Bayesian Active Learning

Assaf Dvora

May 3, 2022

1 Introduction

Bayesian Active Learning by Disagreement (BALD) [1] is an information theoretic approach for active learning designed for the Gaussian Process Classifier. The following sections describes this approach.

2 Bayesian Information Theoretic Active Learning

We consider a fully discriminative model where the goal of active learning is to discover the dependencies of some variable $y \in \mathcal{Y}$ on an input variable $\boldsymbol{x} \in \mathcal{X}$. The key idea in active learning is that the learner chooses the input queries $\boldsymbol{x}_i \in \mathcal{X}$ and observes the system's responses y_i, rather than passively receiving $(\boldsymbol{x}_i, y_i)$ pairs.

Within a Bayesian framework we assume existence of some latent parameters, $\boldsymbol{\theta}$ (e.g. GP latent function), that controls the dependence between inputs and outputs through the conditional distribution $p(y|\boldsymbol{x}, \boldsymbol{\theta})$. Having observed data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, a posterior distribution over the latent parameters is inferred, $p(\boldsymbol{\theta}|\mathcal{D})$. The central goal of information theoretic active learning is to reduce the number of possible hypotheses maximally fast. I.e., minimizing the uncertainty about the parameters using Shannon's entropy. Data points $\mathcal{D}'$ are selected that satisfy

$$\arg\min_{\mathcal{D}'} \mathrm{H}[\boldsymbol{\theta}|\mathcal{D}'] = -\int p(\boldsymbol{\theta}|\mathcal{D}') \log p(\boldsymbol{\theta}|\mathcal{D}') \tag{1}$$

Since solving this problem is NP-hard, a greedy policy is often used. Therefore the objective is to seek that data point $\boldsymbol{x}$ that maximizes the decrease in posterior entropy [1]

$$\arg\max_{\boldsymbol{x}^*} \{\mathrm{H}[\boldsymbol{\theta}|\mathcal{D}] - \mathrm{H}[\boldsymbol{\theta}|y, \boldsymbol{x}, \mathcal{D}]\} \tag{2}$$

In practice, y is unknown and therefore the second term in Eq. (2) is replaced by the expected posterior entropy

$$\arg\max_{\boldsymbol{x}^*} \{\mathrm{H}[\boldsymbol{\theta}|\mathcal{D}] - \mathbb{E}_{y \sim p(y|\boldsymbol{x}, \mathcal{D})}[\mathrm{H}[\boldsymbol{\theta}|y, \boldsymbol{x}, \mathcal{D}]]\} \tag{3}$$

This solution however arises a computational difficulty: if $N_{\boldsymbol{x}}$ data points are under consideration, and N_y responses may be seen, then $\mathcal{O}(N_{\boldsymbol{x}} N_y)$ posterior updates are required.

An important insight arises if we note that the objective in Eq. (2) is equal to the mutual information between the latent parameters $\boldsymbol{\theta}$ and the unknown responses y^*, $I[\boldsymbol{\theta}, y^*|\boldsymbol{x}^*, \mathcal{D}]$. Using this insight it is simple to show that the objective can be can be rearranged to compute the entropies in the y space (see Appendix A.)

$$\arg\max_{\boldsymbol{x}^*}\{\mathrm{H}[y|\boldsymbol{x}, \mathcal{D}] - \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{H}[y|\boldsymbol{\theta}, \boldsymbol{x}]]\} \tag{4}$$

Eq. (4) overcomes the computational difficulty described in Eq. (3). The latent parameter $\boldsymbol{\theta}$ is now conditioned only on $\mathcal{D}$, so only $\mathcal{O}(1)$ posterior updates are required. Eq. (4) also provides us with an interesting intuition about the objective; we seek the $\boldsymbol{x}$ for which the model is marginally most uncertain about y (high $\mathrm{H}[y|\boldsymbol{x}, \mathcal{D}]$), but for which, given individual settings of the parameters, y is confident (low $\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{H}[y|\boldsymbol{\theta}, \boldsymbol{x}]]$). Further analysis of the objective function is given in Section (5). We note that the argument in (4) is non-negative as we show in Appendix B.

3 Gaussian Process Classifier (GPC)

In this section we introduce the Gaussian Process Classifier (GPC) [2]. The probabilistic model underlying GPC is as follows

$$\begin{aligned} f(\boldsymbol{x}) &\sim \mathcal{GP}(0, k(\cdot, \cdot)) \\ y|\boldsymbol{x}, f &\sim Bernoulli(\Phi(f(\boldsymbol{x})) \end{aligned}$$

where Φ is the Gaussian CDF. The latent parameter, now called $f(\boldsymbol{x})$, is a function $\mathcal{X} \to \mathbb{R}$, and is assigned a GP prior.

Inference in the GPC model is non-Gaussian and intractable. Throughout this section we will assume that a Gaussian approximation of the posterior (e.g. Laplace approximation) is used. The posterior predictive distribution of $f(\boldsymbol{x})$ is than given by

$$p(f(\boldsymbol{x})|\boldsymbol{x}, \mathcal{D}) = \mathcal{N}(f(\boldsymbol{x})|\mu_{\boldsymbol{x}}, \sigma^2_{\boldsymbol{x}}) \tag{5}$$

The posterior predictive of y is then given by [2]

$$p(y|\boldsymbol{x}, \mathcal{D}) = \Phi\left(\frac{\mu_{\boldsymbol{x}}}{\sqrt{1+\sigma^2_{\boldsymbol{x}}}}\right) \tag{6}$$

4 GPC Active Learning

In the GPC case, the objective function given in Eq. (4), takes the following form

$$\arg\max_{\boldsymbol{x}^*}\{\mathrm{H}[y|\boldsymbol{x}, \mathcal{D}] - \mathbb{E}_{f_{\boldsymbol{x}}\sim p(f_{\boldsymbol{x}}|\boldsymbol{x}, \mathcal{D})}[\mathrm{H}[y|f_{\boldsymbol{x}}]]\} \tag{7}$$

The first term in (7) which is the entropy of the posterior predictive of y can be handled analytically:

$$\mathrm{H}[y|\boldsymbol{x},\mathcal{D}] = -p(y|\boldsymbol{x},\mathcal{D})\log p(y|\boldsymbol{x},\mathcal{D}) - (1-p(y|\boldsymbol{x},\mathcal{D}))\log(1-p(y|\boldsymbol{x},\mathcal{D}))$$

The second term in (7) involves integration over f space of the entropy of $y|f_{\boldsymbol{x}}$

$$\mathbb{E}_{f_{\boldsymbol{x}}}[\mathrm{H}[y|f_{\boldsymbol{x}}]] = \int \mathrm{H}[y|f_{\boldsymbol{x}}]\mathcal{N}(f_{\boldsymbol{x}}|\mu_{\boldsymbol{x}},\sigma^2_{\boldsymbol{x}})df_{\boldsymbol{x}} \tag{8}$$

where the entropy of $y|f_{\boldsymbol{x}}$ is given by

$$\begin{aligned}\mathrm{H}[y|f_{\boldsymbol{x}}] &= -p(y|f_{\boldsymbol{x}})\log p(y|f_{\boldsymbol{x}}) - (1-p(y|f_{\boldsymbol{x}}))\log(1-p(y|f_{\boldsymbol{x}})) &(9)\\ &= -\Phi(f(\boldsymbol{x})\log\Phi(f(\boldsymbol{x}) - (1-\Phi(f(\boldsymbol{x}))\log(1-\Phi(f(\boldsymbol{x})) &(10)\end{aligned}$$

5 Numerical Integration

To compute the objective function (7), one must compute the expectation $\mathbb{E}_{f_{\boldsymbol{x}}\sim p(f_{\boldsymbol{x}}|\boldsymbol{x},\mathcal{D})}\{\mathrm{H}[y|f_{\boldsymbol{x}}]\}$. Using the strong law of large number (SLLN) one can approximate the expectation by summation over samples from the posterior. Alternatively, numerical integration can be used to solve the integral

$$\begin{aligned}\mathbb{E}_{f_{\boldsymbol{x}}}[\mathrm{H}[y|f_{\boldsymbol{x}}]] = \int \mathrm{H}[y|f_{\boldsymbol{x}}]\mathcal{N}(f_{\boldsymbol{x}}|\mu_{\boldsymbol{x}},\sigma^2_{\boldsymbol{x}})df_{\boldsymbol{x}} = \\ = \frac{1}{\sqrt{2\pi\sigma^2_{\boldsymbol{x}}}}\int \mathrm{H}[y|f_{\boldsymbol{x}}]\exp\left[-\frac{1}{2}\left(\frac{f_{\boldsymbol{x}}-\mu_{\boldsymbol{x}}}{\sigma_{\boldsymbol{x}}}\right)^2\right]df_{\boldsymbol{x}} \qquad (11)\end{aligned}$$

By using change of variables we obtain

$$z = \frac{f_{\boldsymbol{x}}-\mu_{\boldsymbol{x}}}{\sigma_{\boldsymbol{x}}}, \quad \frac{dz}{df_x} = \frac{1}{\sigma_{\boldsymbol{x}}} \tag{12}$$

and (11) becomes

$$\mathbb{E}_{f_{\boldsymbol{x}}}[\mathrm{H}[y|f_{\boldsymbol{x}}]] = \frac{1}{\sqrt{2\pi}}\int \mathrm{H}[y|\sigma_{\boldsymbol{x}}z+\mu_{\boldsymbol{x}}]\exp[-\frac{z^2}{2}]dz \tag{13}$$

The integral can now be approximated using numerical integration

$$\mathbb{E}_{f_{\boldsymbol{x}}}[\mathrm{H}[y|f_{\boldsymbol{x}}]] \approx \frac{1}{\sqrt{2\pi}}\sum_{i=1}^{n}\mathrm{H}[y|\sigma_{\boldsymbol{x}}z_i+\mu_{\boldsymbol{x}}]\exp[-\frac{z_i^2}{2}]\Delta z \tag{14}$$

Appendix A.

Using Bayes rule we can express the objective in Eq. (3) in the y space

$$
\begin{aligned}
&\arg\max_{\boldsymbol{x}}\{H[\boldsymbol{\theta}|\mathcal{D}] - \mathbb{E}_{y\sim p(y|\boldsymbol{x},D)}[H[\boldsymbol{\theta}|y,\boldsymbol{x},\mathcal{D}]]\} = \\
&\arg\min_{\boldsymbol{x}} \mathbb{E}_{y\sim p(y|\boldsymbol{x},D)}[H[\boldsymbol{\theta}|y,\boldsymbol{x},\mathcal{D}]] = \\
&\arg\min_{\boldsymbol{x}} \mathbb{E}_{y\sim p(y|\boldsymbol{x},D)}[\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|y,\boldsymbol{x},\mathcal{D})}[-\log p(\boldsymbol{\theta}|y,\boldsymbol{x},\mathcal{D})]] = \\
&\arg\max_{\boldsymbol{x}} \mathbb{E}_{y\sim p(y|\boldsymbol{x},D)}[\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|y,\boldsymbol{x},\mathcal{D})}[\log p(y|\boldsymbol{\theta},x) + \log p(\boldsymbol{\theta}|\mathcal{D}) - p(y|\boldsymbol{x},\mathcal{D})]] = \\
&\arg\max_{\boldsymbol{x}}\{H[y|\boldsymbol{x},\mathcal{D}]
\end{aligned}
$$

Appendix B.

In this section we show that the argument in the objective function (4) is non-negative.

Denote with $h(\cdot)$ the convex function $h(p) = p\log p, p > 0$ and denote with $\psi_{k,\boldsymbol{x}}(\cdot)$ the random variable $\psi_{k,\boldsymbol{x}}(\boldsymbol{\theta}) = p(y = k|\boldsymbol{\theta},\boldsymbol{x})$, using Jensen inequality we obtain

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{H}[y|\boldsymbol{\theta},\boldsymbol{x}]] &= \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}\{-\sum_{k=1}^{K} h[\psi_{k,\boldsymbol{x}}(\boldsymbol{\theta})]\} = \\
&= -\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}\{h[\psi_{k,\boldsymbol{x}}(\boldsymbol{\theta})]\} = \\
&= -\sum_{k=1}^{K} \int h[\psi_{k,\boldsymbol{x}}(\boldsymbol{\theta})]p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \leqslant \\
&\leqslant -\sum_{k=1}^{K} h[\int \psi_{k,\boldsymbol{x}}(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}] = \\
&= -\sum_{k=1}^{K} h[p(y = k|\boldsymbol{x},\mathcal{D}] = \mathrm{H}[y|\boldsymbol{x},\mathcal{D}]
\end{aligned}
$$

Thus, $\mathrm{H}[y|\boldsymbol{x},\mathcal{D}] \geqslant \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{H}[y|\boldsymbol{\theta},\boldsymbol{x}]]$.

References

[1] Houlsby, Neil, et al. "Bayesian active learning for classification and preference learning." arXiv preprint arXiv:1112.5745 (2011).

[2] Rasmussen, C. and Williams, C. (2005). Gaussian Processes for Machine Learning. The MIT Press.