Bayesian Active Learning for Maximal Information Gain on Model Parameters

Kasra Arnavaz, Aasa Feragen, Oswin Krause, Marco Loog

Paper number: 2937

Active Learning vs Random Sampling

Data Modeling Process

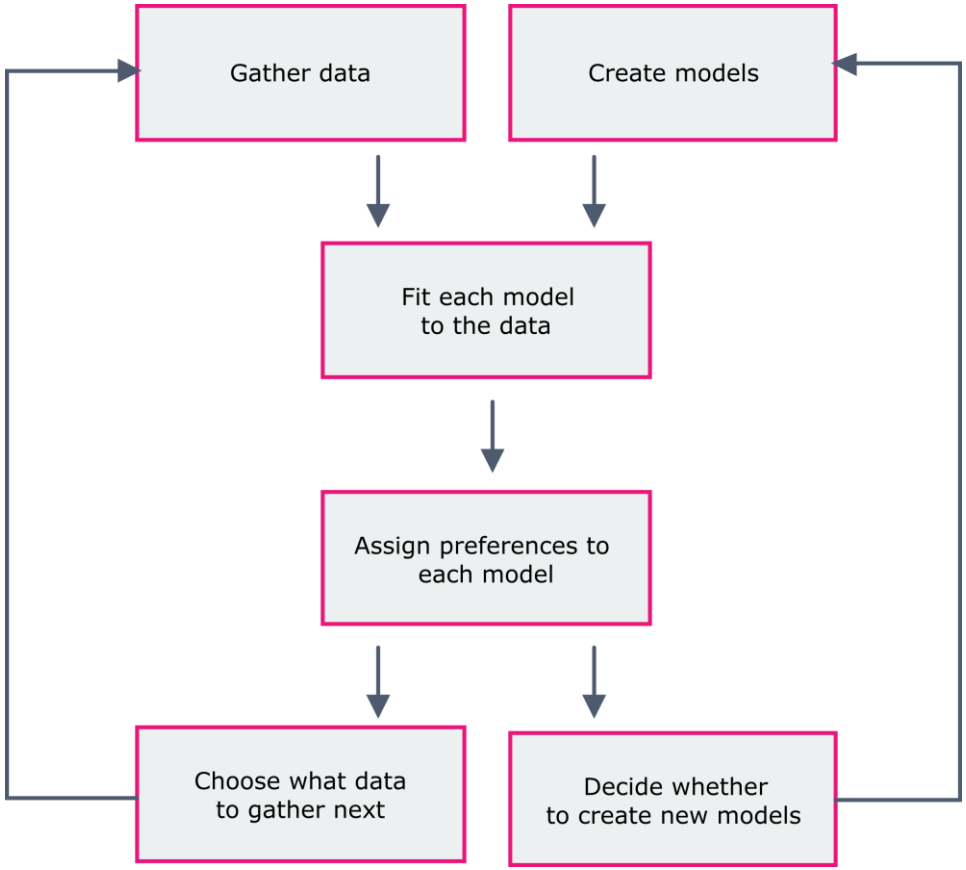


Photo credit: “Bayesian interpolation”, MacKay 1992

Some Notations

Suppose we have observed N input-target pairs as $D = \{x_n, t_n\}$, where $x_n \in \mathbb{R}^k, t_n \in \{0,1\}$, and $n = 1, 2, \cdots, N$.

Some Notations

Suppose we have observed N input-target pairs as $D = \{x_n, t_n\}$, where $x_n \in \mathbb{R}^k, t_n \in \{0,1\}$, and $n = 1, 2, \cdots, N$.

We limit our attention to a logistic regression model with parameters $w \in \mathbb{R}^k$ defined by

$$y(x_n; w) = \frac{1}{1 + \exp(-w^T x_n)} \quad .$$

Bayesian Inference

If we assume a zero-mean Gaussian prior with variance $1/\alpha$ over parameters, our posterior distribution

$$P(w \mid D, \alpha) = \frac{1}{Z} \exp(-M(w)),$$

where

$$M(w) = -\sum_n t_n \log y(x_n; w) + (1 - t_n) \log(1 - y(x_n; w)) + \frac{1}{2} \alpha w^T w.$$

Bayesian Inference

If we assume a zero-mean Gaussian prior with variance $1/\alpha$ over parameters, our posterior distribution

$$P(w \mid D, \alpha) = \frac{1}{Z} \exp(-M(w)),$$

where

$$M(w) = -\sum_n t_n \log y(x_n; w) + (1 - t_n) \log(1 - y(x_n; w)) + \frac{1}{2} \alpha w^T w.$$

$$M(w) \approx M(w_0) + \nabla M(w_0)^T (w - w_0) + \frac{1}{2} (w - w_0)^T \nabla \nabla M(w_0)(w - w_0)$$

Bayesian Inference

If we assume a zero-mean Gaussian prior with variance $1/\alpha$ over parameters, our posterior distribution

$$P(w \mid D, \alpha) = \frac{1}{Z} \exp(-M(w)),$$

where

$$M(w) = -\sum_n t_n \log y(x_n; w) + (1 - t_n) \log(1 - y(x_n; w)) + \frac{1}{2} \alpha w^T w.$$

$$M(w) \approx M(w_0) + \nabla M(w_0)^T (w - w_0) + \frac{1}{2} (w - w_0)^T \nabla\nabla M(w_0)(w - w_0)$$

normalizing constant

0

mean

inverse of covariance matrix

Entropy of a Gaussian

Entropy of a k -dimensional Gaussian distribution with covariance matrix A^{-1} is

$$S = \frac{k}{2}(1+\log 2\pi) + \frac{1}{2}\log(\det A^{-1})$$

Bayesian Active Learning

If we select change in entropy $(S_N - S_{N+1})$ as the measure for information gain, our objective is to select x_{N+1} that gives maximal expected information gain, i.e.

$$x_{N+1} = \arg\max_{x \in Q} (E_{P(t|x,D)}[S_N - S_{N+1}]).$$

Bayesian Active Learning

Therefore, the change in entropy would equal to

$$\Delta S = \frac{1}{2}\log(1+m)$$

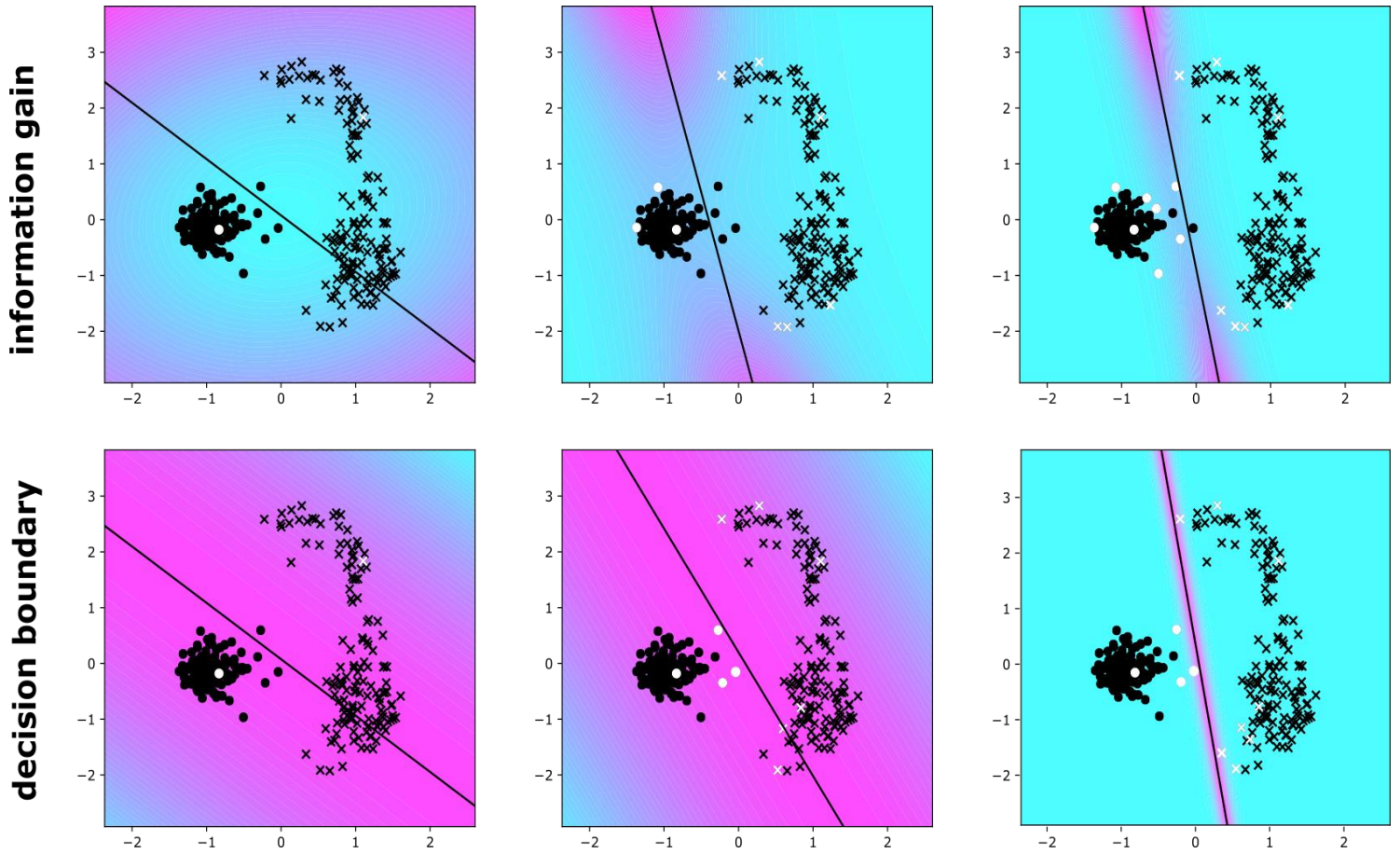
where

$$m = y(x_{N+1}; w_{MAP})[1 - y(x_{N+1}; w_{MAP})]x_{N+1}^T A_N^{-1} x_{N+1}.$$

This term does not depend on t_{N+1}, and thus $E[\Delta S] = \Delta S$.

Bayesian Active Learning

$$m = y(x_{N+1}; w_{MAP})[1 - y(x_{N+1}; w_{MAP})]x_{N+1}^T A_N^{-1} x_{N+1}.$$



Turning Inference into Prediction

To turn inference into prediction, we must take the expectation of our model output when the parameters are drawn from the posterior, i.e.

$$P(t=1|\,x,D)=\int \overbrace{P(t=1|\,x,w)}^{y(x,w)}P(w\,|\,D)dw.$$

Turning Inference into Prediction

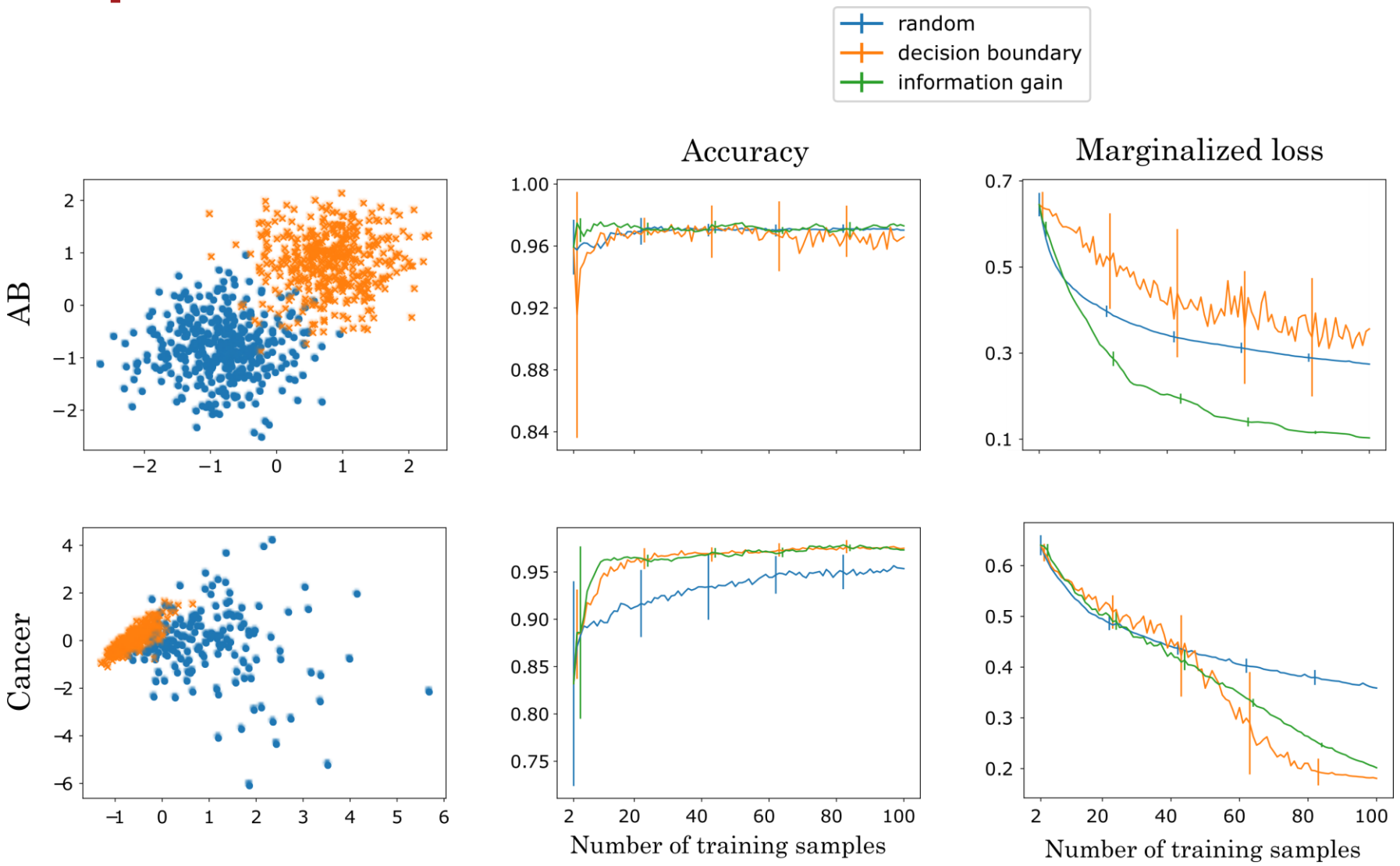
To turn inference into prediction, we must take the expectation of our model output when the parameters are drawn from the posterior, i.e.

$$P(t=1|\,x,D)=\int P(t=1|\,x,w)P(w|\,D)dw.$$

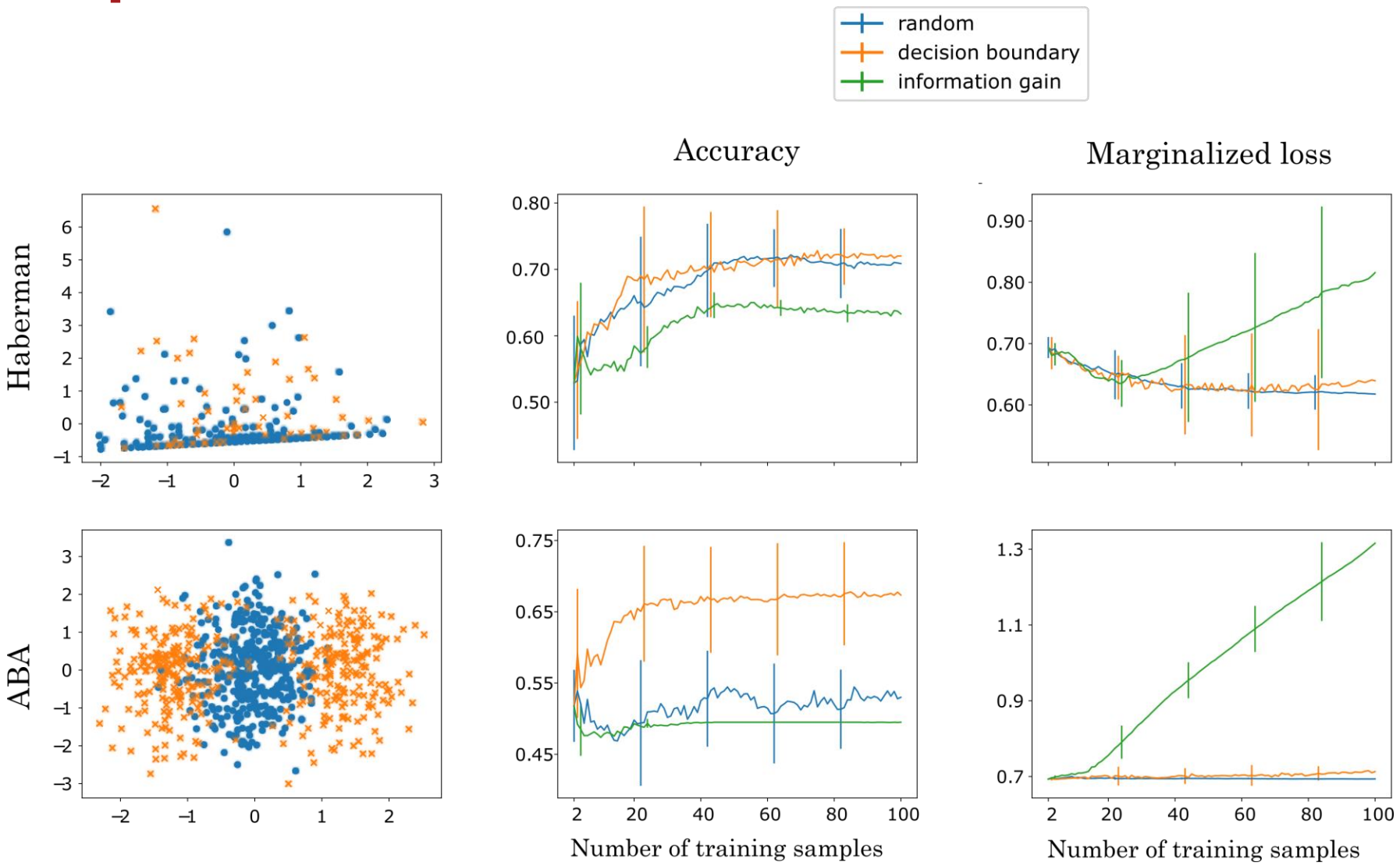
An approximation is given as

$$P(t=1|\,x,D)=\frac{1}{1+\exp(-w_{\mathrm{MAP}}^{T}x\,/\,\sqrt{1+\frac{\pi}{8}x^{T}A^{-1}x})}\,.$$

Experiments



Experiments



Discussions

- All our derivations were under the assumption that our model is well-matched to the data.

- Bayesian hypothesis testing is through model comparison:

$$P(H_i \mid D) \propto P(D \mid H_i)P(H_i)$$

Bayesian Active Learning for Maximal Information Gain on Model Parameters

Paper number: 2937

Kasra Arnavaz, Aasa Feragen, Oswin Krause, Marco Loog

Project Number:
NNF17OC0028360