Machine learning L12 Active learning

Frank Miller, Department of Statistics

May 25, 2021

Stockholm University

Active machine learning

- Situation: Huge amount of unlabeled data available

Picture from: Lindholm et al. (2021). http://smlbook.org

- A human annotator ("oracle") can label some data points
- Should we randomly pick data points for labeling?
- Or can we do better?
- We have x-data for all data-points and could decide: Which are the most informative data points for labeling?

Stockholm University

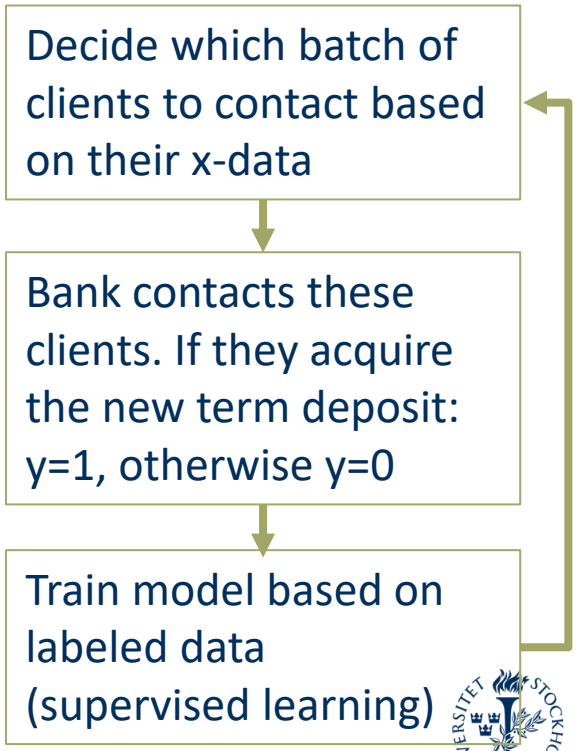
Active machine learning – Example: Optimizing marketing campaigns

- Situation (Löv, 2019, Master's thesis, SU): Bank launches marketing campaign for new term deposit
- It wants to contact its clients about their interest – contacting is costly
- They have a register of clients including their savings, marital status, employment status, types of loans, age, education, … (8 input variables/features; >45000 clients)
- Based on a model trained from data, the bank wants to predict clients with higher interest in new term deposit
- Then, they can focus on contacting the more interested clients, reducing the number of client contacts

Active machine learning – Example: Optimizing marketing campaigns

- x-data (8 input variables) of >45000 clients available, y-data (if client acquires new term deposit) not available

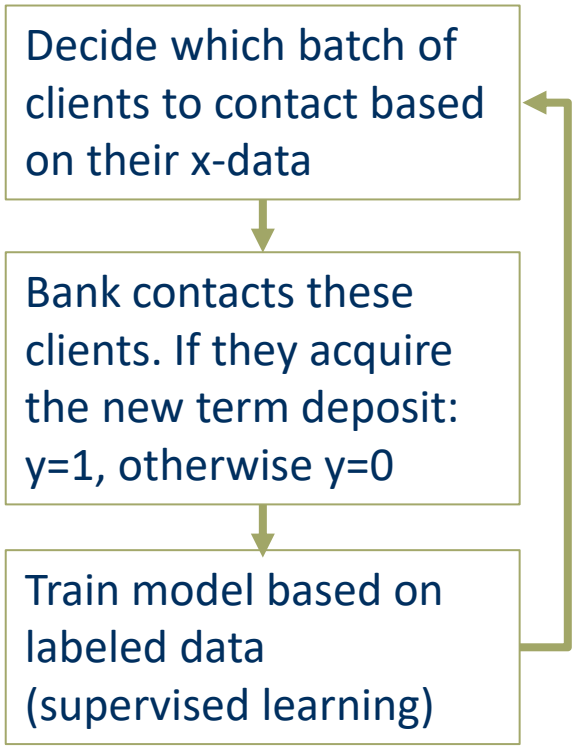
age	job	marital	education	balance	housi	loan	poutcome	y
56	employed	married	tertiary	1476	no	no	no	
32	unemployed	married	tertiary	744	yes	no	no	
39	employed	married	primary	592	yes	no	no	
54	employed	divorced	secondary	447	yes	no	no	
55	employed	married	primary	1691	yes	no	failure	
33	employed	divorced	tertiary	893	yes	no	failure	
42	unemployed	married	tertiary	576	no	no	no	
59	employed	married	primary	-187	yes	no	no	
→ 52	employed	married	secondary	3332	yes	yes	no	0
54	employed	married	secondary	171	no	yes	no	
30	employed	married	secondary	486	yes	no	no	
→ 58	unemployed	divorced	secondary	5920	yes	no	no	0
25	employed	single	primary	-65	yes	no	no	
52	employed	divorced	secondary	123	no	no	no	
→ 31	employed	single	tertiary	2589	no	no	no	1
32	employed	single	secondary	706	yes	no	no	
33	employed	married	secondary	3132	yes	no	no	
47	unemployed	divorced	secondary	-157	no	yes	no	
29	employed	single	secondary	612	no	no	no	
38	employed	married	tertiary	276	no	no	no	



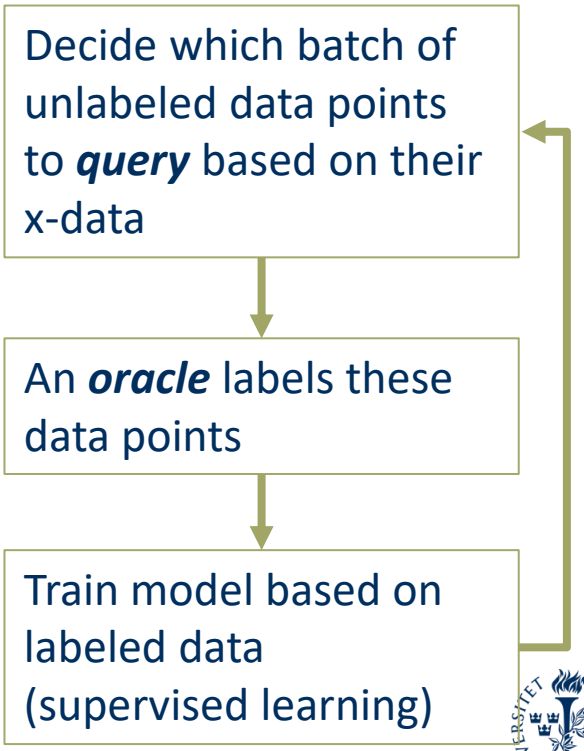
Stockholm University

Active machine learning – the process

- Marketing example



General active learning



Stockholm University

Active machine learning – Query strategies

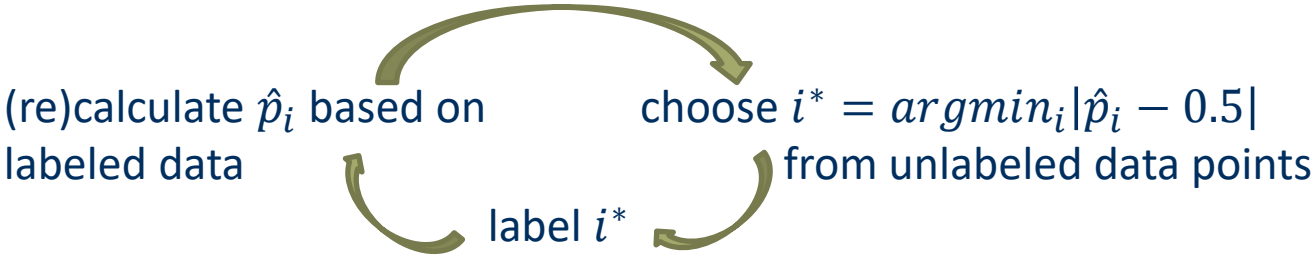
- Which (batch of) data points should be queried?
- Different strategies:

- Uncertainty sampling
  - Least confident, margin, and entropy method
- Variance reduction
  - Optimality criteria from Optimal Experimental Design (D-, A-, and E-optimality)

- For simplicity, we assume that we can query one data point at the time (batch size = 1)

Active machine learning – Query strategy: Uncertainty sampling

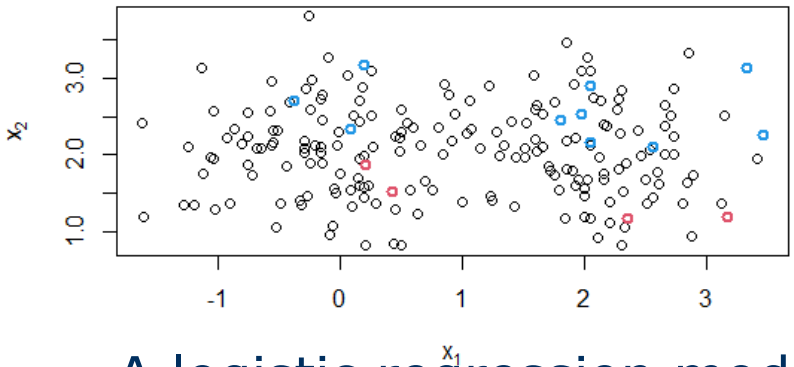
- Label sequentially the "most informative" unlabeled data points $\boldsymbol{x}_i$
- One method: choose data point $\boldsymbol{x}_i$ with the most uncertain predicted class
- In a binary classification problem ($y_i \in \{0,1\}$), let $\hat{p}_i$ be the current predicted probability for $y_i = 1$; logistic regression: $\hat{p}_i = 1/\{1 + \exp(-\hat{\boldsymbol{\theta}}^T \boldsymbol{x}_i)\}$
- Then, the data point with $\hat{p}_i$ closest to 0.5 is most uncertain: choose $i^* = argmin_i|\hat{p}_i - 0.5|$



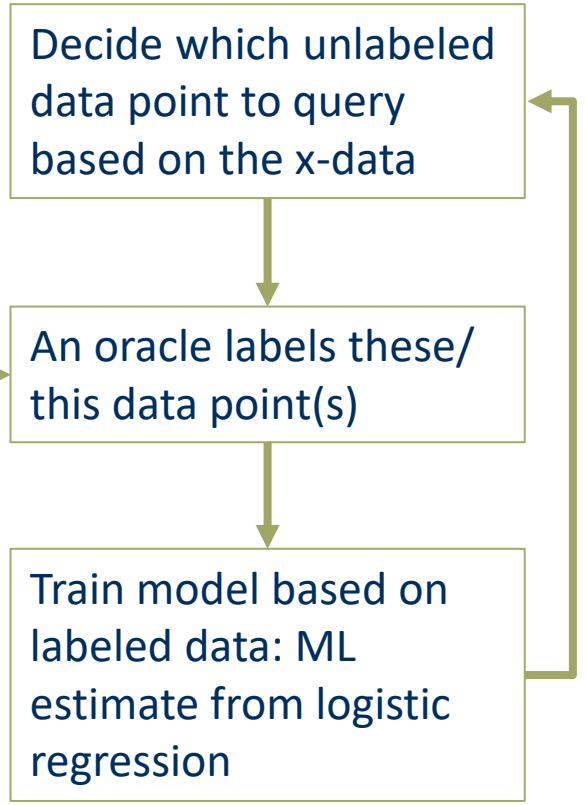
Stockholm University

Active machine learning – Illustrating example

- To illustrate, we consider a dataset of n=100 unlabeled data points with 2 input variables/features x_1 and x_2
- 15 data points randomly selected and labeled (class 0/1)

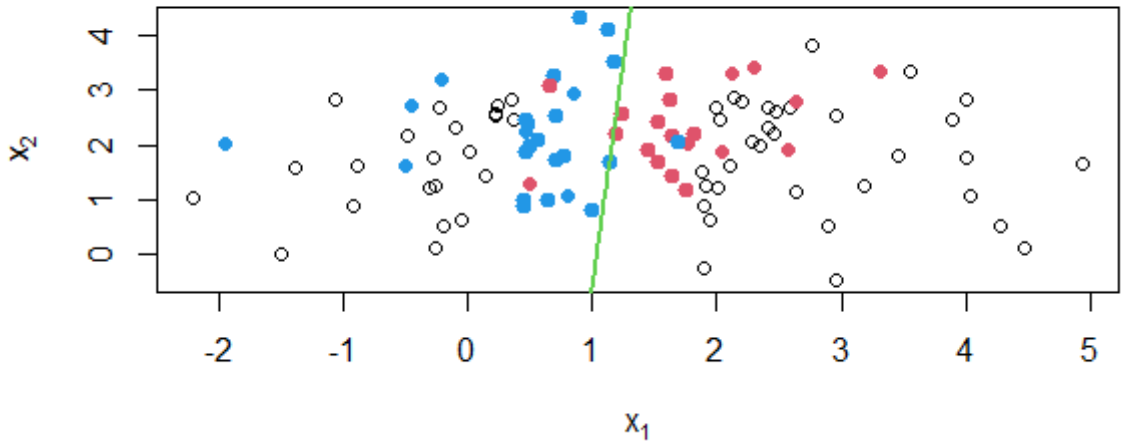


- A logistic regression model is used for classification



Stockholm University

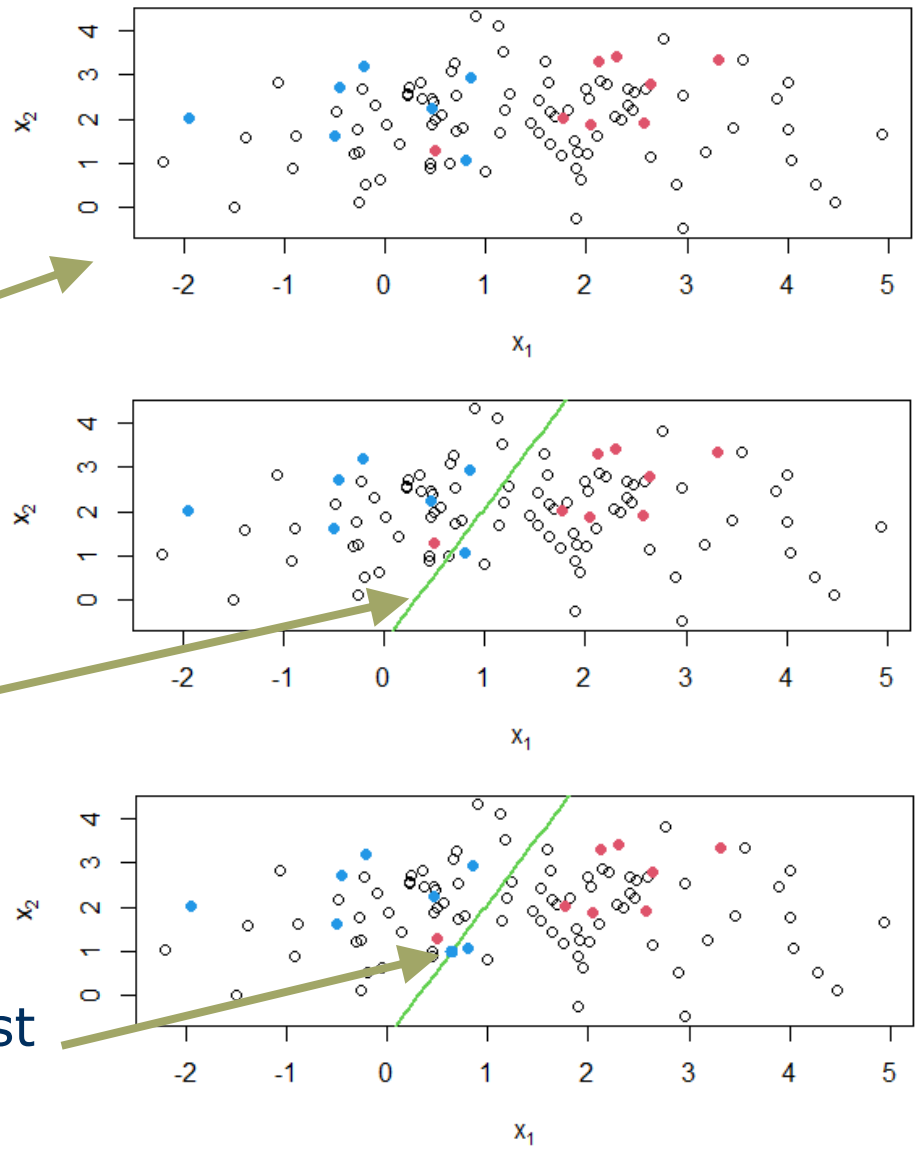
Active machine learning example



Stockholm University

Active machine learning example

- Pick randomly some observations, label them
- Analyze with logistic regression model (supervised learning, labeled data only) and determine decision boundary (where both classes are equally likely)
- Choose observation with largest uncertainty (closest to the boundary), label it

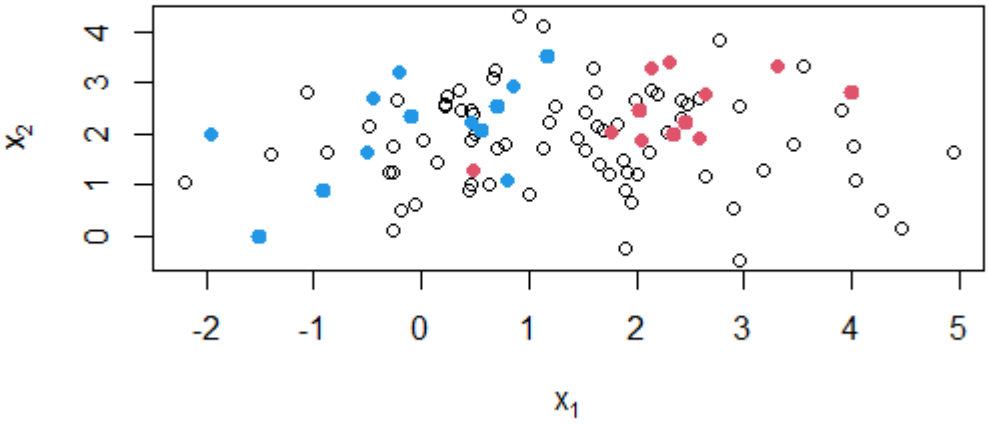


Stockholm University

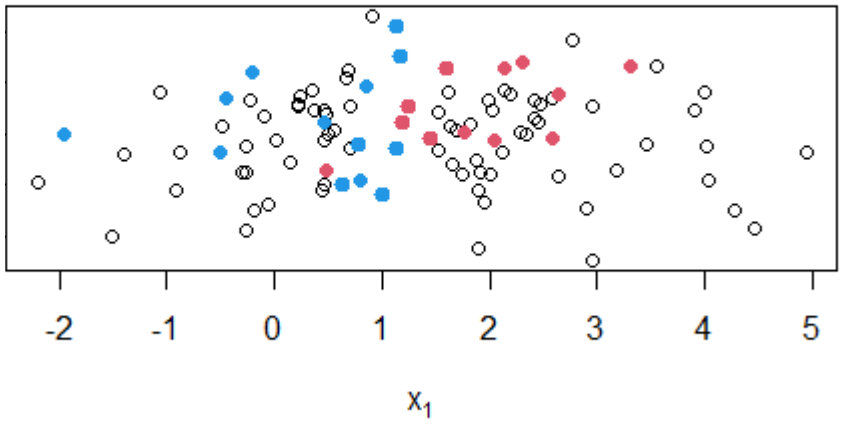
Active machine learning – Query strategy: Variance reduction

- 15 random data points; 10 with active/passive learning

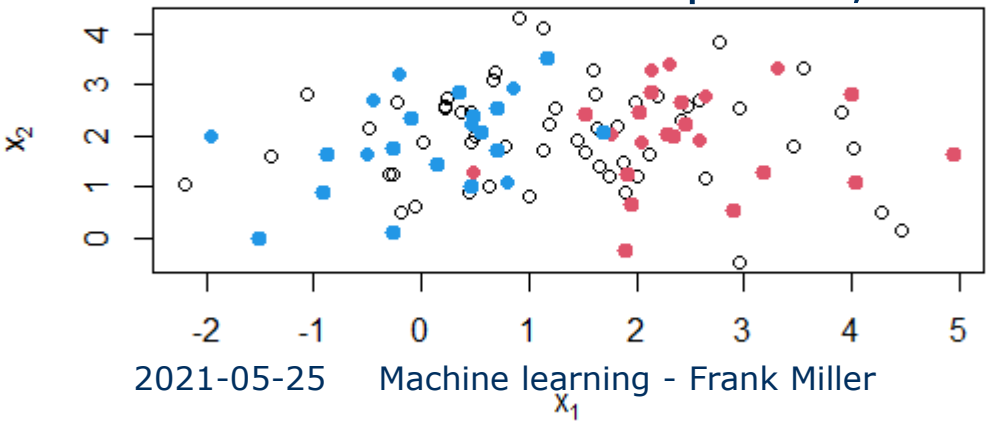
Passive learning

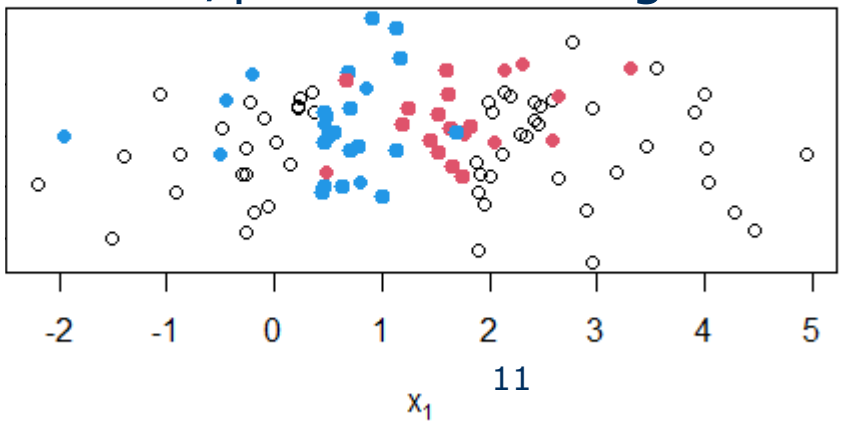


Active learning (uncertainty sampling)



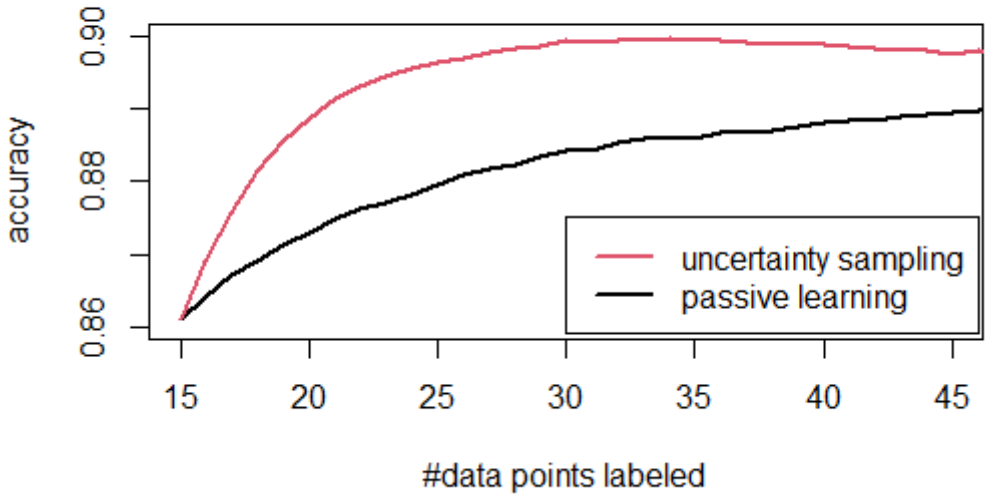
- 15 random data points; 30 with active/passive learning





Active and passive learning

- Accuracy (percentage of correctly classified labels) for active learning with uncertainty sampling and passive learning with random choice



The results here are based on the average of 1000 simulations

- Accuracy increased faster with active learning compared to passive learning
- Other quality measures than accuracy possible

Stockholm University

Uncertainty sampling for more than 2 classes

- Let $\hat{y} = argmax_y p(y|x,\theta)$ be the class with the highest posterior probability for data point *x* under model θ and $\hat{y}_S = argmax_{y \neq \hat{y}} p(y|x,\theta)$ be the second-best class
- ***Least confident*** method
  Choose data point *i** with
  $x_{i^*} = argmax_i\{1 - p(\hat{y}|x_i,\theta)\}$

Probabilities for class m

	$\hat{y}_m$	m=1	m=2	m=3
Data point	i=1	0.4	0.35	0.25
	i=2	0.45	0.45	0.1

- ***Margin*** method
  Choose *i** with $x_{i^*} = argmin_i\{p(\hat{y}|x_i,\theta) - p(\hat{y}_S|x_i,\theta)\}$
- ***Entropy*** method
  Choose *i** with $x_{i^*} = argmax_i\{-\sum_m p(y_m|x_i,\theta)\log p(y_m|x_i,\theta)\}$

- Compare Figure 5 in Settles (2010)
- Note: all three methods coincide for K=2 classes

Variance of LS estimate in linear regression and optimal design

- Linear model: $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{\epsilon}$
- The LS estimate for $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ (see ch. 3.A)
- The covariance matrix of the LS estimator is: $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

  $(Cov(\widehat{\boldsymbol{\theta}}) = E(\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^T) = E\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\right)^T\right) =$

  $E\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\,\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{y}\,\boldsymbol{y}^T)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} =$

  $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1})$

- If we have the possibility to choose the x-variables, we can minimize $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
- x-variables can be chosen in the planning of experiments and in active learning

Variance of LS estimate in linear regression and optimal design

- We want to minimize $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
- If we have more than one parameter, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is a matrix and we have to decide in which way to minimize
- Optimality criteria:
  - Minimize Average (or sum) of variances of all $\hat{\theta}_i$; these are in the diagonal of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$:
    $$\text{Minimize trace}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$
  - Minimize Determinant of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ which corresponds to the volume of a confidence ellipsoid for $\boldsymbol{\theta}$:
    $$\text{Minimize } \det(\boldsymbol{X}^T\boldsymbol{X})^{-1} = 1/\det(\boldsymbol{X}^T\boldsymbol{X})$$
  - Minimize the largest Eigenvalue of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$:
    $$\text{Minimize } \lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = 1/\lambda_{\min}(\boldsymbol{X}^T\boldsymbol{X})$$
- These are called A-, D-, and E-optimality, respectively

Variance of ML estimate in logistic regression and optimal design

- Logistic regression model:

$$E(y_i) = g(\boldsymbol{x}_i; \boldsymbol{\theta}) = 1/\{1 + \exp(-\boldsymbol{\theta}^T \boldsymbol{x}_i)\}$$

- The covariance of the ML estimator $\widehat{\boldsymbol{\theta}}$ is $(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}$ with a diagonal matrix

$$\boldsymbol{W} = \mathrm{diag}\big(v(\boldsymbol{x}_1; \boldsymbol{\theta}), v(\boldsymbol{x}_2; \boldsymbol{\theta}), \ldots, v(\boldsymbol{x}_n; \boldsymbol{\theta})\big)$$

$$v(\boldsymbol{x}_i; \boldsymbol{\theta}) = g(\boldsymbol{x}_i; \boldsymbol{\theta}) * (1 - g(\boldsymbol{x}_i; \boldsymbol{\theta}))$$

- Consequently, we want to minimize $\mathrm{trace}(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}$, $1/\det(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})$, or $1/\lambda_{\min}(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})$ for A-, D-, or E-optimality

Active machine learning – Query strategy: Variance reduction

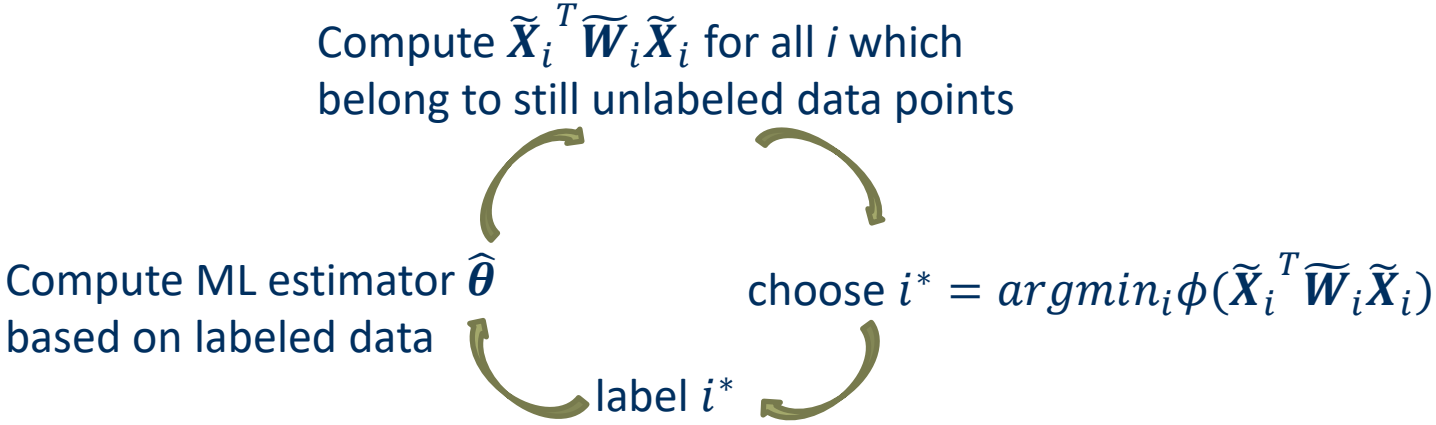
- In active learning, we have labeled data with design matrix $\boldsymbol{X_L}$ and variance matrix $\boldsymbol{W_L}$ for the labeled data points
- The variance reduction method compares the possible matrices

$$\widetilde{\boldsymbol{X}}_i = \begin{pmatrix} \boldsymbol{X_L} \\ \boldsymbol{x}_i^T \end{pmatrix}$$

  with *i* from unlabeled data points
- Search *i* among unlabeled data points with e.g. $1/\det\left(\widetilde{\boldsymbol{X}}_i{}^T \widetilde{\boldsymbol{W}}_i \widetilde{\boldsymbol{X}}_i\right)$ minimal (in case of D-optimality); $\widetilde{\boldsymbol{W}}_i$ is here the diagonal matrix with same diagonal elements as $\boldsymbol{W}$ and an additional element $v\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}\right)$ for the new data point

Active machine learning – Query strategy: Variance reduction

- The matrix $\boldsymbol{W}$ and $\widetilde{\boldsymbol{W}}_i$ depend on the unknown parameter vector $\boldsymbol{\theta}$, but we can compute them based on the current estimate $\hat{\boldsymbol{\theta}}$ I



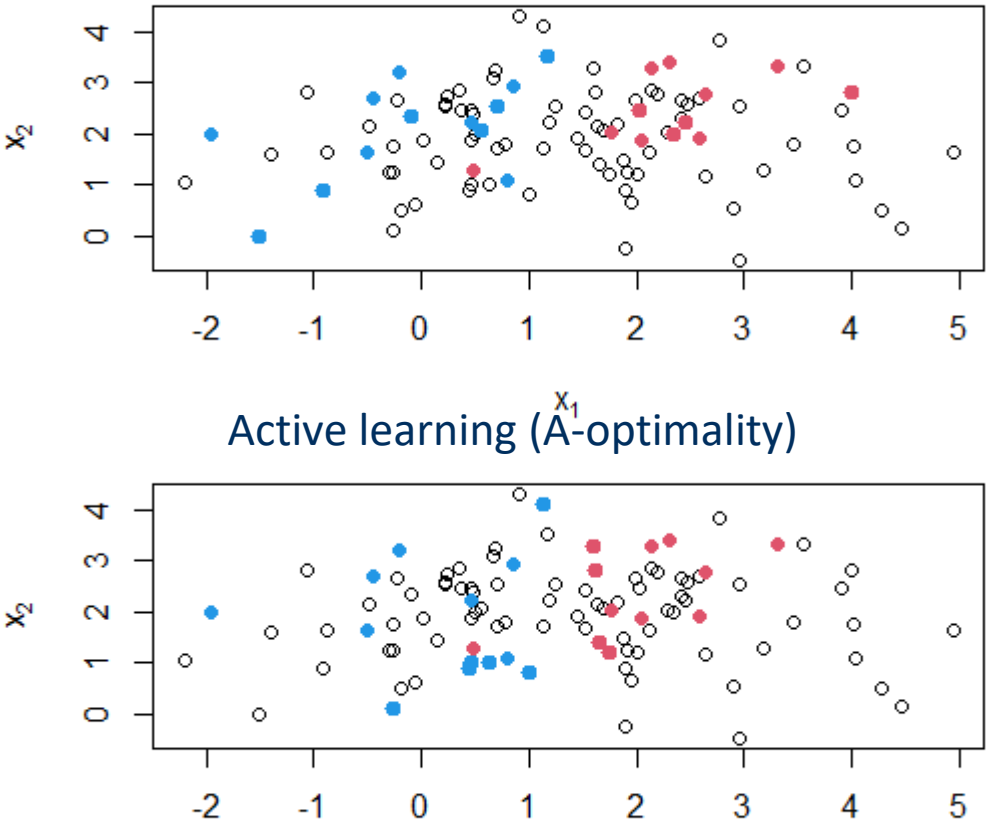
where $\phi(M) = \text{trace}(M^{-1})$ or $\frac{1}{\det(M)}$ or $= 1/\lambda_{\min}(M)$

Stockholm University

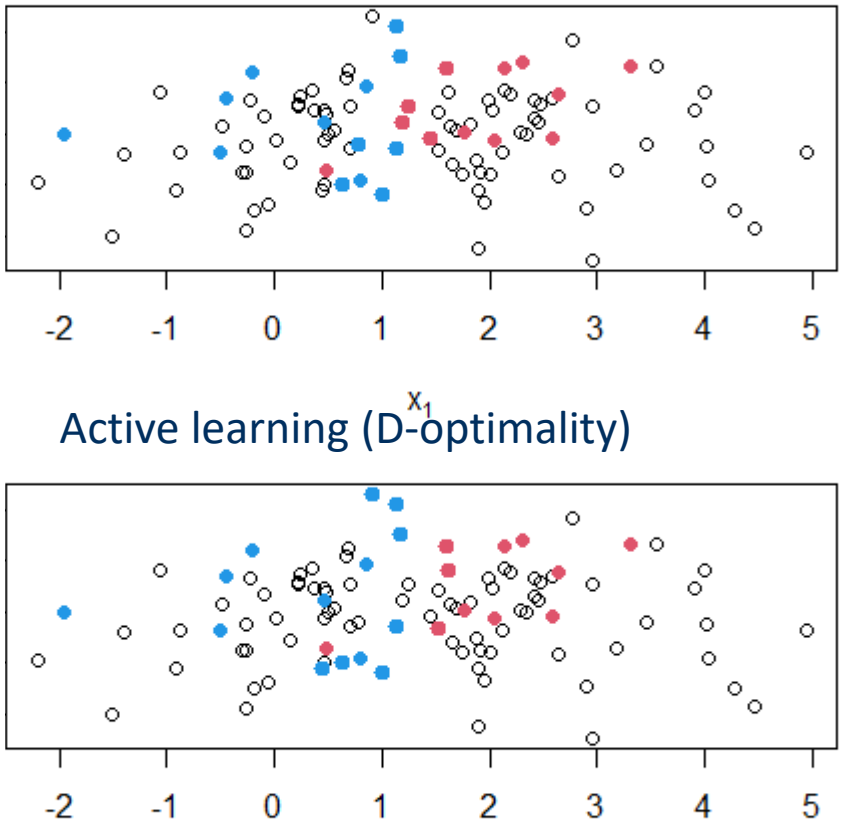
Active machine learning – Query strategy: Variance reduction

- 15 random data points; 10 with active/passive learning

Passive learning



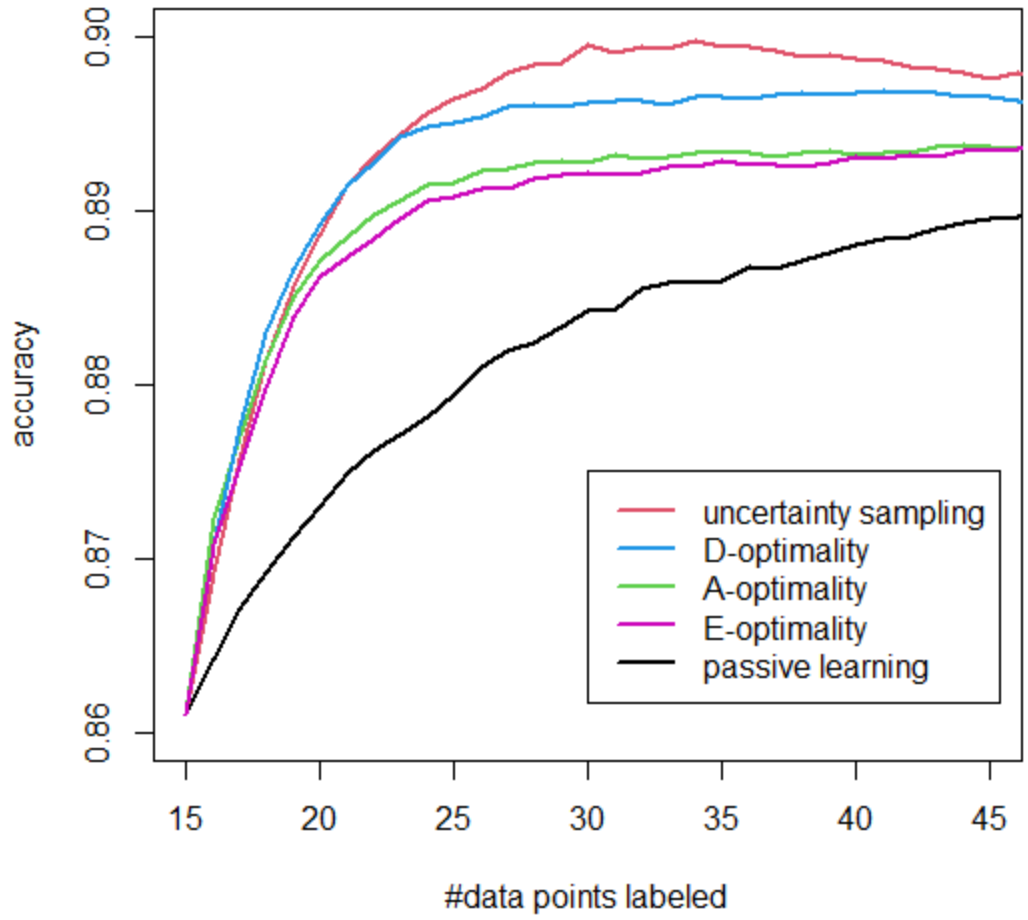
Active learning (uncertainty sampling)



Active learning (A-optimality)

Active learning (D-optimality)

Comparison of different active strategies

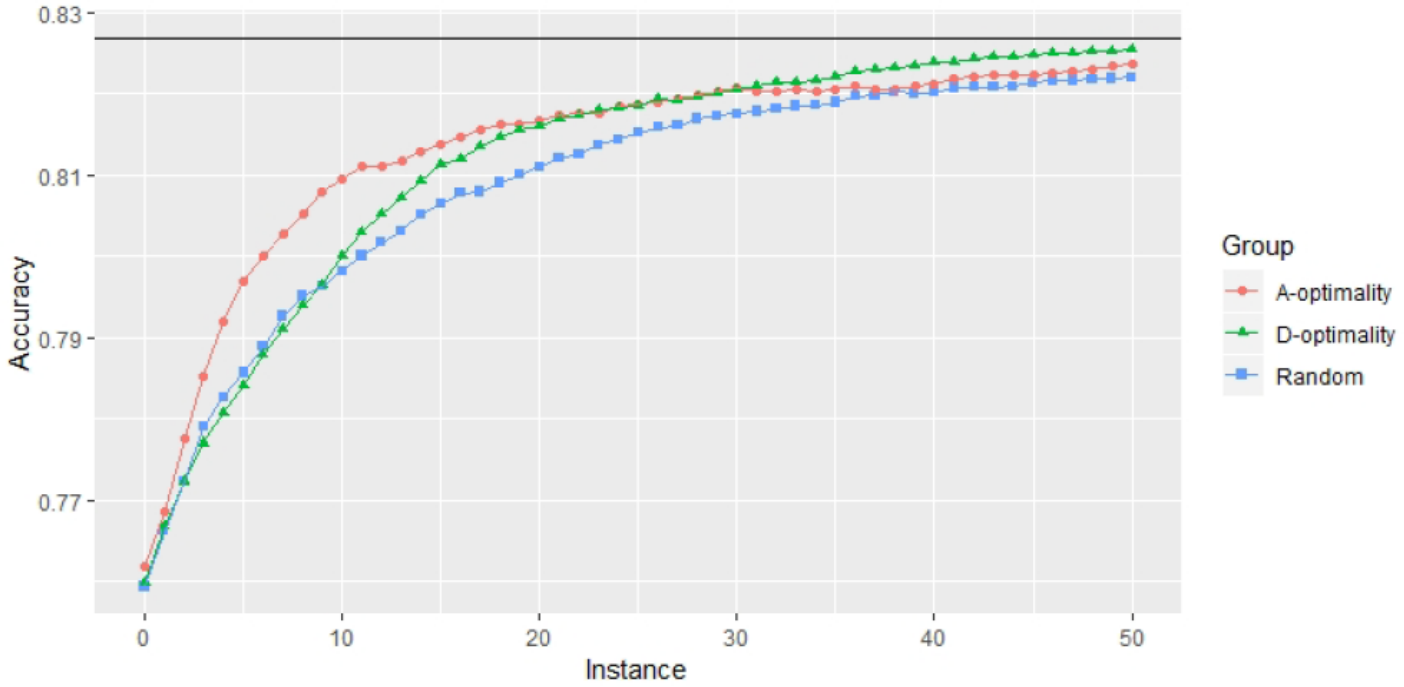


The results here are based on the average of 1000 simulations

Stockholm University

Example: Optimizing marketing campaigns

- A- and D-optimal active learning better than passive learning ("random")



From: Löv T (2019). *Optimizing marketing campaigns with active machine learning.* Master's thesis. Stockholm University.

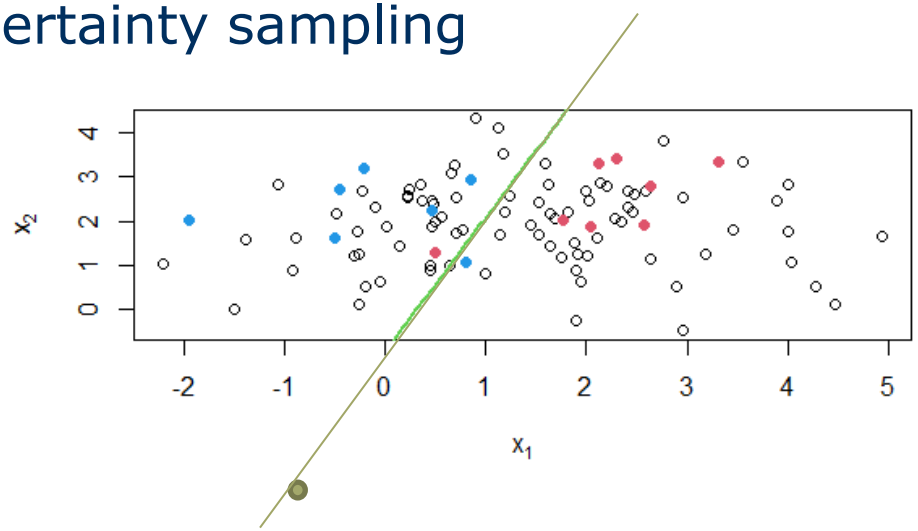
Stockholm University

Comparison of different active strategies

- Uncertainty sampling is computationally less complex
- For variance reduction method, computation of inverse, determinant, or eigenvalues required n_U times in each iteration (n_U = number of unlabeled data points)
- If model has many parameters, this can become intractable
- Take care to program the code for selection most efficient, e.g. compute `1/det(M)` instead of `det(solve(M))`
- Variance reduction method can use a taylor-made optimality criterion

Comparison of different active strategies

- Outliers (in x-distribution) might have impact especially on uncertainty sampling



- Variance reduction methods can better cope with this

- Density-weighted methods can deal with this issue; they downweigh outliers

Other active learning scenarios

- We have discussed so-called pool-based sampling;
  a static pool of unlabeled data available
- Other scenarios:
  - Stream-based selective sampling
    Unlabeled data arrives one at a time; model decides then whether to query the data point (label it) or to discard and wait for the next
  - Membership query synthesis
    Model creates x-data *de novo* for a query
    Example:
    - x-data are angles of a robot's arm and y-data are its coordinates
    - Can be challenging/impossible for recognition of handwritten characters