BAYESIAN ACTIVE LEARNING FOR MAXIMAL INFORMATION GAIN ON MODEL PARAMETERS

Acknowledgment: This work was supported by the Novo Nordisk Foundation grant NNF17OC0028360.

Kasra Arnavaz, Aasa Feragen, Oswin Krause, Marco Loog

kasra@di.ku.dk, afhar@dtu.dk, oswin.krause@di.ku.dk, m.loog@tudelft.nl

Active learning investigates whether with fewer samples we could reach at least the same performance as random sampling if we had the control over which samples to gather.

Data Modeling Process

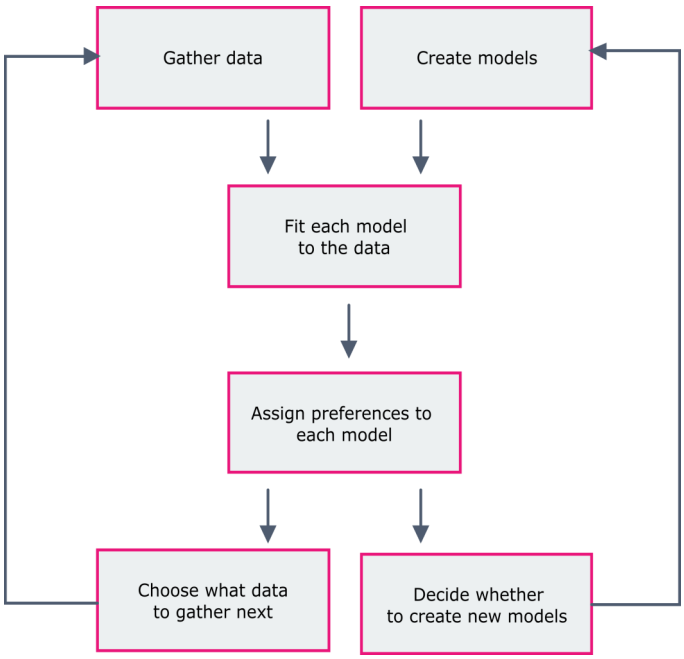


Photo credit: "Bayesian interpolation", MacKay 1992

We assume we have already decided on one particular model we want to be using.

We would then aim to gather data which gives us maximal expected information on the parameters of that particular model.

Bayesian Inference

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 $n = 1,..,N$.

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

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

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

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

where

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

We approximate our posterior distribution over parameters as a Gaussian with mean $w_{\text{MAP}} = \arg\min M(w)$ and covariance A^{-1}, where

$$A = \sum_n y(x_n; w_{\text{MAP}})[1 - y(x_n; w_{\text{MAP}})]x_n x_n^T + \alpha I_k.$$

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.

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

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

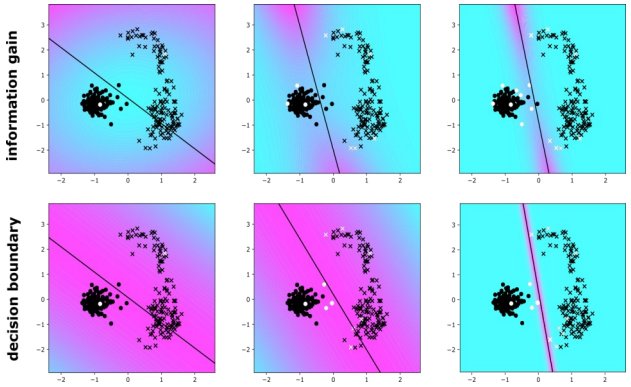
$$S = \frac{k}{2}(1 + \log 2\pi) + \frac{1}{2}\log(\det A^{-1}).$$

Therefore, the change in entropy would equal to

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

where

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



The contour colors indicate the value (pink=high, blue=low) of the property being maximized, i.e. the information gain (top) and the distance to decision boundary (bottom). The black points indicate unlabeled samples and white points indicate labeled samples. We show the progression after 2, 12, and 22 labeled samples.

Experiments

We compare the derived ‘information gain’ strategy for data gathering on roughly linearly separable data sets (top 5 rows) as well as non-linearly separable data sets (bottom 5 rows).