2.11.2018

Active Learning

Data Mining Guestlecture

Georg Krempl

Algorithmic Data Analysis
Information & Computing Sciences Department
Utrecht University, The Netherlands

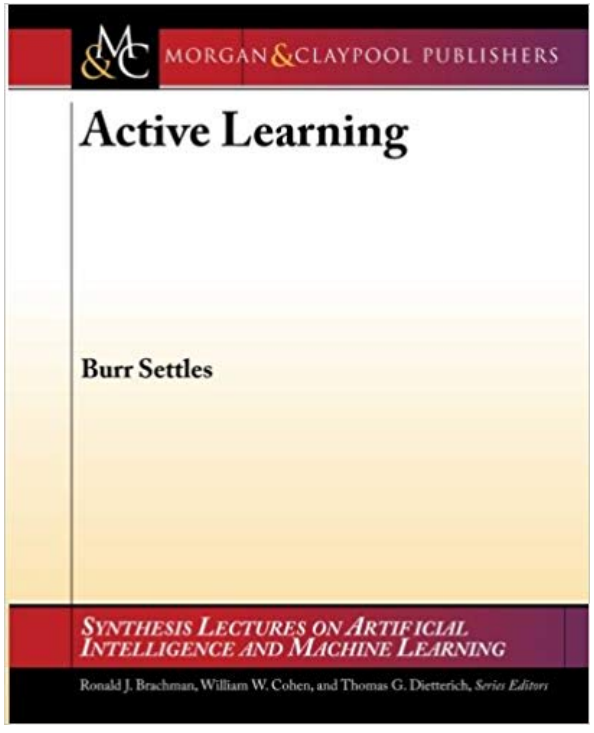
Lecture Outline

- Lecture's Context
- Semi-Supervised Learning
- Active Learning
  - Motivation
  - Approaches
    - Uncertainty Sampling
    - Expected Error Reduction
    - Ensemble / Query by Committee
    - Probabilistic Active Learning
  - Evaluation
- Summary and Questions

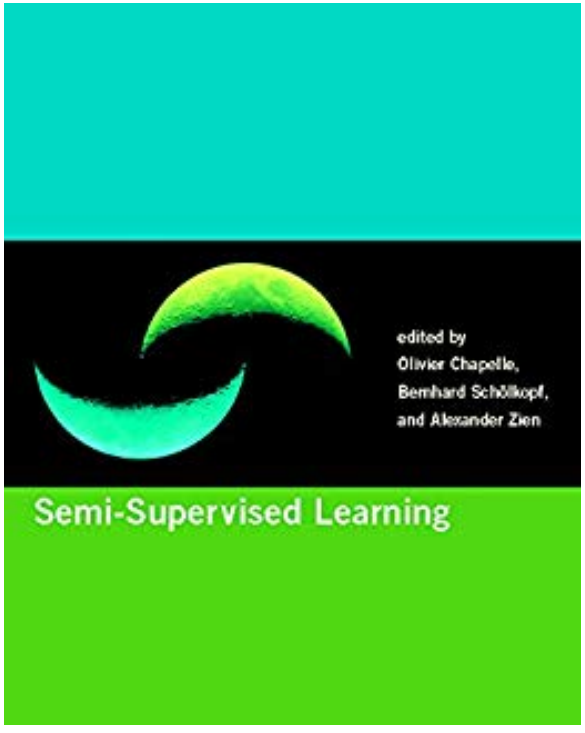
Lecture Outline

- Lecture's Context
- Semi-Supervised Learning
- Active Learning
  - Motivation
  - Approaches
    - Uncertainty Sampling
    - Expected Error Reduction
    - Ensemble / Query by Committee
    - Probabilistic Active Learning
  - Evaluation
- Summary and Questions

Recommended Literature on Active and Semi-Supervised Learning



(a) [Settles, 2012]



(b) [Chapelle et al., 2006]

Context of Active Learning within Data Mining / Machine Learning

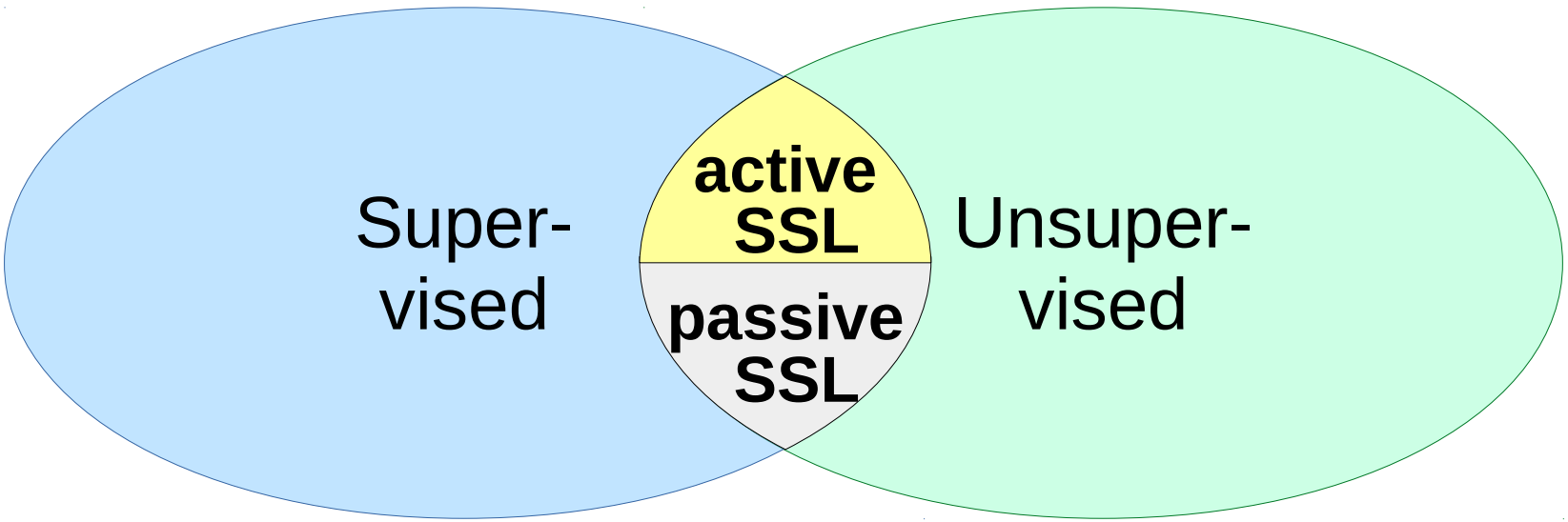
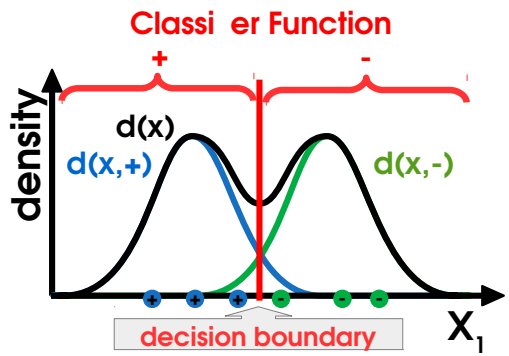


Figure: Context of Active Learning within Machine Learning

Supervised (Passive) Classification: Recapitulation

Credit Scoring: will a new order be paid?

amount X_1	...	default?
2.2	...	no
1.7	...	yes
1.3	...	yes
3.1	...	no
...	...	...



- **Historical Data**
  e.g. previous customer's records
- **Generate Training Sample** $\mathcal{L}$ with feature variables (e.g. order value X) and class label (e.g. default Y)
- **Aim: Classifier function** $f : X \to Y$
  E.g. using Bayes theorem

$$Pr(Y|X) = \frac{Pr(X|Y) \cdot Pr(Y)}{Pr(X)}$$

- **Estimate Distributions**
  joint distributions $d(X, Y)$ or posterior distributions
  $d(Y|X) = \frac{d(X,Y)}{d(X)}$
- **Derive Decision Boundary**
  at intersections of posterior distributions

Supervised Classification: Limitations

Common assumptions about the training data:

Data (features, labels) on each instance is

- *unbiased*
- *correct*
- *complete*
- and *available at once*
- available at *no cost* and *without the classifier controlling label selection*

What if these assumptions do not hold?

- Challenge for conventional supervised learning approaches[1]
  Requires techniques such as adaptive, active, semi-supervised, transfer learning

[1] See [Krempl et al., 2014b] in SIGKDD Explorations.

Lecture Outline

- Lecture's Context
- Semi-Supervised Learning
- Active Learning
  - Motivation
  - Approaches
    - Uncertainty Sampling
    - Expected Error Reduction
    - Ensemble / Query by Committee
    - Probabilistic Active Learning
  - Evaluation
- Summary and Questions

Label Paucity:

- *Completeness* of information is not met:
  Only some labels are available
- **Semi-Supervised Classification**

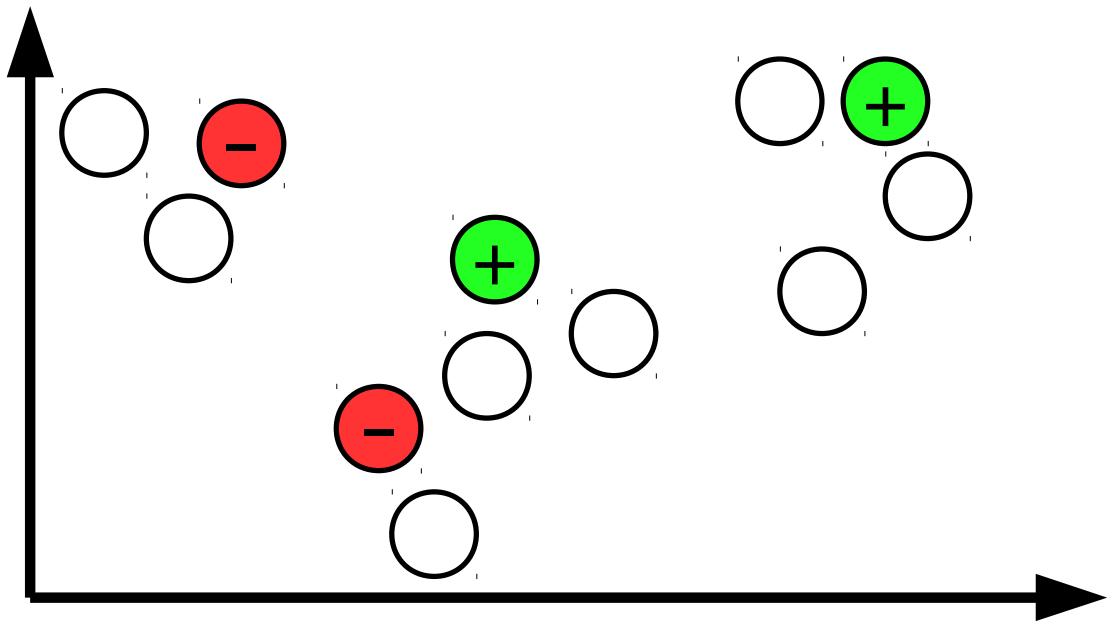


Figure: Classification with Partially Labelled Data

Semi-Supervised Learning[2]

Basic situation

- Not all data are labelled

Objective

- Use unlabelled data to improve classification model

Assumptions

Basic Assumption: There is some structure in the data

- Smoothness assumption
  - Neighbouring instances are similar w.r.t. class label, i.e.the closer, the more likely is a similar class label
  - We need to find low density regions
  - Decision boundary is placed in low-density regions
- Cluster assumption
  - Instances form clusters (subpopulations)
  - Within one cluster, similar label is more likely
  - In extreme case, label Y is conditionally independent of X given cluster Z
- Manifold assumption
  - Data are on a manifold of much lower dimension than the feature space
  - Learning the manifold eases classification (remedy against curse of dimensionality)

▸ Details of SSL Approaches

[2] See for example [Chapelle et al., 2006, Zhu, 2008].

Semi-Supervised Learning[3]

Basic situation

- Not all data are labelled

Objective

- Use unlabelled data to improve classification model

Assumptions

Basic Assumption: There is some structure in the data

- Smoothness assumption
- Cluster assumption
- Manifold assumption

Methods

- Self-Learning, Co-Learning
- Generative Models
- Low-density Separation
- Graph-Based Methods

▸ Details of SSL Approaches

[3] Literature survey by Zhu 2005 [Zhu, 2008].

SSL - Limitations

- Ex-post, passive viewpoint (we work with the data we have)
- What if we can influence which labels we obtain?

Lecture Outline

- Lecture's Context
- Semi-Supervised Learning
- Active Learning
  - Motivation
  - Approaches
    - Uncertainty Sampling
    - Expected Error Reduction
    - Ensemble / Query by Committee
    - Probabilistic Active Learning
  - Evaluation
- Summary and Questions

Active Learning: Motivation[4]

Motivation

- some labels are expensive to obtain

Active Learning in Exemplary Applications

- Credit Scoring:
  E.g. costly to accept high risk clients for model building
- Brain Computer Interfaces:
  E.g. performing tasks for calibration can be tedious for user

Selection is important

- Even (or in particular!) in "Big Data" applications
- Efficient allocation of limited resources
- Sample where we **expect something interesting**

[4] See, e.g., [Settles, 2012, Fu et al., 2012, Zliobaitė et al., 2013].

Active Learning: Context & History

Context & Aim

- unlabelled data $\mathcal{U}$ is abundant,
- labelling is costly (small set of labelled data $\mathcal{L}$)
- control over the label selection process
- select the most valuable (informative) instances for labelling

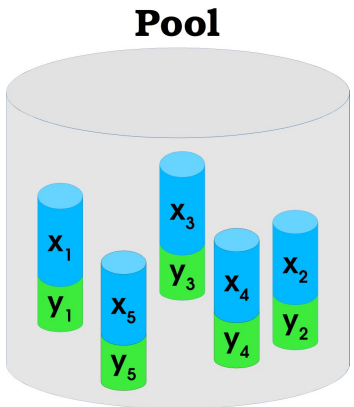
History

- Optimal experimental design [Fedorov, 1972]
- Learning with queries/query synthesis [Angluin, 1988]
- Selective sampling [Cohn et al., 1990]

Selective Data Acquisition Tasks[5]

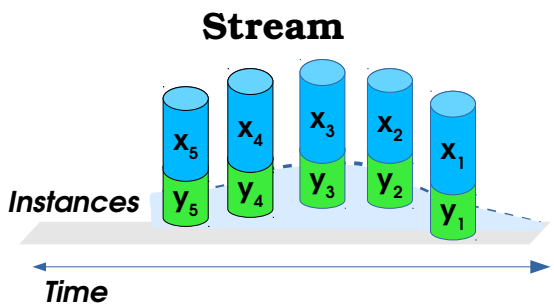
Active Learning Scenarios

- **Query synthesis**: example generated upon query
- **Pool** $\mathcal{U}$ of unlabelled data: static, repeated access
- **Stream**: sequential arrival, no repeated access



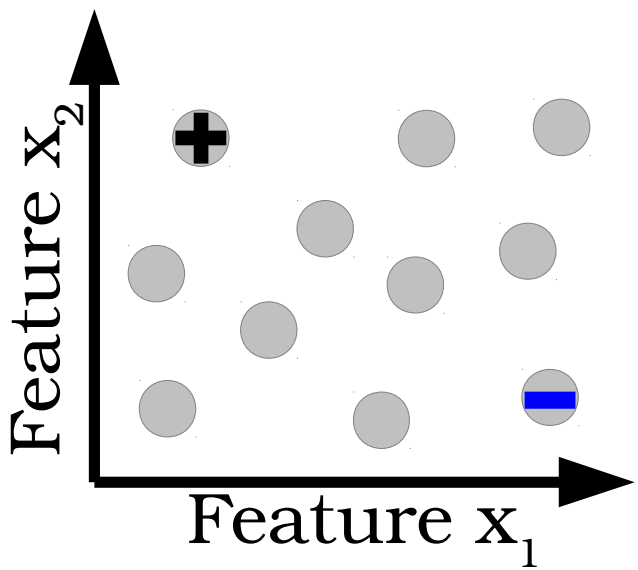
Type of Selected Information

- Active label acquisition
- Active feature (value) acquisition
- Active class selection, also denoted
  Active class-conditional example acquisition
- ...



[5] Own categorization, inspired by [Attenberg et al., 2011, Saar-Tsechansky et al., 2009, Settles, 2009].

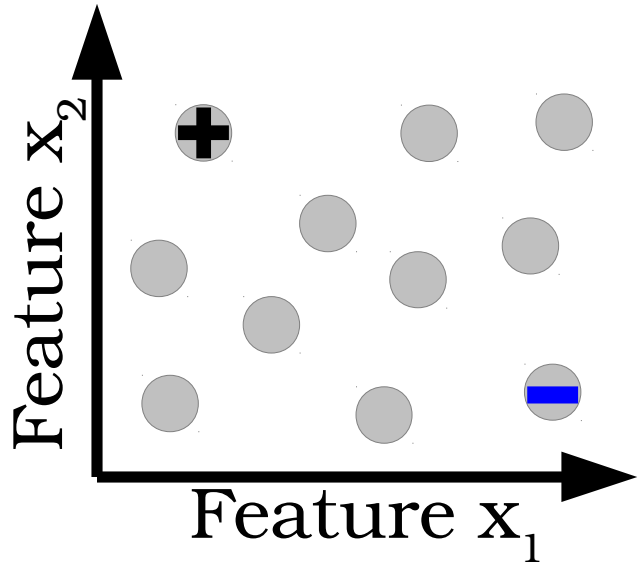
Active Learning Strategies: Motivating Example



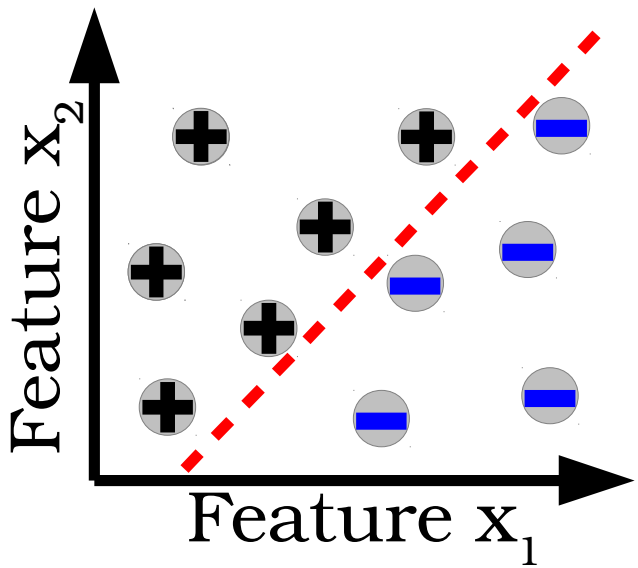
Example

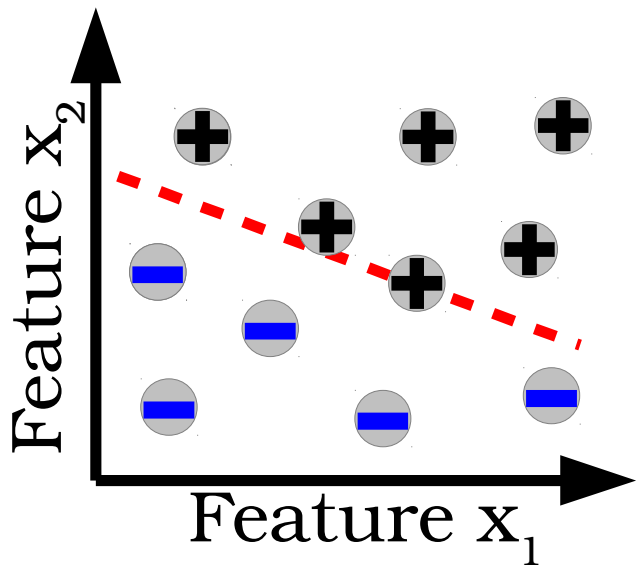
- Uniform distributed instances
- Which label to request?
- Idea: Try to *maximise diversity* of requested labels

Active Learning Strategies: Motivating Example



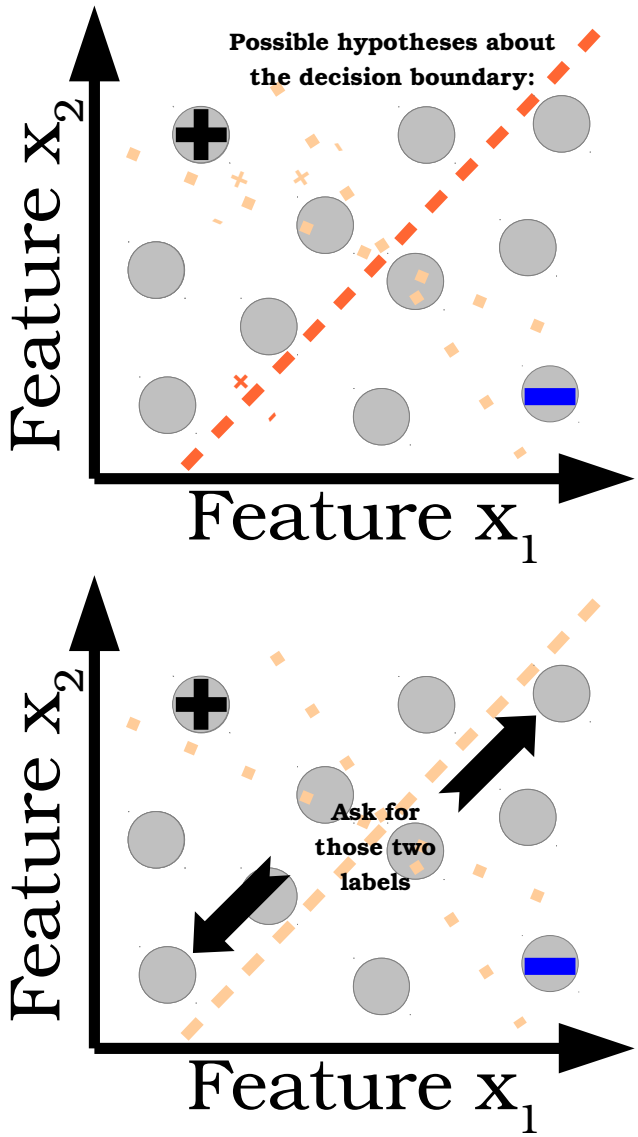
Based on these two labels, what are possible true label distributions?





and
many
others
. . .

Active Learning Strategies: Motivating Example



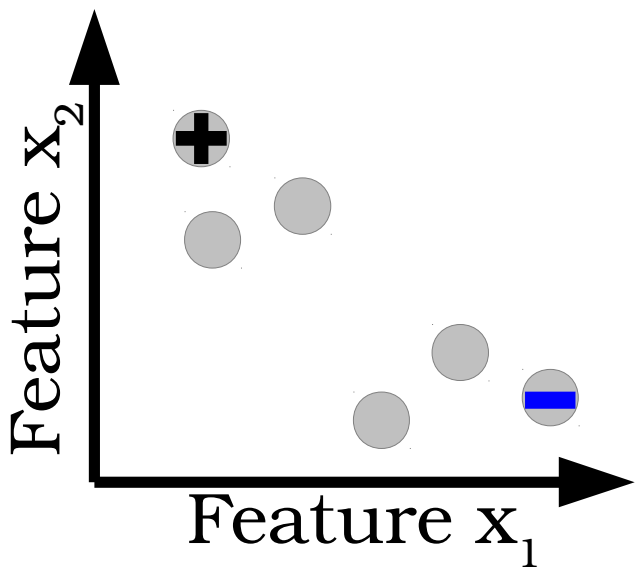
Problem

- Many distributions could have generated the data (see figure on the top left)
- How to know which one is the true one?

Core Ideas:

- *Maximise diversity* of requested labels
  - **Exploration** of feature space!
- *Reduce* the number of *remaining hypotheses* i.e. use the requested data to eliminate unlikely hypotheses
  - **Exploitation** of acquired knowledge!
  - Figure on the bottom left: Requesting labels at the top-right and bottom-left corners helps eliminating many possible hypotheses)
- But: Be aware of this **exploration vs. exploitation tradeoff**
- Furthermore, consider *density* for *sampling in high-density areas* rather than focusing on potentially irrelevant outliers

Active Learning Strategies: Motivating Example (2)



Example

- Multimodal distributed instances
- Which label to request?
- Idea: Try to *exploit structure* of data

Lecture Outline

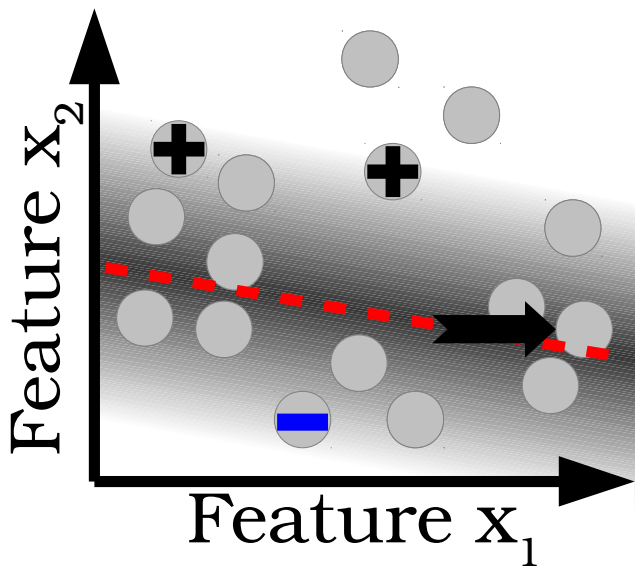
- Lecture's Context
- Semi-Supervised Learning
- Active Learning
  - Motivation
  - Approaches
    - Uncertainty Sampling
    - Expected Error Reduction
    - Ensemble / Query by Committee
    - Probabilistic Active Learning
  - Evaluation
- Summary and Questions

Active Learning Approaches

Selected Active Learning Approaches

- Uncertainty Sampling
- Expected Error Reduction
- Ensemble / Query by Committee
- Probabilistic Active Learning

Uncertainty-Based Strategy



Idea

Select those instances where we are least certain about the label

Approach:

- 3 labels preselected
- Linear classifier
- Use *distance to the decision boundary* as *uncertainty measure*

- **But: Is this a good uncertainty measure?**
- **Are there other ways of measuring uncertainty?**

Uncertainty Measures:

Three families of measures have been proposed[6]:

Margin

- Uncertainty is inverse proportional to the distance to the decision boundary
- *the closer* to the boundary, the *less certain*

Confidence (Posterior)

- Difference of the posteriors of different classes
- $abs\left(P(y=+|x) - P(y=-|x)\right)$
- *the higher* the difference, the *more certain*

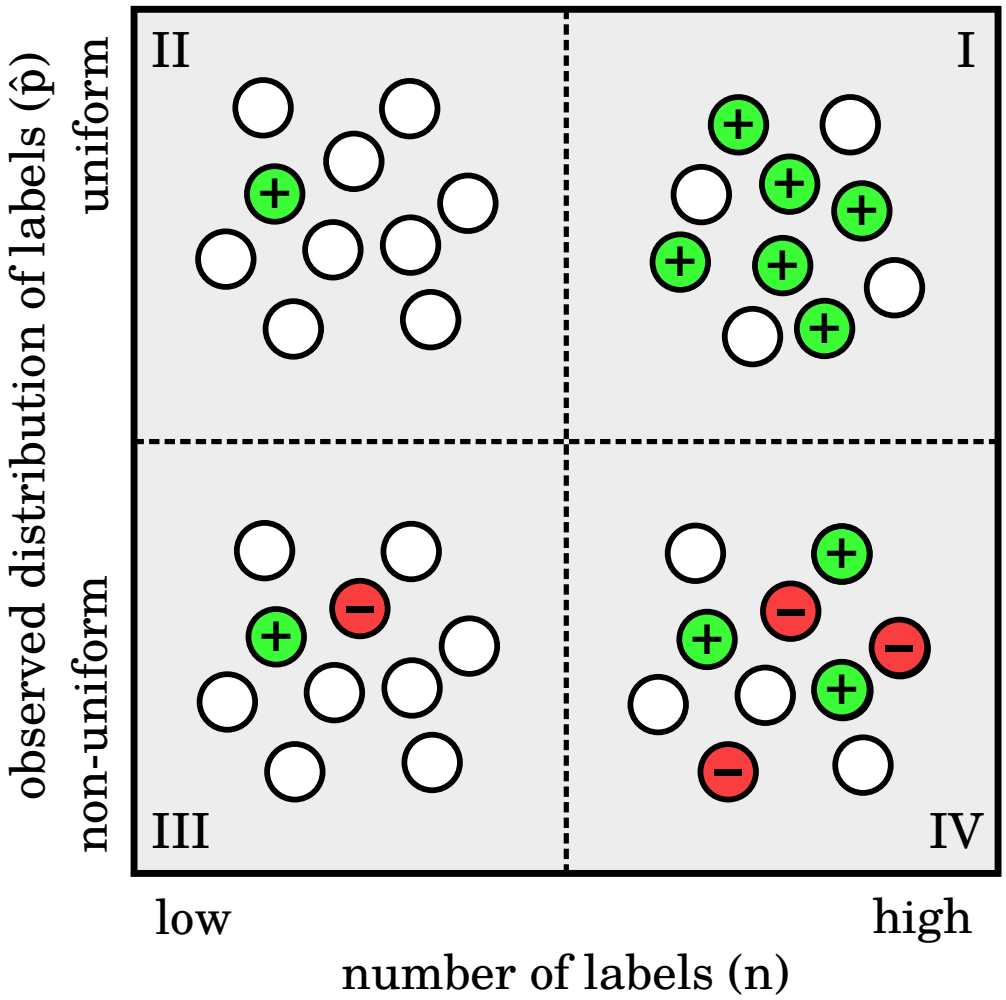
Entropy

- $-\sum_{y\in\{+,-\}} p(y|x)\log\left(p(y|x)\right)$
- *the higher* the entropy, the *less certain*

- **But how suitable are they?**

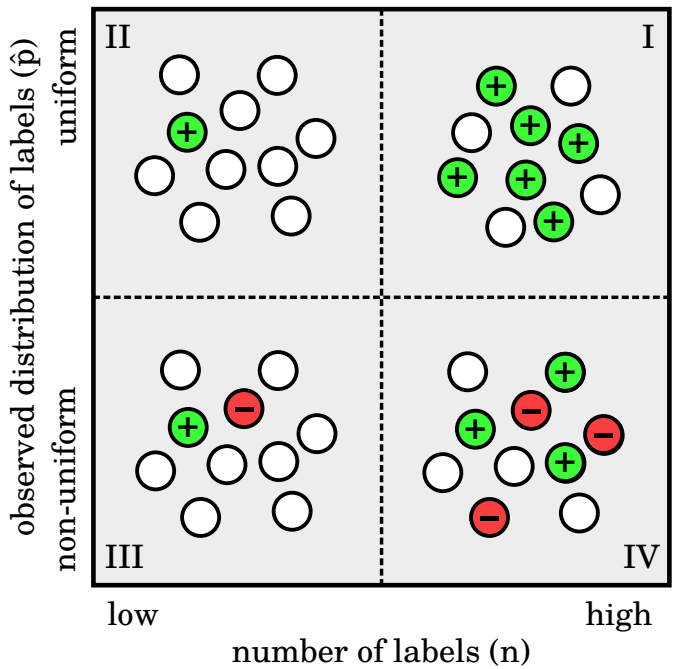
[6] See [Settles, 2012].

Exemplary AL Situations



- a label's value depends on the label information in its neighbourhood
- label information
  - number of labels
  - share of classes
- uncertainty sampling ignores **the number of similar labels**

Uncertainty Measures: Discussion



Margin

- Requests instances close to the decision boundary
- Avoids regions of low variance, ignores number of already requested labels

Confidence

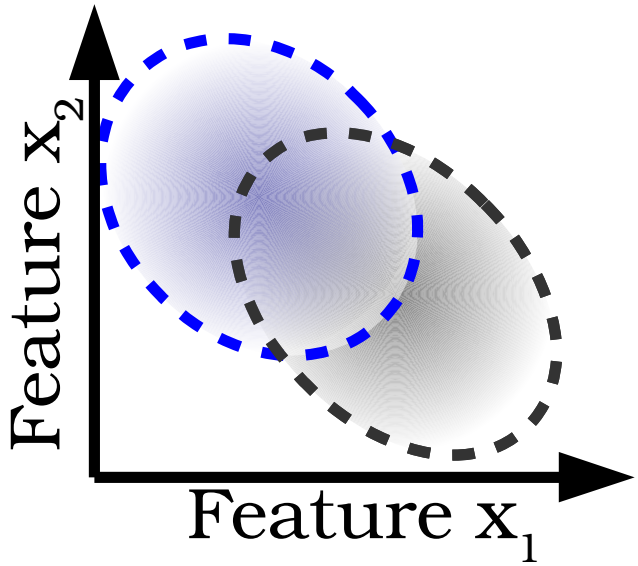
- Requests instances with low posterior difference
- Results in requests close to the decision boundary
- Again ignores the number of already requested labels

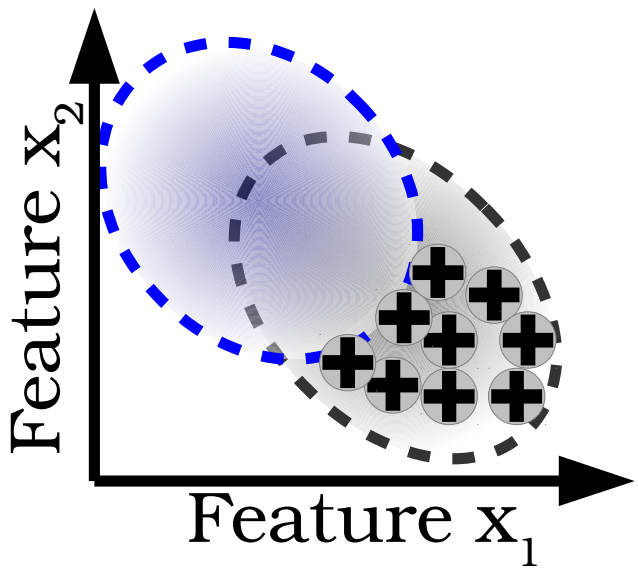
Entropy

- As above.

- **All those approaches ignore the number of previously requested labels!**

Error-Reduction Strategy





Observation

Valuable are instances that help us to improve.

Idea

Select instances that reduce the error the most

Observation

- Given the data on the left,
- generated by two Gaussian clusters,
- **where will we need many labels**?

- **Center: High variance**
  Some labels needed to learn that posteriors are equal, further labels will not improve classifier
- **Top-left, bottom-right: Low variance**
  Labels here are valuable, but a few may suffice
- **Top-right, bottom-left: Not frequent**
  The performance here is irrelevant

Error-Reduction Strategy

Formalisation: Compute *expected error reduction*

Approach

1. given current error rate e
2. for each possible outcome $\hat{y} \in \{+, -\}$ do
3. update classifier with $\hat{y}$
4. compute new error rate $e_{\hat{y}}$
5. compute error reduction $\delta_{\hat{y}} = e - e_{\hat{y}}$
6. compute weight $w_{\hat{y}}$ for this outcome $\hat{y}$
7. end
8. compute expected error reduction $E[\delta] = \sum_{\hat{y} \in \{+,-\}} w_{\hat{y}} \cdot \delta_{\hat{y}}$
9. request the label that maximises $E[\delta]$

Questions

1. How compute outcome-weights $w_{\hat{y}}$?
2. How compute new error rate $e_{\hat{y}}$?
   Specifically: On which data?

Error-Reduction Strategy

Outcome-Weights Computation

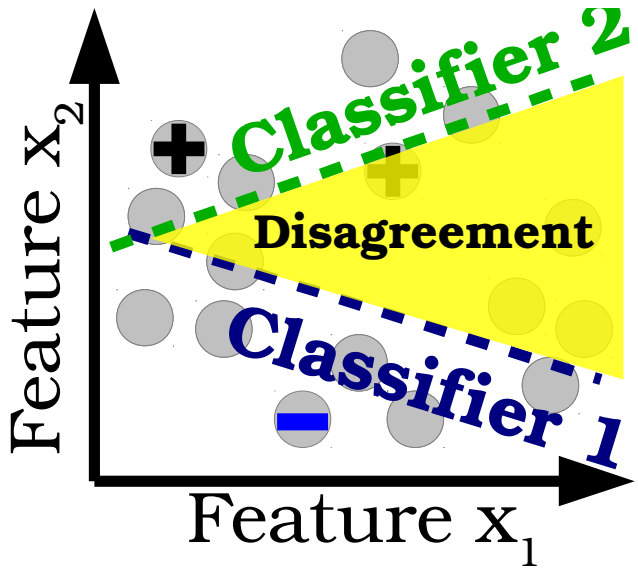
1. Average: Equal weights $w_{\hat{y}} = 0.5$ for all $\hat{y}$
2. Most likely: Weight the most likely outcome one
   (most likely based on the current classifiers prediction), the other outcome zero
3. Posterior-based: Use posterior estimates by current classifier as weights:
   $w_{\hat{y}} = \hat{p}(y = \hat{y}|x)$

- Variant 1: Favours exploration
  Only good for cases of high label variance
- Variant 2: Favours exploitation, similar to self-training
  Risk of getting stuck in the wrong hypothesis
- Variant 3: Inbetween one and two

Error-Rate Computation

- Use current classifier to label the unlabelled instances
  Similar to *self-training*
- Good estimate of the *generalisation capability*
- But: Might again lead to getting stuck in the wrong hypothesis

Ensemble-Based Strategy



Idea

Use disagreement between base classifiers

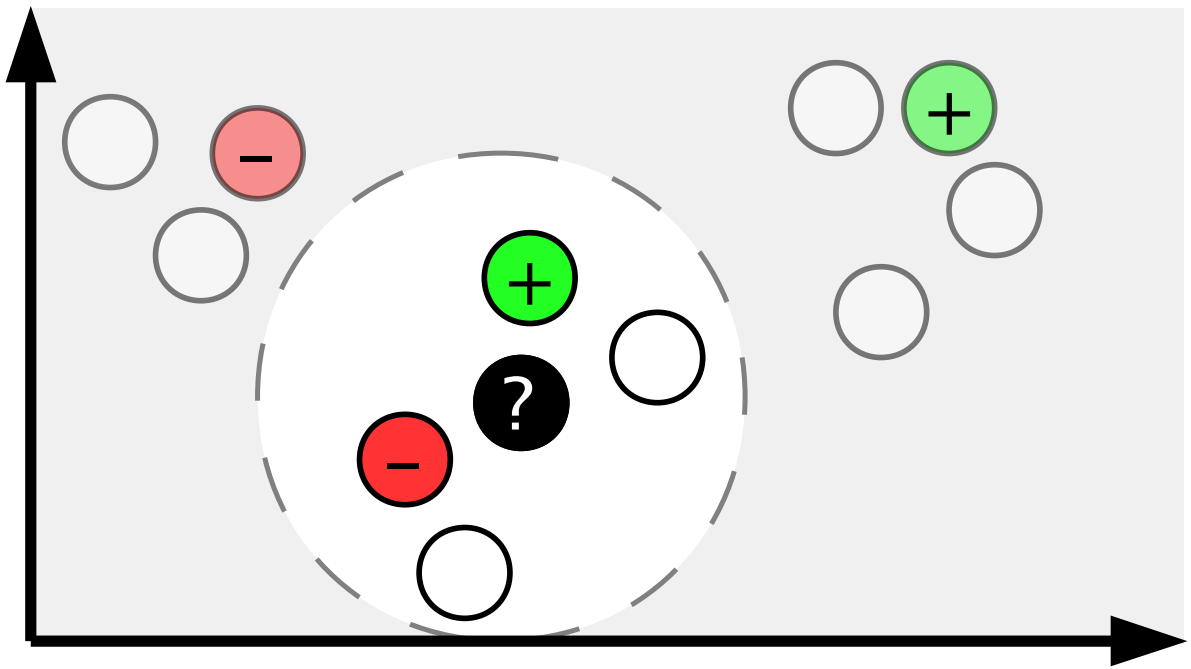
Approach

1. Get an initial set of labels
2. Split that set into (overlapping) subsets
3. On each subset, train a different base-classifier
4. Repeat
5. On each unlabelled instance do
6. Apply all base-classifiers
7. Request label, if base-classifiers disagree
8. Update all base-classifiers
9. Go to step 4
10. Until convergence

Probabilistic Active Learning: Illustrative Example

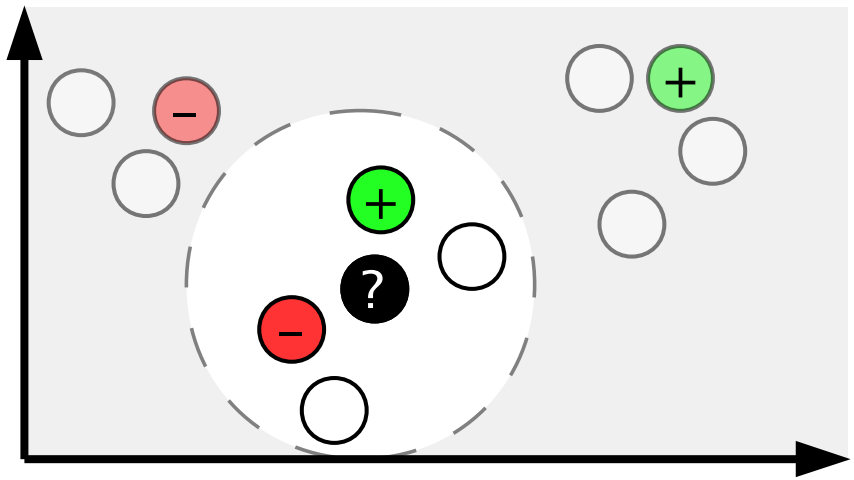
- Given: Dataset with labelled (⊖/⊕) and unlabelled (○) instances
- Objective: Select the most valuable candidate for labelling

How?



- **Decision Theoretic Approach:** Determine the gain in classification performance from labelling a candidate such as, e.g., $x_?$ (❓)

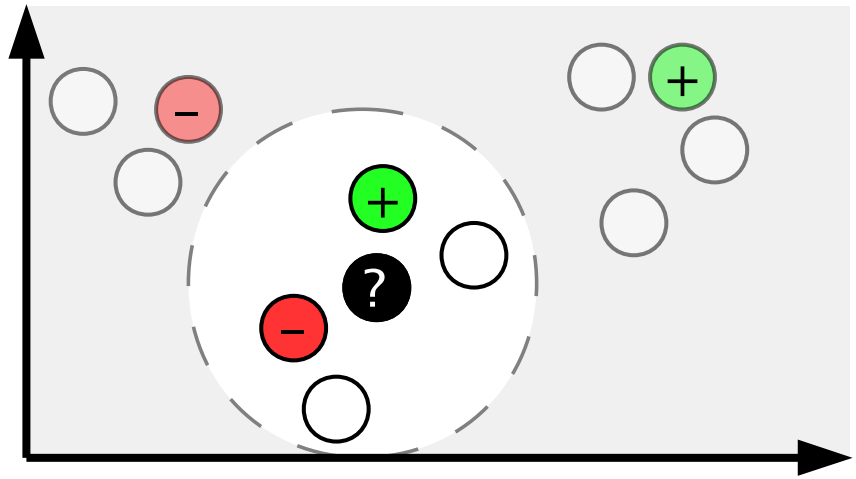
Principles of Probabilistic Active Learning [7]



- **How would we classify instances at the position of ?**
- Two neighbouring labels, half of them positive, half of them negative (**observed posterior** probability $\hat{p} = \hat{Pr}(Y = +|X = x_?) = 0.5$)
- Tie, e.g. let's classify instances here as negative

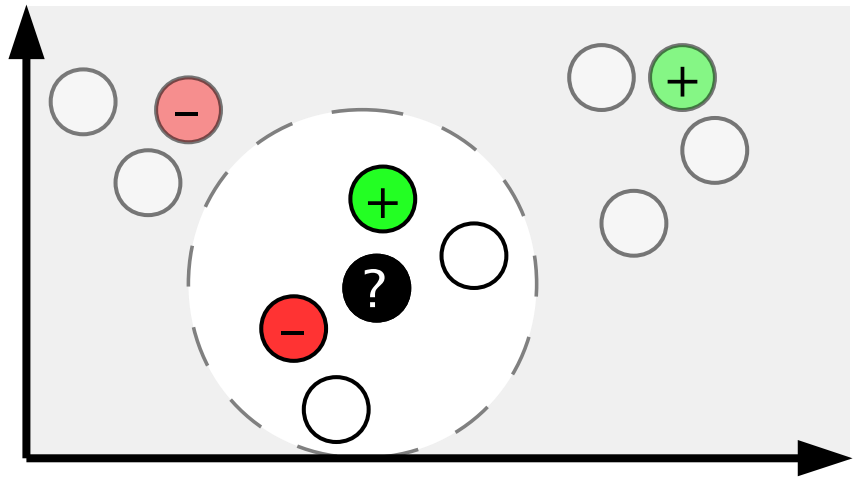
[7] See [Krempl et al., 2014a].

Principles of Probabilistic Active Learning



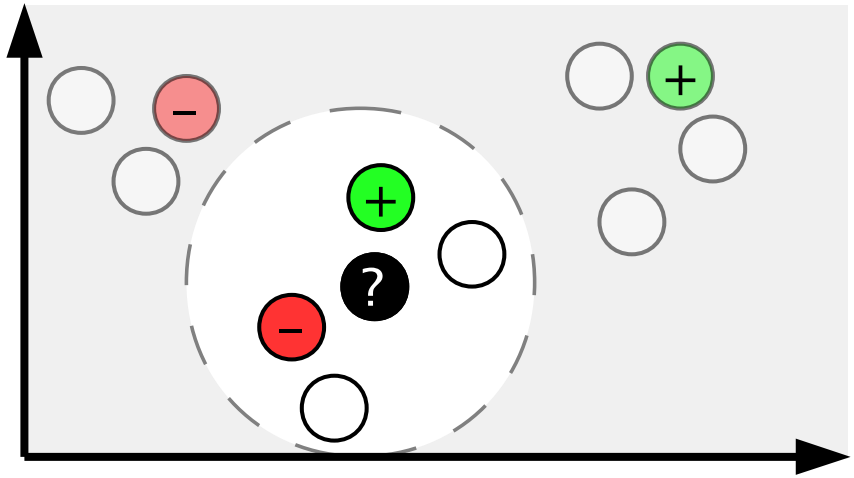
- **What is our current classification performance here, e.g. accuracy?**
- Depends on the **true posterior** probability $p = Pr(Y = +|X = x_?)$ here
- Unknown, but let's assume it to be, e.g., $p = 0.6$
- Then, our current accuracy is 40%

Principles of Probabilistic Active Learning



- **What happens if we get the label $y_?$ of $x_?$?**
- Depends on the **label's realisation**
- Unknown (yet), but let's assume it to be, e.g., positive: $y_? = +$
- Then, our classification rule changes to predicting positive labels
- Consequently, our current accuracy increases to 60%
- This results in a **gain in classification performance** by $+0.2$

Principles of Probabilistic Active Learning

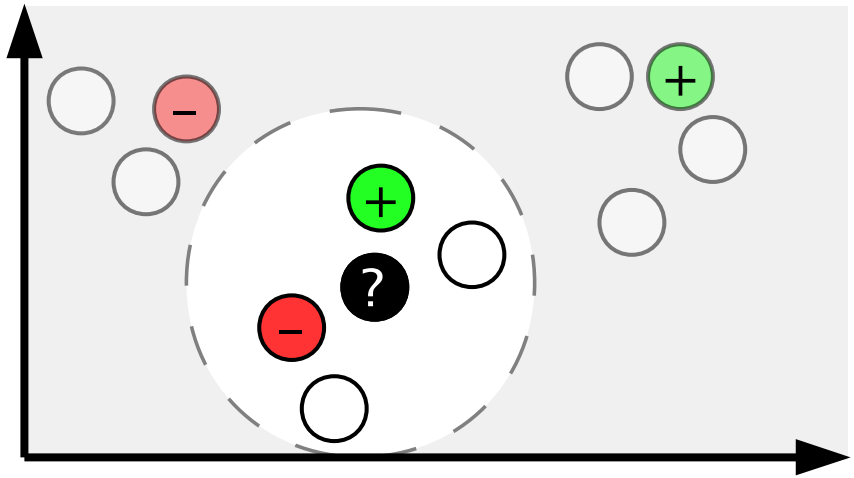


- **In reality we do not know the label realisation $y_?$ yet!**
- However, $y_?$ is the realisation of a **Bernoulli trial**:

$$\begin{array}{ccc} & y_? & \\ p \swarrow & & \searrow (1-p) \\ + & & - \end{array} \tag{1}$$

- Probability of success is the **true posterior** p
- E.g., in the case above, we get with $p = 0.6$ a positive label

Principles of Probabilistic Active Learning



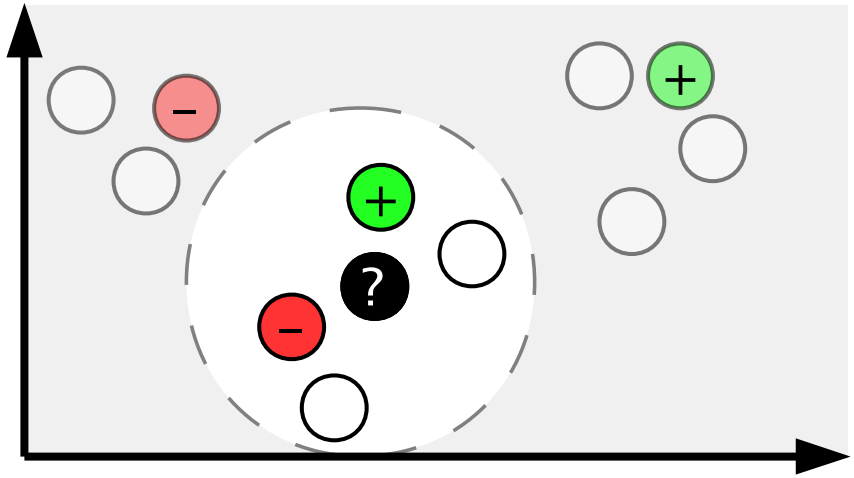
- **But how to determine p?**
- Unknown, but the already available labels provide some information:

  Let's summarise them by **label statistics** $\mathcal{ls} = (n_+, n_-)$, where
  - $n = n_+ + n_- = 2$ is the number of labels
  - $\hat{p} = \frac{n_+}{n_+ + n_-} = \frac{1}{2}$ is the share of positives therein (posterior estimate)

- These labels are the realisation of n Bernoulli trials with success probability p (*cmp. urn model*):

$$n_+ \sim \text{Binomial}_{n,p} = \text{Bin}_{n,p} \tag{2}$$

Principles of Probabilistic Active Learning



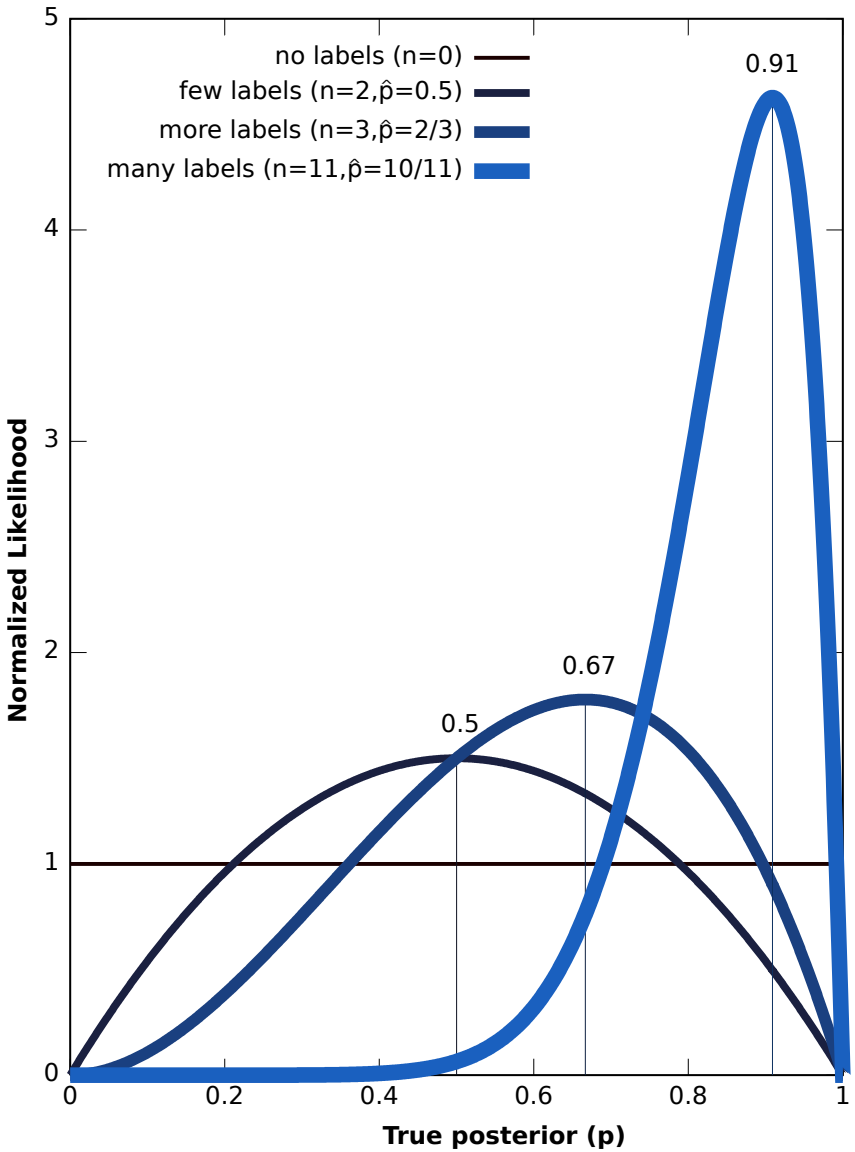
- The likelihood of p (as **parameter** of the above Binomial distribution), given the data $\mathcal{L} = (n_+, n_-)$ is

$$L(p|\mathcal{L}) = \mathrm{Bin}_{n_+ + n_-, p}(n_+) = \frac{\Gamma(n_+ + n_- + 1)}{\Gamma(n_+ + 1) \cdot \Gamma(n_- + 1)} \cdot p^{n_+} \cdot (1-p)^{n_-} \qquad (3)$$

- Normalising this likelihood yields the Beta distribution:

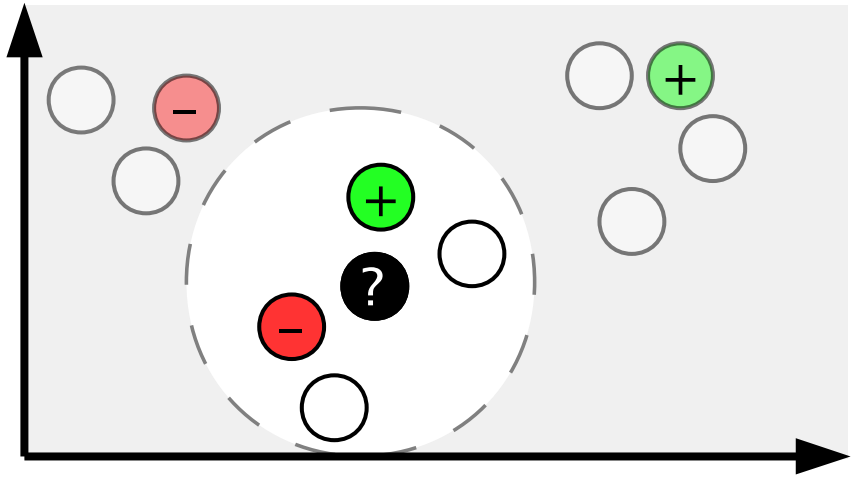
$$\omega_{\mathcal{L}}(p) = \frac{L(p|\mathcal{L})g(p)}{\int_0^1 L(\psi|\mathcal{L})g(\psi)\mathrm{d}\psi} = (n_+ + n_- + 1) \cdot L(p|\mathcal{L}) = \mathrm{Beta}_{\alpha,\beta}(p) \qquad (4)$$

Probabilistic Active Learning – Excursus: Beta Prior



- Uniform prior: Prior to the first label's arrival, all values of p are assumed equally plausible.
- A Bayesian approach yields for the normalised likelihood corresponds a beta distribution with parameters:
  - α Number of positive labels plus one
  - β Number of negative labels plus one
- Left: Plot of normalised likelihoods for different values of α, β
- The peak of this function becomes the more distinct, the more labels are obtained.

Principles of Probabilistic Active Learning



- Thus, given label statistics $\mathcal{ls} = (n_+, n_-)$ summarising labelled data,
- we have **two random variables**
  - the true posterior $P \sim \mathrm{Beta}_{n_+ +1, n_- +1}$
  - the label realisation $Y \sim \mathrm{Bernoulli}_p = \mathrm{Ber}_p$
- and calculate the **performance gain in expectation** over p, y

$$\mathrm{pgain}(\mathcal{ls}) = \mathrm{E}_p\left[\mathrm{E}_y\left[\mathrm{performancegain}_p(\mathcal{ls}, y)\right]\right] \tag{5}$$

$$= \int_0^1 \mathrm{Beta}_{\alpha,\beta}(p) \cdot \sum_{y \in \{0,1\}} \mathrm{Ber}_p(y) \cdot \mathrm{gain}_p(\mathcal{ls}, y)\,\mathrm{d}p \tag{6}$$

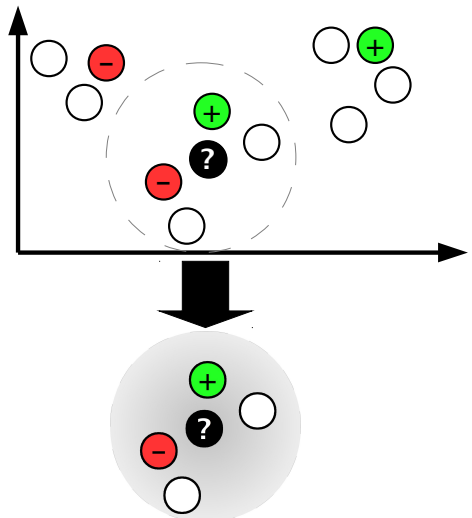
Probabilistic Gain

This **probabilistic gain** quantifies

- the *expected change* in classification performance
- at the *candidate's position* in feature space,
- in each and every *future classification* there,
- given that *one additional label* is acquired.

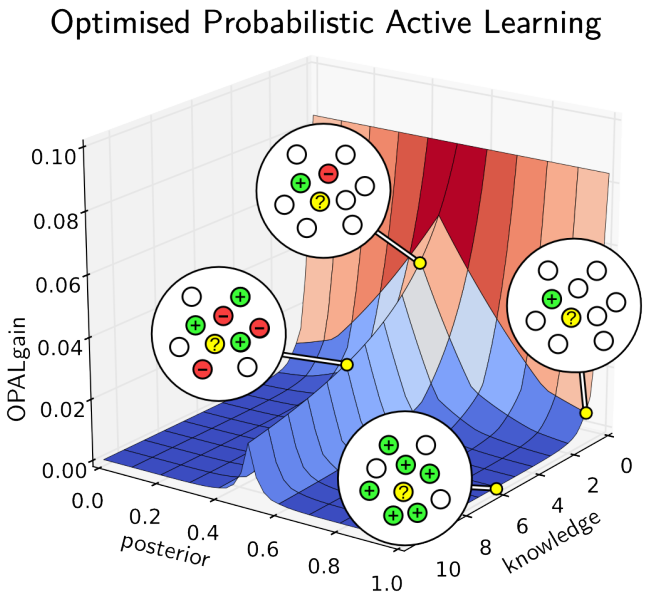
- Weight pgain with the density d_x *over labelled and unlabelled data* at the candidate's position.

- Select the candidate with highest density-weighted probabilistic gain.



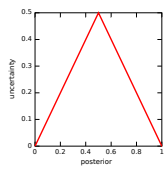
Probabilistic Gain – G_{OPAL}

Probabilistic Gain for Equal Misclassification Costs



- The probabilistic gain in accuracy as a function of $\mathcal{ls} = (n, \hat{p})$ is
- monotone with variable n,
- symmetric with respect to $\hat{p} = 0.5$,
- zero for irrelevant candidates.
- Compare to uncertainty:

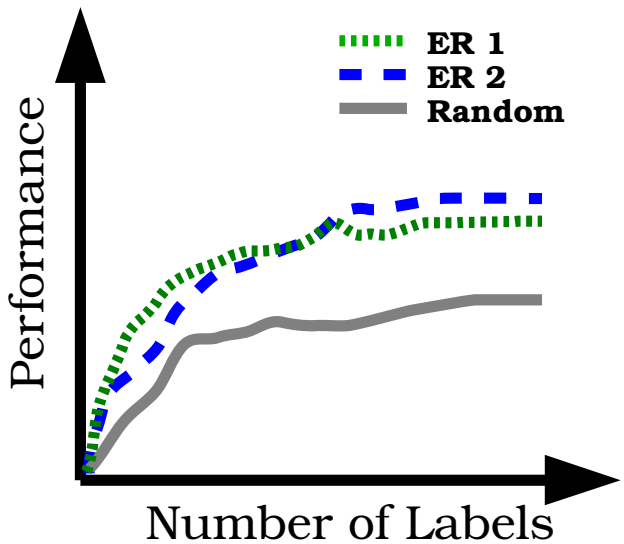
(in confidence) const. w.r.t. n:



Active Learning Strategies - Discussion

- Uncertainty-Based:
  fast and simple, *waste labels* in regions with high Bayesian error rate
- Error-Reduction:
  good quality, *high computational cost*
- Ensemble-Based:
  good quality, *convergence avoids unnecessary sampling* in regions with high Bayesian error rate, *conceptually easily understandable*, *multiple classifiers needed*
- Probabilistic Approach:
  good quality, *considering labelstatistics avoids unnecessary sampling* in regions with high Bayesian error rate

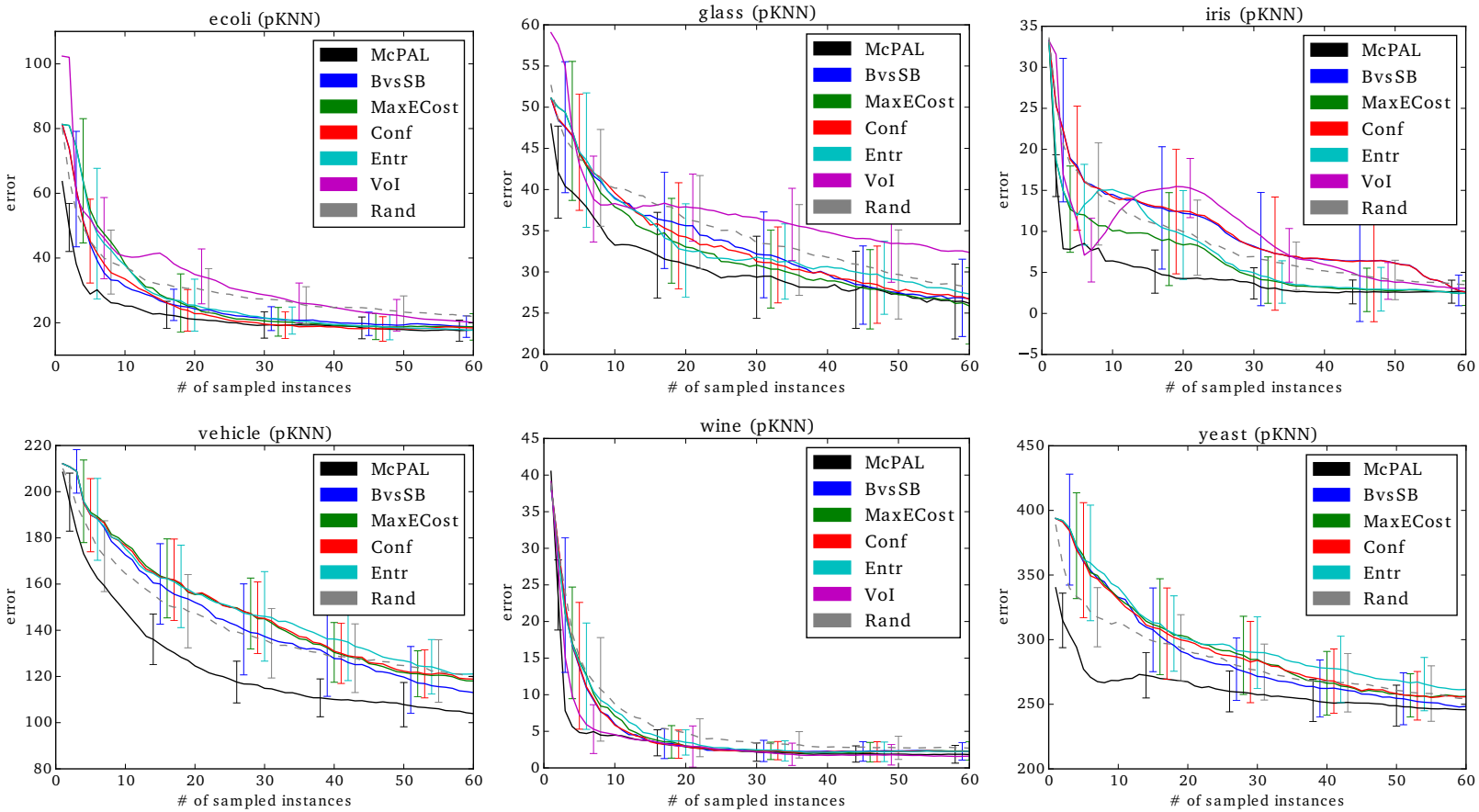
Evaluating Active Learning Approaches



Learning Curves

- Plot performance against the number of requested labels
- Expected Behaviour:
  - Performance increases with number of labels
  - Convergence: after ∞ label requests, all strategies should have the same performance
- Caveat: Allways compare using the same classifier!

Evaluation Example: Multi-Class Probabilistic AL (McPAL): (Selection[8])



Learning curves of mean misclassification cost (including standard deviation as error bars) of McPAL and its competitors on all six datasets using the pKNN classifier.

[8] For more results, please consult ECAI paper [Kottke et al., 2016].

Lecture Outline

- Lecture's Context
- Semi-Supervised Learning
- Active Learning
  - Motivation
  - Approaches
    - Uncertainty Sampling
    - Expected Error Reduction
    - Ensemble / Query by Committee
    - Probabilistic Active Learning
  - Evaluation
- Summary and Questions

Summary and Questions

Summary

- Motivation & Aim: Not all labels are available
- General Idea: Use structure of the data!
- Semi-Supervised Learning:
  - Assumptions: Smoothness, Clusters, or Manifold
  - Approaches: Self-Learning, Generative Models, Low-Density Separation, Graph-Based Models
- Active Learning:
  - Scenarios: Query Synthesis, Pool, Stream
  - Approaches: Uncertainty Sampling, Expected Error Reduction, Query by Committee, Probabilistic Active Learning
  - Evaluation: Learning Curves

Questions?

Bibliography I


Angluin, D. (1988).
Queries and concept learning.
Machine Learning, 2:319–342.


Attenberg, J., Melville, P., Provost, F., and Saar-Tsechansky, M. (2011).
Selective data acquisition for machine learning.
In Krishnapuram, B., Yu, S., and Rao, R. B., editors, *Cost-Sensitive Machine Learning*. CRC Press, Boca Raton, FL, USA, 1st edition.


Chapelle, O., Schölkopf, B., and Zien, A., editors (2006).
Semi-supervised Learning.
MIT Press.


Cohn, D., Atlas, L., Ladner, R., El-Sharkawi, M., Marks, R., Aggoune, M., and Park, D. (1990).
Training connectionist networks with queries and selective sampling.
In *Advances in Neural Information Processing Systems (NIPS)*. Morgan Kaufmann.


Fedorov, V. V. (1972).
Theory of Optimal Experiments Design.
Academic Press.


Fu, Y., Zhu, X., and Li, B. (2012).
A survey on instance selection for active learning.
Knowledge and Information Systems, 35(2):249–283.

Bibliography II


Kottke, D., Krempl, G., Lang, D., Teschner, J., and Spiliopoulou, M. (2016).
Multi-class probabilistic active learning.
In Fox, M., Kaminka, G., Hüllermeier, E., and Bouquet, P., editors, *Proc. of the 22nd Europ. Conf. on Artificial Intelligence (ECAI2016), 2016*, Frontiers in Artificial Intelligence and Applications. IOS Press.


Krempl, G., Kottke, D., and Spiliopoulou, M. (2014a).
Probabilistic active learning: Towards combining versatility, optimality and efficiency.
In Dzeroski, S., Panov, P., Kocev, D., and Todorovski, L., editors, *Proceedings of the 17th Int. Conf. on Discovery Science (DS), Bled*, volume 8777 of *Lecture Notes in Computer Science*, page 168–179. Springer.


Krempl, G., Zliobaitė, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M., and Stefanowski, J. (2014b).
Open challenges for data stream mining research.
SIGKDD Explorations, 16(1):1–10.
Special Issue on Big Data.


Saar-Tsechansky, M., Melville, P., and Provost, F. (2009).
Active feature-value acquisition.
Management Science, 55(4):664–684.

Bibliography III


Settles, B. (2009).
Active learning literature survey.
Computer Sciences Technical Report 1648, University of Wisconsin-Madison, Madison, Wisconsin, USA.


Settles, B. (2012).
Active Learning.
Number 18 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers.


Zhu, X. (2008).
Semi-supervised learning literature survey.
Technical Report 1530, University of Wisconsin.


Zliobaitė, I., Bifet, A., Pfahringer, B., and Holmes, G. (2013).
Active learning with drifting streaming data.
IEEE Transactions on Neural Networks and Learning Systems, 25(1):27–39.

SSL: Approaches

- Self-Learning, Co-Learning
- Generative Models
- Low-density Separation
- Graph-Based Methods

Generative Models

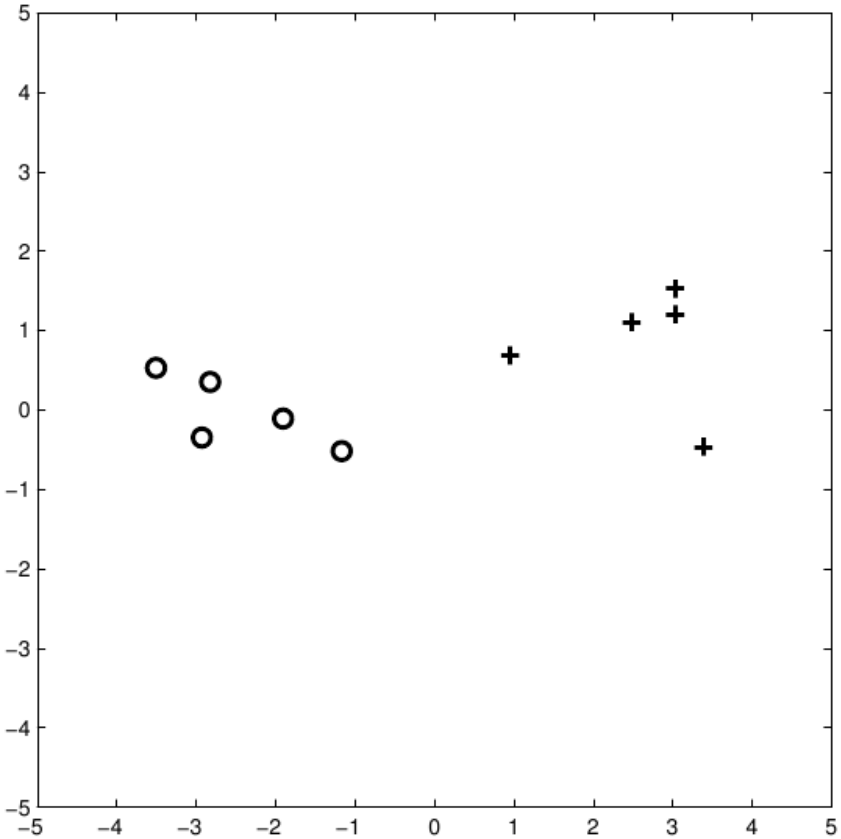
- **Assumption:**
  $P(X, Y) = P(Y)P(X|Y)$ with $P(X|Y)$ being an identifiably mixture distribution
- **Objective:**
  Identify generative model that produces the data
- **Basic Idea:**
  Learn distribution $P(X|Y)$ of each class $y \in Y$
  This requires some assumptions upon the generative model
- **Probabilistic approach:**
  Find likelihood–maximising parameter set Θ, e.g. using expectation–maximisation
- **Algorithmic approach:**
  Cluster – and – Label

Back to SSL Discussion

Generative Models - Example

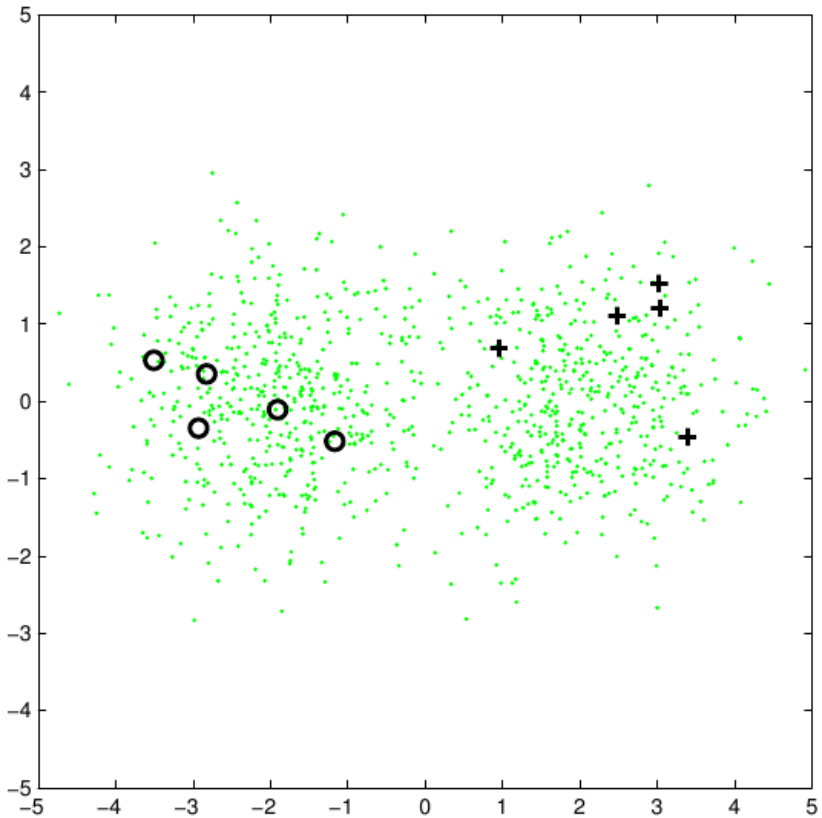
Gaussian Mixture Model Example[9]

Labelled data



Back to SSL Discussion

Labelled and unlabelled data

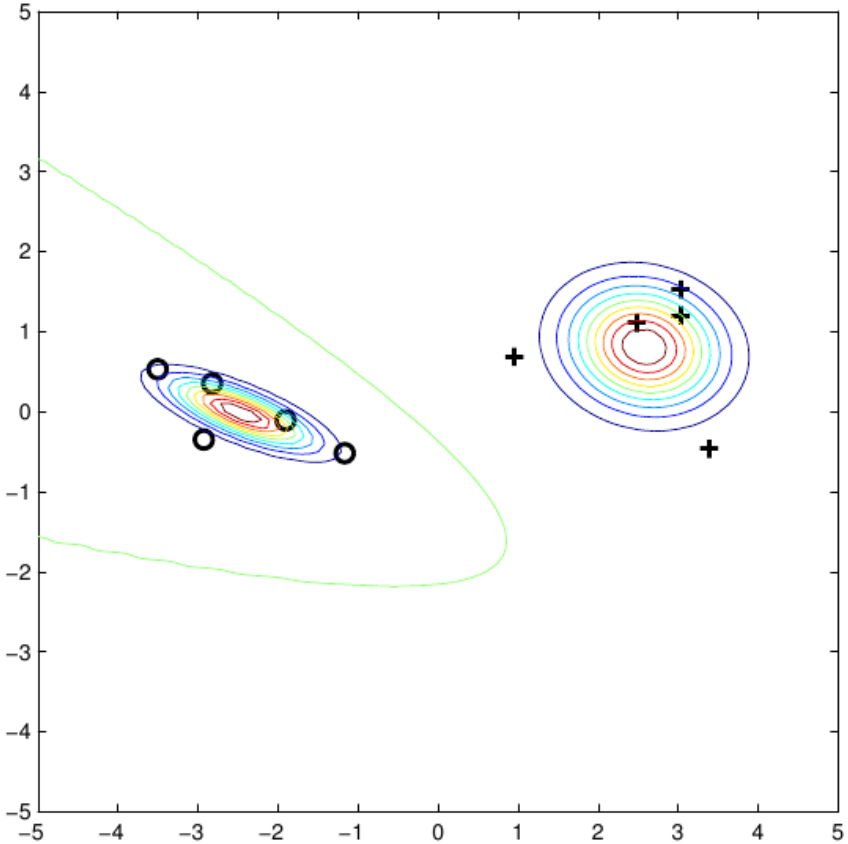


[9] Illustrations from Zhu 2005 [Zhu, 2008].

Generative Models - Example (2)

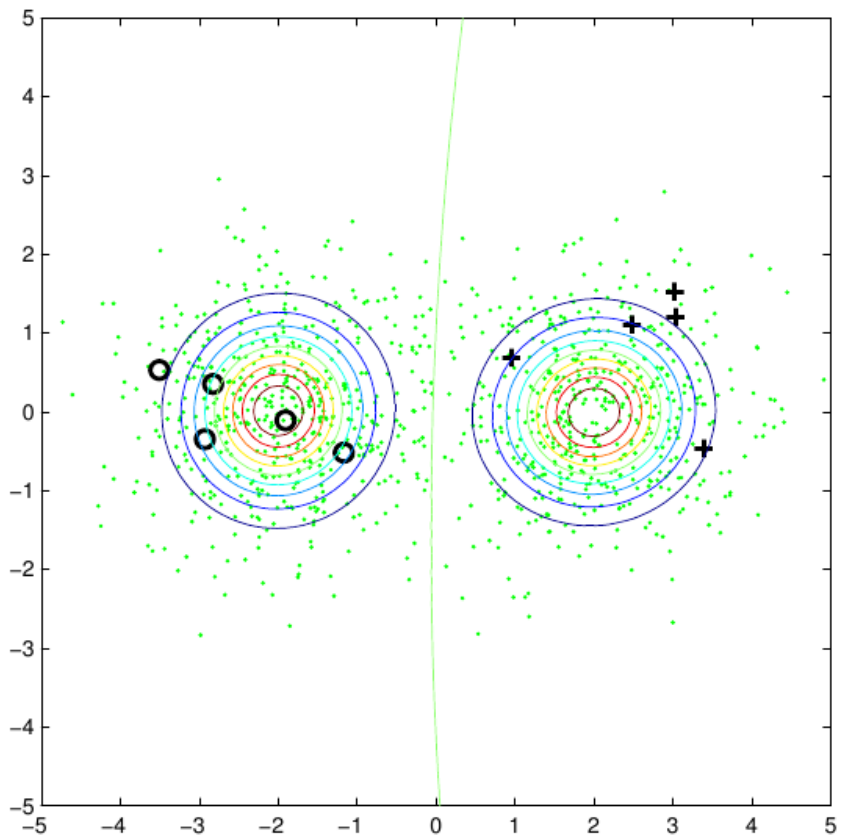
Gaussian Mixture Model Example[10]

Decision model learnt solely on labelled data



Back to SSL Discussion

Decision model learnt on all data



[10] Illustrations from Zhu 2005 [Zhu, 2008].

Generative Models - Challenges

- Identifiability
  It may happen, that solution is not unique, even if type of distributions involved is known

  Example from Zhu 2005:
  - Given $P(X) \sim U(0;1)$ and two labelled instances $(x_1 = 0.1; y_1 = +)$, $(x_2 = 0.9; y_2 = -)$
  - We know that $P(X|Y = +) \sim U(0;\tau)$ and $P(X|Y = -) \sim U(\tau;1)$, but we do not know value of τ
  - Different values of τ result in concurring models
    We can not identify the true model given the two instances
- Model correctness
  - If assumptions in the model are met, using unlabelled data increases quality of prediction model
  - But: Otherwise, model can be worse than a model based solely on labelled data
- Model fitting
  Parameter set might not be *global* optimum

Back to SSL Discussion

Low-Density Separation

- **Assumption:**
  Classes are separated by low density regions
- **Objective:**
  Placement of decision boundary in regions of low density
- **Approaches:**
  Transductive Support Vector Machines (TSVMs)
  Conventional SVM: aximize margin between labelled instances
  Transductive SVM: Maximize margin between all instances
  (i.e. labelled and unlabelled ones)

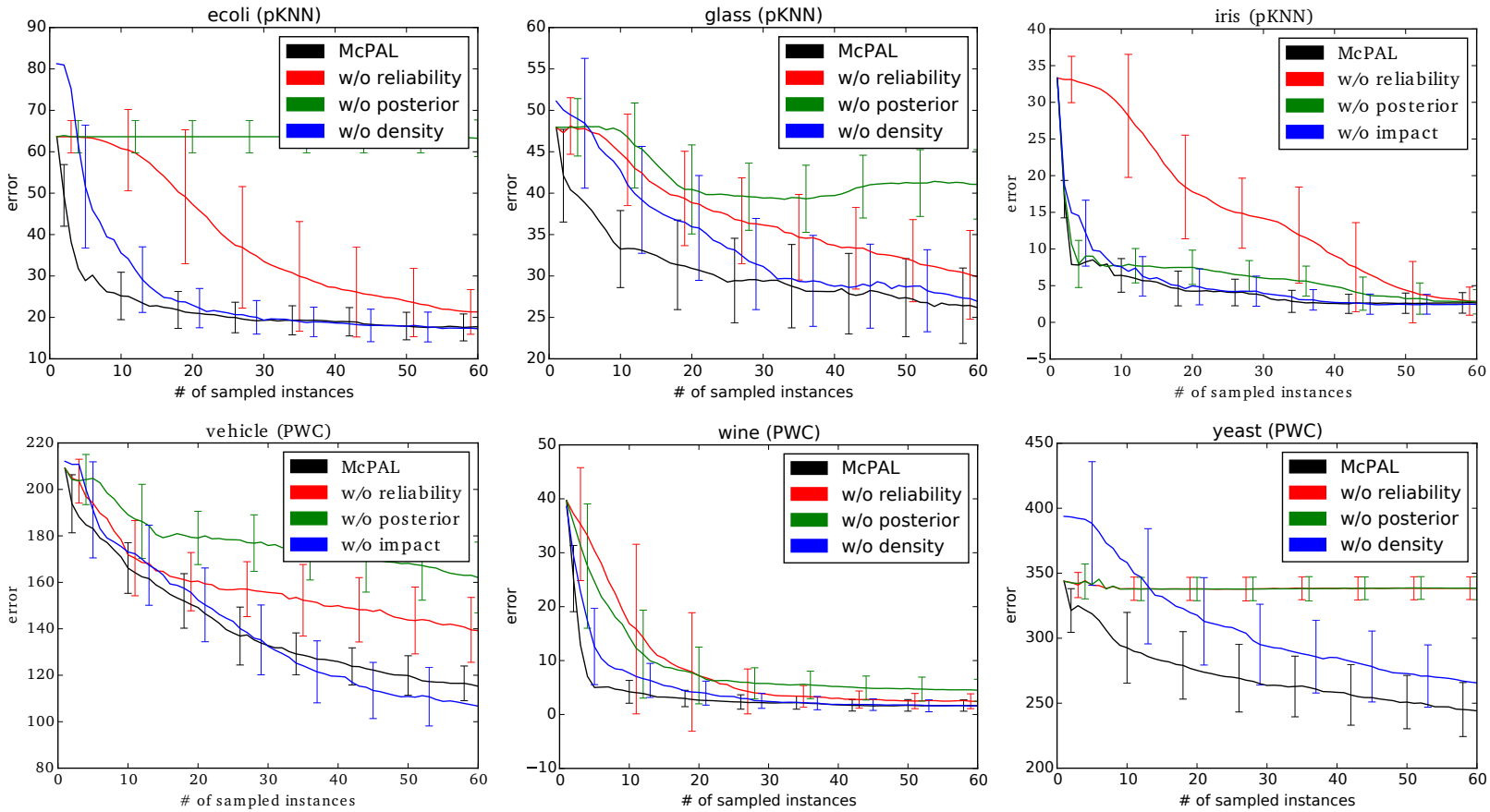
Graph-Based Methods

- **Assumption:**
  Manifold assumption, smoothness over graph
- **Approach:**
  - Represent data by a graph
  - Vertices/Nodes are labelled and unlabelled instances
    (e.g. connect an instance to its k nearest neighbours, or to all neighbours within some ϵ-neighbourhood)
  - Edges represent similarities
  - Estimate a function f on the graph,
    that is close to given labels and smooth to the whole graph
  - Use f to propagate label to neighbouring instances in graph

Back to SSL Discussion

Evaluation of Influence Factors

- Original McPAL compared to variants
- ignoring reliability: normalising the $\vec{k}$ vector to $\sum \vec{k} = n = 1$
- ignoring posterior: using uniform frequency estimate $k_i = n/C, 1 \leq i \leq C$
- ignoring density: setting it to a constant value



Learning curves in mean misclassification cost (including standard deviation as error bars).