82	0.51	0.86	0.83	1.00

Conclusions

- RMGPC considers only **the number of errors made**, and not the distance of such errors to the decision boundaries of the resulting classifier.
- RMGPC can **identify** the training instances that are **more likely to be outliers**.
- **Approximate inference** can be efficiently carried out using **expectation propagation**.
- When **noise is injected** in the labels, RMGPC **performs better** than other alternatives which consider latent Gaussian noise or noise which follows a heavy-tail distribution.
- When there is **no noise** in the data, RMGPC performs **better or equivalent** to these other alternatives.

References

1. Christopher K. I. Williams and David Barber. Bayesian classification with Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12):1342-1351, 1998.
2. Hyun-Chul Kim and Zoubin Ghahramani. Bayesian Gaussian process classification with the EM-EP algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12):1948-1959, 2006.
3. F. L. Wauthier and M. I. Jordan. Heavy-Tailed Process Priors for Selective Shrinkage. In J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor, and A. Culotta, editors, Advances in Neural Information Processing Systems 23, pages 2406-2414, 2010.
4. Matthias Seeger and Michael I. Jordan. Sparse Gaussian process classification with multiple classes. Technical report, University of California, Berkeley, 2004.
5. T. Minka. A Family of Algorithms for approximate Bayesian Inference. PhD thesis, Massachusetts Institute of Technology, 2001. Computer Science, pages 896-905. Springer Berlin / Heidelberg, 2008.
6. Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary Gaussian process classification. Journal of Machine Learning Research, 6:1679-1704, 2005.
7. Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2006.