Bayesian Active Learning for Gaussian Process Classification

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, Máté Lengyel

Computational and Biological Learning Lab, Department of Engineering, University of Cambridge

Computational and Biological Learning
University of Cambridge

Introduction to Active Learning

- AL concerns designing learners that choose their training data.
- Applications include: sensor placement, information extraction, speech recognition, cognitive science, collaborative filtering, quantum tomography etc.
- Referred to as 'Optimal Experimental Design' in statistics.
- We revisit the Information Theoretic approach.

Bayesian Information Theoretic AL

Latent parameters $\boldsymbol{\theta} \in \Theta$ govern dependence of $\boldsymbol{y} \in \mathcal{Y}$ on input $\boldsymbol{x} \in \mathcal{X}$ (discriminative model). Observe data, $\mathcal{D}$, Bayes rule yields the posterior distribution over parameters $p(\theta|\mathcal{D})$. Select $\boldsymbol{x}_i$ (myopically) to minimize the posterior entropy:

$$\boldsymbol{x}_{\text{new}} = \arg\max_{\boldsymbol{x}} \mathrm{H}[\boldsymbol{\theta}|\mathcal{D}] - \mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x}\mathcal{D})}\left[\mathrm{H}[\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x},\mathcal{D}]\right]$$

Problems:

- Parameter space is often high dimensional, for GPs, it is infinite dimensional.
- Posterior updates required for all input/output combinations ($\mathcal{O}(N_{\boldsymbol{x}}N_{\boldsymbol{y}})$).

Solution: Rearrange to Dataspace

$$\begin{aligned}
&\mathrm{H}[\boldsymbol{\theta}|\mathcal{D}] - \mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x}\mathcal{D})}\left[\mathrm{H}[\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x},\mathcal{D}]\right]\\
&= \mathrm{I}[\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x},\mathcal{D}]\\
&= \mathrm{H}[\boldsymbol{y}|\boldsymbol{x},\mathcal{D}] - \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}\left[\mathrm{H}[\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}]\right]
\end{aligned}$$

- Output space is often low dimensional and $\mathcal{O}(1)$ posterior updates required.
- We call this Bayesian Active Learning by Disagreement (BALD).
- Aside: equivalent to the Jensen-Shannon divergence.

Review of Gaussian Processes

GPs provide a prior over functions $f : \mathcal{X} \to \mathbb{R}$:

$$f \sim \mathrm{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$

For regression/classification define likelihood functions respectively:

$$\boldsymbol{y}|\boldsymbol{x}, f \sim \mathcal{N}(f(\boldsymbol{x}), \sigma^2), \quad \boldsymbol{y}|\boldsymbol{x}, f \sim \mathrm{Bernoulli}(\Phi(f(\boldsymbol{x})))\ \Phi(z) = \int_{-\infty}^{z} \mathcal{N}(0,1)\mathrm{d}z$$

For classification, posterior is intractable, make a Gaussian approximation (Expectation Propagation (EP), the Laplace approximation, Variational methods).

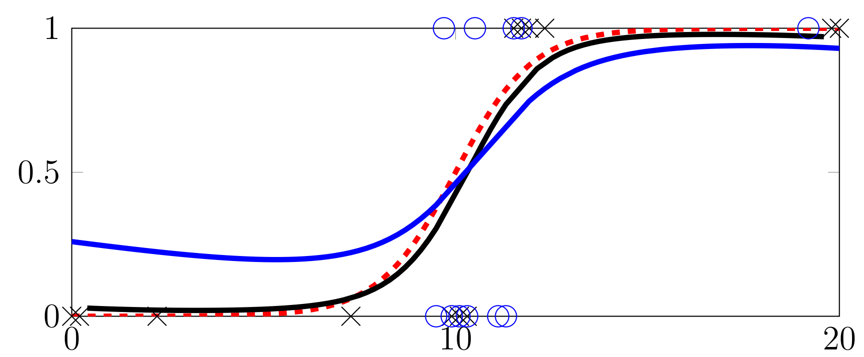

Figure 1: Toy active GPC problem. True generating function is (- - -). 15 actively selected samples are drawn using both BALD (×) and Maximum Entropy Sampling (○). The predictive distributions from BALD and MES are (——) and (——) respectively.

BALD for GPC

Two terms need to be computed:

$$\begin{aligned}
\mathrm{H}[y|\boldsymbol{x},\mathcal{D}] &\overset{1}{\approx} \mathrm{h}\left(\int \Phi(f_{\boldsymbol{x}})\mathcal{N}(f_{\boldsymbol{x}}|\mu_{\boldsymbol{x},\mathcal{D}},\sigma^2_{\boldsymbol{x},\mathcal{D}})df_{\boldsymbol{x}}\right)\\
&= \mathrm{h}\left(\Phi\left(\frac{\mu_{\boldsymbol{x},\mathcal{D}}}{\sqrt{\sigma^2_{\boldsymbol{x},\mathcal{D}}+1}}\right)\right)\\
\mathbb{E}_{f\sim p(f|\mathcal{D})}\left[\mathrm{H}[\boldsymbol{y}|f]\right] &\overset{1}{\approx} \int \mathrm{h}(\Phi(f_{\boldsymbol{x}}))\mathcal{N}(f_{\boldsymbol{x}}|\mu_{\boldsymbol{x},\mathcal{D}},\sigma^2_{\boldsymbol{x},\mathcal{D}})df_{\boldsymbol{x}}\\
&\overset{2}{\approx} \int \exp\left(-\frac{f_{\boldsymbol{x}}^2}{\pi\ln 2}\right)\mathcal{N}(f_{\boldsymbol{x}}|\mu_{\boldsymbol{x},\mathcal{D}},\sigma^2_{\boldsymbol{x},\mathcal{D}})df_{\boldsymbol{x}}\\
&= \frac{C}{\sqrt{\sigma^2_{\boldsymbol{x},\mathcal{D}}+C^2}}\exp\left(-\frac{\mu^2_{\boldsymbol{x},\mathcal{D}}}{2\left(\sigma^2_{\boldsymbol{x},\mathcal{D}}+C^2\right)}\right)
\end{aligned}$$

where:

$$\mathrm{h}(p) = -p\log p - (1-p)\log(1-p), \quad C = \sqrt{\frac{\pi\ln 2}{2}}$$

- $\overset{1}{\approx}$ is a Gaussian approximation to intractable posterior.
- $\overset{2}{\approx}$ is a squared exponential approximation to $\mathrm{h}(\Phi(f_{\boldsymbol{x}}))$ (binary entropy of Normal cdf).
- The objective function is smooth and differentiable.

Summary

1. Apply an approximate inference algorithm to get $\mu_{\boldsymbol{x},\mathcal{D}}$ and $\sigma_{\boldsymbol{x},\mathcal{D}}$ for each point of interest $\boldsymbol{x}$.
2. Select $\boldsymbol{x}$ that maximises:

$$\mathrm{h}\left(\Phi\left(\frac{\mu_{\boldsymbol{x},\mathcal{D}}}{\sqrt{\sigma^2_{\boldsymbol{x},\mathcal{D}}+1}}\right)\right) - \frac{C}{\sqrt{\sigma^2_{\boldsymbol{x},\mathcal{D}}+C^2}}\exp\left(-\frac{\mu^2_{\boldsymbol{x},\mathcal{D}}}{2\left(\sigma^2_{\boldsymbol{x},\mathcal{D}}+C^2\right)}\right) \tag{1}$$

Related Algorithms

The following algorithms are closely related, often approximating the BALD objective:

- Uncertainty Sampling [Lewis and Gale, 1994] / Maximum Entropy Sampling [Sebastiani and Wynn, 2000].
- The Informative Vector Machine [Lawrence and Herbrich, 2001].
- Query by Committee [Freund et al., 1997].
- SVM-based active learning [Tong and Koller, 2001].

Extensions

Further work that we have performed:

- Comparison to decision theoretic algorithms.
- Hyperparameter learning ($\boldsymbol{\theta}^+, \boldsymbol{\theta}^-$ = params of interest, nuisance parameters): $\mathrm{H}\left[\mathbb{E}_{p(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-|\mathcal{D})}[\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}^+,\boldsymbol{\theta}^-]\right] - \mathbb{E}_{p(\boldsymbol{\theta}^+|\mathcal{D})}\left[\mathrm{H}\left[\mathbb{E}_{p(\boldsymbol{\theta}^-|\boldsymbol{\theta}^+,\mathcal{D})}[\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}^+,\boldsymbol{\theta}^-]\right]\right]$
- Multiclass: combine criteria for K one-versus-all classifiers.
- Preference Learning: extend GP methods of [Chu and Ghahramani, 2005].

Results

Experiments run on *pool-based* active learning. Test set accuracy plotted is against number of queries.

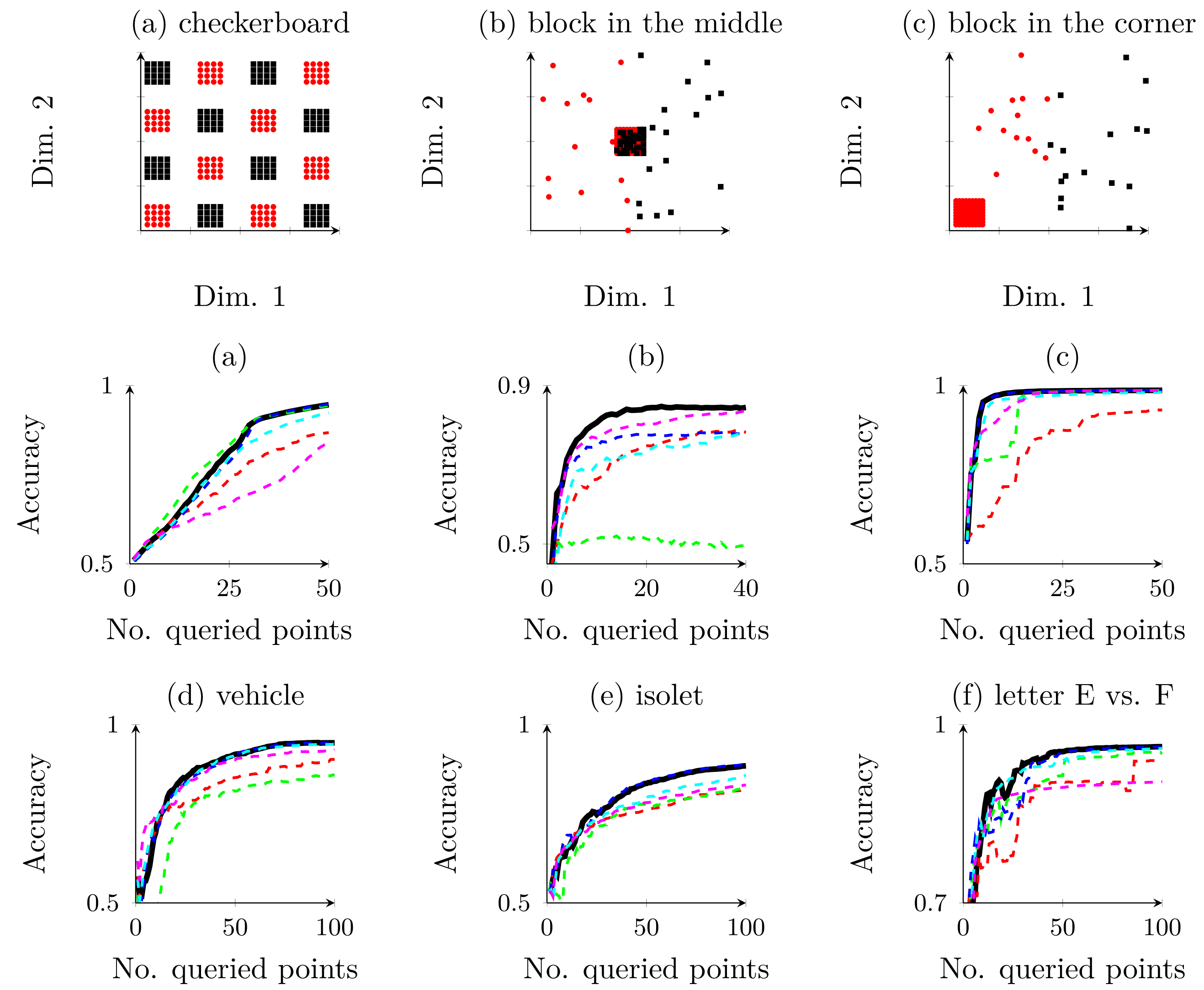

Figure 2: *Top:* Three 2D artificial datasets designed to test the algorithms in pathological scenarios. *Middle:* Results for corresponding artificial datasets using BALD (——), random query (- - -), MES (- - -), IVM (- - -), QBC (- - -), active SVM (- - -). *Bottom:* Results on three real-world datasets.

Acknowledgements

This work is made possible by our sponsors: Google, Europe (Neil Houlsby) and Trinity College, Cambridge (Ferenc Huszár) and the Wellcome Trust (Máté Lengyel).

References

[Chu and Ghahramani, 2005] Chu W, Ghahramani Z. 2005.
Preference learning with gaussian processes.
In: Proceedings of the 22nd international conference on Machine learning. ACM, pp 137–144.

[Freund et al., 1997] Freund Y, Seung H, Shamir E, Tishby N. 1997.
Selective sampling using the query by committee algorithm.
Machine Learning 28:133–168.

[Lawrence and Herbrich, 2001] Lawrence N, Herbrich R. 2001.
A sparse Bayesian compression scheme-the informative vector machine.
In: NIPS 2001 workshop on kernel methods. Citeseer.

[Lewis and Gale, 1994] Lewis D, Gale W. 1994.
A sequential algorithm for training text classifiers.
In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Springer-Verlag New York, Inc., pp 3–12.

[Sebastiani and Wynn, 2000] Sebastiani P, Wynn H. 2000.
Maximum entropy sampling and optimal Bayesian experimental design.
Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62:145–157.

[Tong and Koller, 2001] Tong S, Koller D. 2001.
Support vector machine active learning with applications to text classification.
Journal of Machine Learning Research 2:45–66.