Active Learning in Regression Tasks

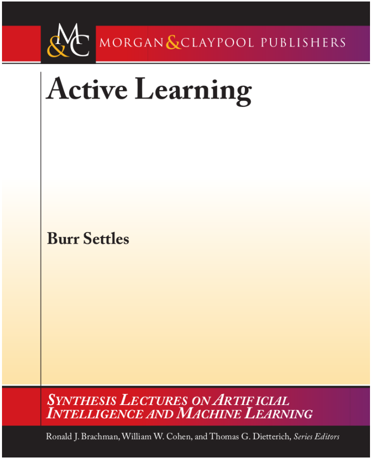
Jakub Repický

Faculty of Mathematics and Physics,
Charles University

Institute of Computer Science,
Czech Academy of Sciences

Selected Parts of Data Mining
Dec 01 2017, Prague

Bibliography



Burr Settles. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 6 (1), 1-114.

Definition

Active learning

Machine learning algorithms that aim at reducing the training effort by posing queries to an oracle.

Targets tasks, in which:

- Unlabeled data are abundant
- Obtaining unlabeled instances is cheap
- Labeling is expensive

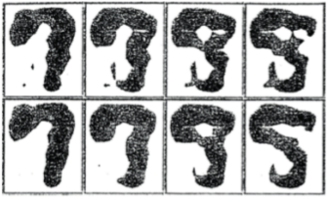
Motivation

Examples of expensive labeling tasks

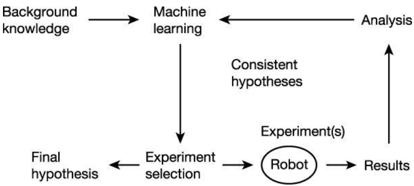
- Annotation of domain-specific data
- Extracting structured information from documents or multi-media
- Transcribing speech
- Testing scientific hypotheses
- Evaluating engineering designs by numerical simulations
- ...

Query Synthesis

- Learner may inquire about any instance from the input space
- May create uninterpretable queries
- Applicable for non-human oracles (e. g., scientific experiments)



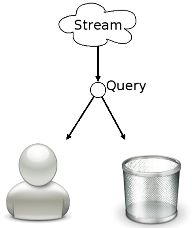
(Lang and Baum, 1992)



(King, 2004)

Selective (Stream-Based) Sampling

- Drawing (observing) instances from an input source
- The learner decides whether to discard or query the instance
- Applicable on sequential or large data



Pool-Based Sampling

- A small set $\mathcal{L}$ of labeled instances
- A large pool $\mathcal{U}$ of unlabeled instances
- Instances selected from $\mathcal{L}$ according to a utility measure evaluated on $\mathcal{U}$
- Most widely used in applications (information extraction, text classification, speech recognition, . . .)

Pool-Based Uncertainty Sampling

1. $\mathcal{L}$ – initial set of labeled instances
2. $\mathcal{U}$ – pool of unlabeled instances
3. **while true**
    1. $\theta \leftarrow$ model trained on $\mathcal{L}$
    2. $\boldsymbol{x}^* \leftarrow$ the most uncertain instance according to θ
    3. $y^* \leftarrow$ label for $\boldsymbol{x}^*$ from the oracle
    4. $\mathcal{L} \leftarrow \mathcal{L} \cup (\boldsymbol{x}^*, y^*)$
    5. $\mathcal{U} \leftarrow \mathcal{U}\ (\boldsymbol{x}^*)$

Uncertainty Measures – Least confident

$$\begin{aligned}\boldsymbol{x}^*_{\mathrm{LC}} &= \arg\min_{\boldsymbol{x}} P_\theta(\hat{y}|\boldsymbol{x}) \\ &= \arg\max_{\boldsymbol{x}} 1 - P_\theta(\hat{y}|\boldsymbol{x})\end{aligned}$$

- $\hat{y} = \arg\max_y P_\theta(y|\boldsymbol{x})$
  - minimizes the expected zero-one loss
- Only the most likely prediction is considered

Uncertainty Measures – Margin

$$\begin{aligned}\boldsymbol{x}^*_M &= \arg\min_{\boldsymbol{x}} \left(P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|\boldsymbol{x})\right) \\ &= \arg\max_{\boldsymbol{x}} \left(P_\theta(\hat{y}_2|\boldsymbol{x}) - P_\theta(\hat{y}_1|\boldsymbol{x})\right)\end{aligned}$$

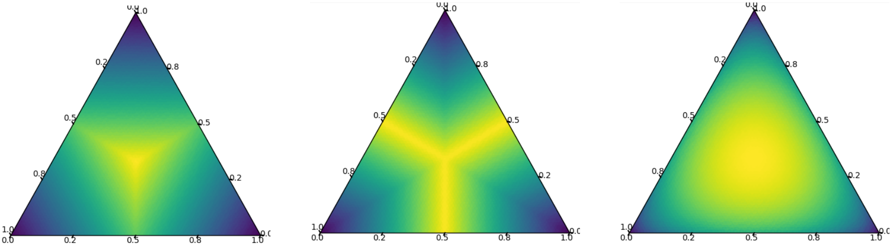
- $\hat{y}_1$ and $\hat{y}_2$ – the first and second most likely classes, respectively
- Still ignores the remainder of the predictive distribution

Uncertainty Measures – Entropy

$$\begin{aligned} \boldsymbol{x}^*_H &= \arg\max_{\boldsymbol{x}} H(Y|\boldsymbol{x}) \\ &= \arg\max_{\boldsymbol{x}} -\sum_y P_\theta(y|\boldsymbol{x}) \log P_\theta(y|\boldsymbol{x}) \end{aligned}$$

- Maximizes the expected log-loss
- Shannon entropy H – the expected self-information of a random variable

Uncertainty Measures

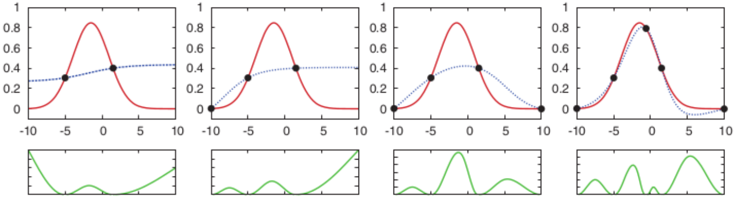


least confident | margin | entropy

Ternary distributions

Uncertainty Sampling in Regression

- Normal distribution maximizes entropy given a variance
- Variance-based uncertainty sampling equivalent to entropy-based sampling under assumption of normality
- Requires estimation of variance



(Settles, 2012)

Variance-based sampling for a 2-layer perceptron

Uncertainty Sampling Caveats

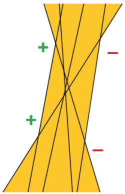
- Utility measures based on a single hypothesis
- Training set $\mathcal{L}$ is very small
- As a result, sampling bias is introduced

(a) target function (b) initial sample (c) uncertainty-based selective sampling over time

(Settles, 2012)

Version Space

- Hypothesis H – a concrete model parametrization
- Hypothesis space $\mathcal{H}$ – the set of all hypotheses allowed by the model class
- Version space $\mathcal{V} \subseteq \mathcal{H}$ – the set of all hypotheses consitent with data
- Active learning $\rightarrow$ try to reduce $\mathcal{V}$ as quickly as possible



(Settles, 2012)

Query by Disagreement

1. $\mathcal{V} \subseteq \mathcal{H}$ – the version and hypothesis spaces, resp.
2. $\mathcal{L}$ – the initial set of labeled instances
3. **repeat**
    1. receive $\boldsymbol{x} \sim \mathcal{X}$ {the stream scenario}
    2. **if** $\exists\, h_1, h_2 \in \mathcal{V},\ h_1(\boldsymbol{x}) \neq h_2(\boldsymbol{x})$ **then**
        - query label y for $\boldsymbol{x}$
        - $\mathcal{L} \leftarrow \mathcal{L} \cup (\boldsymbol{x}, y)$
        - $\mathcal{V} \leftarrow \{h \colon\ h \text{ consistent with } \mathcal{L}\}$
    3. **else**
        - discard $\boldsymbol{x}$
4. return $\mathcal{L}$

Practical Query by Disagreement

Version space $\mathcal{V}$ might be uncountable and thus unrepresentable

- Speculative hypotheses approach
  - $h_1 \leftarrow \mathrm{train}(\mathcal{L} \cup (\boldsymbol{x}, \oplus))$
  - $h_2 \leftarrow \mathrm{train}(\mathcal{L} \cup (\boldsymbol{x}, \ominus))$
- Specific-General ($\mathcal{SG}$) approach
  - A conservative h_S and a liberal h_G hypothesis
  - Approximation of region of disagreement by $\mathrm{DIS}(\mathcal{V}) \approx \{\boldsymbol{x} \in \mathcal{X} : h_S(\boldsymbol{x}) \neq h_G(\boldsymbol{x})\}$
  - Obtaining h_S and h_G: assign $\oplus$ and $\ominus$, in turn, to a sample of background points $\mathcal{B} \subseteq \mathcal{U}$

Query by Disagreement – Example

(f) disagreement-based selective sampling over time

(g) uncertainty-based selective sampling over time

(Settles, 2012)

- Previous heuristics were not aimed at predictive accuracy
- The goal: select points that minimize the *future* expected error
- Equivalent to reducing output variance (Geman et al., 1992):

$$x^*_{\mathrm{VR}} = \arg\min_{\boldsymbol{x} \in \mathcal{L}} \sum_{x' \in \mathcal{U}} \mathrm{Var}_{\theta^+}(Y|\boldsymbol{x}')$$

  - θ^+ – model after retraining on $\mathcal{L} \cup (\boldsymbol{x}, y)$
- A straightforward implementation leads to complexity explosion

Score

Given a model of random variable Y with parameters θ, the score is the gradient of the log likelihood w. r. t. θ:

$$\begin{aligned} u_\theta(\boldsymbol{x}) &= \nabla_\theta \log L(Y|\boldsymbol{x};\theta) \\ &= \frac{\partial}{\partial\theta} \log P_\theta(Y|\boldsymbol{x}) \end{aligned}$$

Fisher information is the variance of the score

$$F(\theta) = \mathrm{Var}(u_\theta(\boldsymbol{x})).$$

Under some mild assumptions, $E[u_\theta(\boldsymbol{x})] = 0$. Further, it can be shown:

$$\begin{aligned} F(\theta) &= E\left[\left(\frac{\partial}{\partial\theta}\log P_\theta(Y|\boldsymbol{x})\right)^2\right] \\ &= -E\left[\frac{\partial^2}{\partial\theta^2}\log P_\theta(Y|\boldsymbol{x})\right] \end{aligned}$$

- Expected value of negative Hessian matrix of log likelihood
- Expresses the amount of sensitivity of log likelihood w. r. t. to changes in θ

Optimal Experimental Design

Cramér–Rao bound

$F(\theta)^{-1}$ is a lower bound on the variance of any unbiased estimator $\hat{\theta}$ of parameters θ.

- "Minimize" Fisher information matrix inverse
- In general, F is a covariance matrix – what to optimize?
- Optimal Experimental Design (Fedorov, 1972) – strategies of optimizing real-valued statistics of Fisher information
- Using Fisher information, $\mathrm{Var}_{\theta^+}(Y|\boldsymbol{x})$ can be estimated without retraining at each $\boldsymbol{x}$

D-Optimal Design

$$\boldsymbol{x}_D^* = \arg\min_{\boldsymbol{x}} \det\left(\left(F_{\mathcal{L}} + u_\theta(\boldsymbol{x}) u_\theta(\boldsymbol{x})^T\right)^{-1} \right)$$

- Can be viewed as a version space reduction strategy
- Reduces the amount of uncertainty in the parameter estimates

A-Optimal Design

$$\boldsymbol{x}_A^* = \arg\min_{\boldsymbol{x}} \operatorname{tr}(AF_{\mathcal{L}}^{-1})$$

- A – a reference matrix
- Using $A_{\boldsymbol{x}} = u_\theta(\boldsymbol{x})u_\theta(\boldsymbol{x})^T$ as the reference matrix leads to a variance sampling strategy

$$\operatorname{tr}(A_{\boldsymbol{x}}F_{\mathcal{L}}^{-1}) = u_\theta(\boldsymbol{x})^T F_{\mathcal{L}}^{-1} u_\theta(\boldsymbol{x})$$

- Minimizes the average variance of the parameter estimates

Fisher information ratio

$$\begin{aligned}
\boldsymbol{x}^*_{\mathrm{FIR}} &= \arg\min_{\boldsymbol{x}} \sum_{\boldsymbol{x}' \in \mathcal{U}} \mathrm{Var}_{\theta^+}(Y|\boldsymbol{x}') \\
&= \arg\min_{\boldsymbol{x}} \sum_{\boldsymbol{x}' \in \mathcal{U}} \mathrm{tr}\Big(A_{\boldsymbol{x}'} \left(F_{\mathcal{L}} + u_\theta(\boldsymbol{x}) u_\theta(\boldsymbol{x})^T \right)^{-1} \Big) \\
&= \arg\min_{\boldsymbol{x}} \ \mathrm{tr}\Big(F_{\mathcal{U}} \left(F_{\mathcal{L}} + u_\theta(\boldsymbol{x}) u_\theta(\boldsymbol{x})^T \right)^{-1} \Big)
\end{aligned}$$

- $A_{\boldsymbol{x}'} = u_\theta(\boldsymbol{x}') u_\theta(\boldsymbol{x}')^T$
- Indirectly reduces the future output variance after labeling $\boldsymbol{x}$

Comparison of Reviewed Strategies (Settles, 2012)

Uncertainty sampling

+ simple, fast
− myopic, might be overly confident about incorrect predictions

Query by committee / disagreement

+ usable with any learning algorithm, some theoretical guarantees
− difficult to train multiple hypotheses, does not try to reduce the expected error

Error / variance reduction

+ optimizes the objection of interest, empiricaly successful
− computationally expensive, difficult to implement

Definition

Optimize $f\colon \mathcal{X} \to \mathbb{R}$ on compact $\mathcal{X} \subseteq \mathbb{R}^D$

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}),$$

under conditions

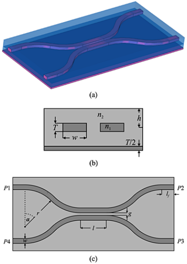
- Unknown analytical definition of f
- Unknown (analytical) derivatives, continuity, convexity properties
- f considered expensive to evaluate
- Observations of f-values possibly noisy

Motivation

Optimization of

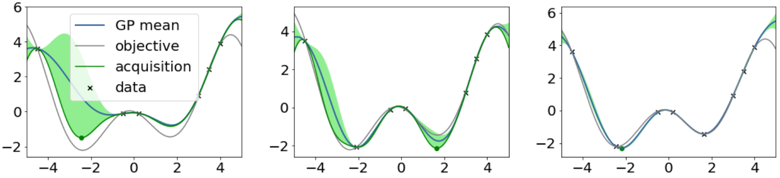
- Empirical functions: material science, chemistry,. . .
- Numerically simulated functions: engineering design optimization

Example: Photonic coupler design



(Bekasiewicz and Koziel, 2017)

1. f – the objective function
2. $\mathcal{A}$ – initial set of labeled instances
3. **repeat**
   1. $\hat{f} \leftarrow$ build the acquisition function on $\mathcal{A}$
   2. $\boldsymbol{x}^* \leftarrow \arg\min_{\boldsymbol{x}} \hat{f}$ {optimize $\hat{f}$}
   3. $y \leftarrow f(\boldsymbol{x}^*)$ {expensive evaluation}
   4. $\mathcal{A} \leftarrow \mathcal{A} \cup (\boldsymbol{x}^*, y)$



Acquisition Functions

Lower Confidence Bound:

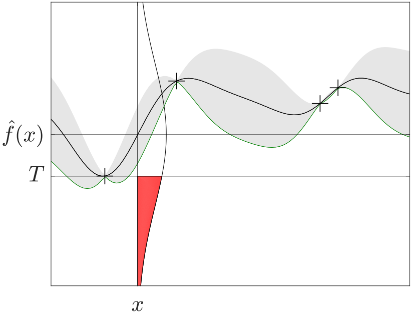
$$\mathrm{LCB}(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}) - \alpha \mathrm{Var}(Y|\boldsymbol{x})$$

Probability of Improvement

$$\mathrm{POI}(\boldsymbol{x}) = P_Y(f(\boldsymbol{x}) \leqslant T)$$

Expected Improvement

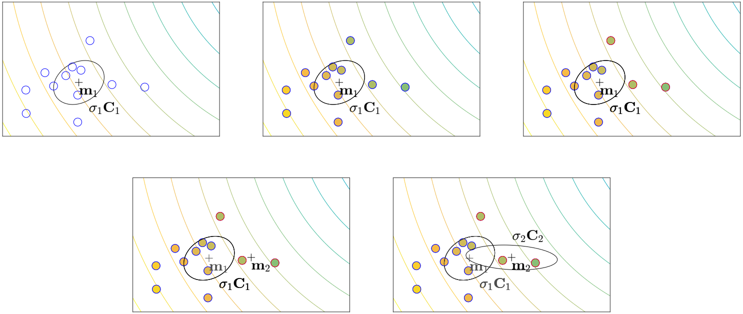
$$\mathrm{EI}(\boldsymbol{x}) = E\left(\max\left\{y^{\min} - f(\boldsymbol{x}), 0\right\}\right)$$



Evolution Strategies

- Population-based randomized search using operators of *selection*, *mutation* and *recombination*
- Covariance Matrix Adaptation Evolution Strategy – one of the most successful continuous black-box optimizer
  - Derandomized mutative parameters
  - Invariant towards rigid transformations of the input space
  - Invariant towards strictly monotonic transformations of the output space

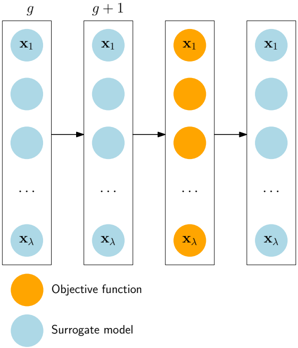
$(\mu\,,\,\lambda)$-CMA-ES (Hansen, 2001)



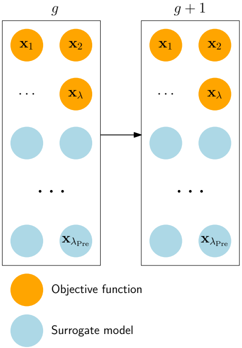
Surrogate modeling

- Stochastic optimization still requires large no. of function evaluations
- Surrogate models of the objective can be utilized as a heuristic
- Two levels of evolution control (EC) are distinguished (Jin, 2002)
  - Generation-based – a fraction of populations is wholly evaluated with the objective function
  - Individual-based – a fraction of each population is evaluated with the objective function

Evolution Control



Generation-based EC



Individual-based EC

Active Learning in Individual-Based EC

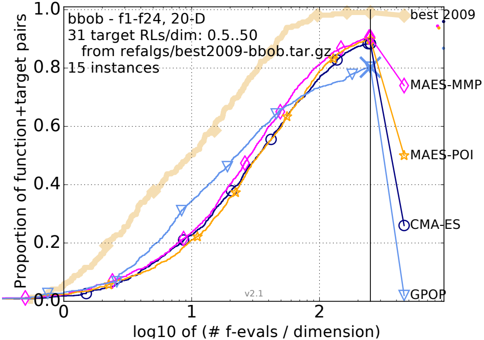
Given an extended population and a surrogate model of the objective function

- Select the most promising points
  - Combine optimality w. r. t. to the objective and utility for improving the model
- The same functions as in Bayesian optimization may be used
  - Lower confidence bound
  - Probability of improvement
  - Expected improvement

Example – Metamodel Assisted Evolution Strategy (Emmerich, 2002)

1. pop – an initial population
2. f – the objective function
3. $\mathcal{C}$ – a pre-selection criterion
4. μ – parent number
5. $\lambda, \lambda_{\mathrm{Pre}}$ – population number, extended pop. number
6. **repeat**
    1. offspring $\leftarrow$ **reproduce**(pop)
    2. offspring $\leftarrow$ **mutate**(pop)
    3. offspring $\leftarrow$ **select** λ best according to $\mathcal{C}$
    4. pop $\leftarrow$ **select** μ best according to f

Experimental comparison



Selected model-based optimizers and CMA-ES compared on the Black-Box optimization benchmarking framework

Further Reading I

Robert Burbidge, Jem J. Rowland, and Ross D. King, *Active learning for regression based on query by committee*, pp. 209–218, Springer Berlin Heidelberg, 2007.

David A. Cohn, *Neural network exploration using optimal experiment design*, Neural Networks **9** (1996), no. 6, 1071 – 1083.

Valerii Fedorov, *Theory of optimal experiments designs*, Academic press, 01 1972.

Stuart Geman, Elie Bienenstock, and René Doursat, *Neural networks and the bias/variance dilemma*, Neural Computation **4** (1992), no. 1, 1–58.

Further Reading II

David J. C. MacKay, *Information-based objective functions for active data selection*, Neural Computation **4** (1992), no. 4, 590–604.

Burr Settles, *Active learning*, Morgan & Claypool Publ., 2012.

Thank you!

repicky at cs.cas.cz