Active Robust Learning on Distributions

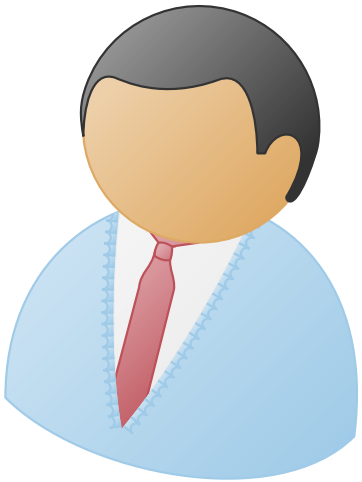
Seyed Hossein Ghafarian

This work has been done at
Ferdowsi University of Mashhad

Table of Content

Motivation: Speed and Accuracy of Learning in Human and Machines

Why we are able to learn faster and more accurate?

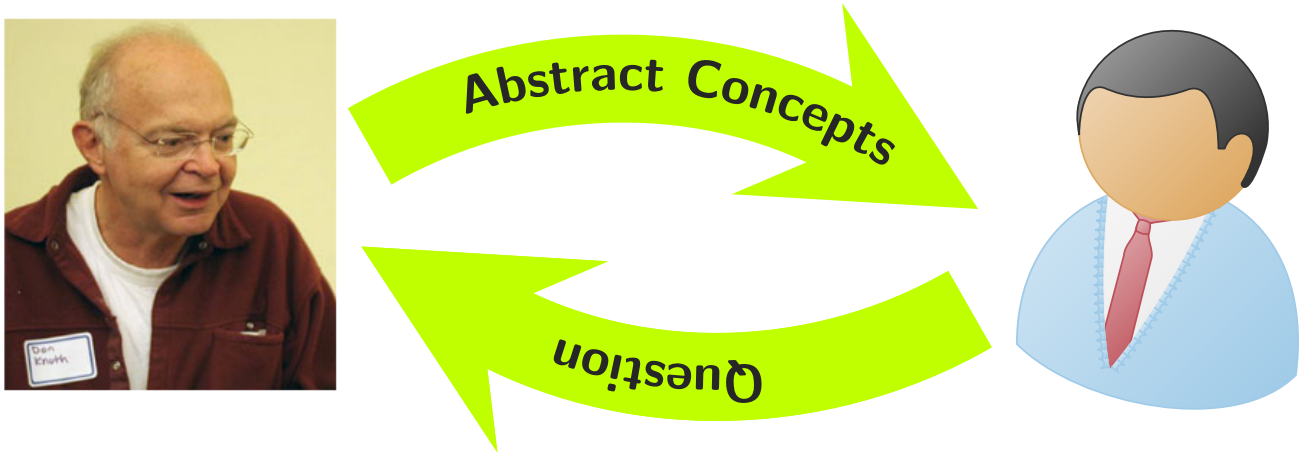


Human Learner

vs

Machine Learner

Knowledge Transfer Between Learner and Teacher



Question and answer in more abstract forms

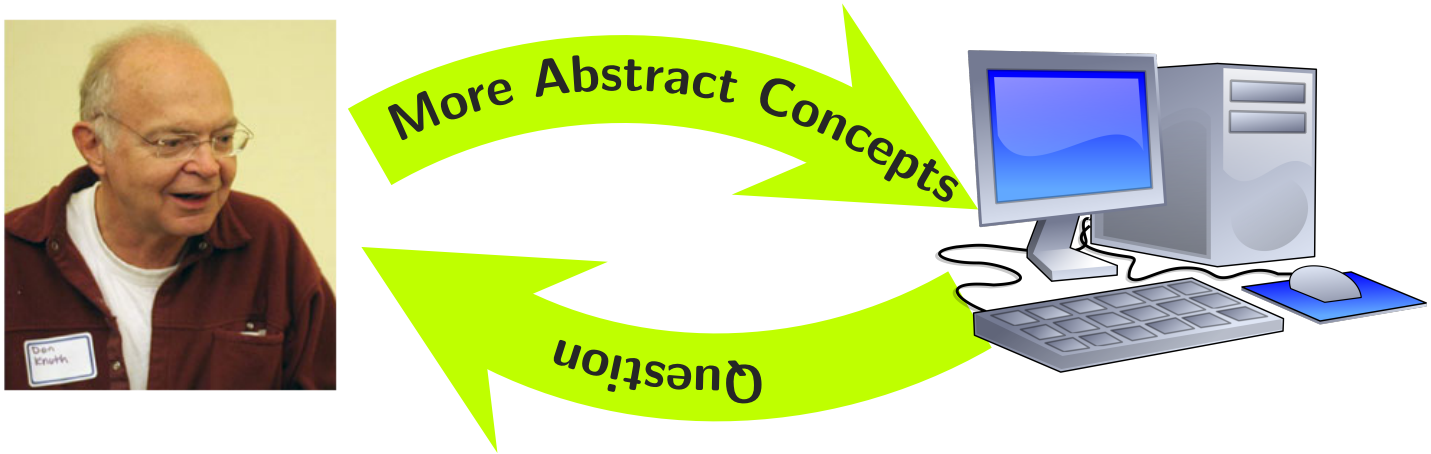
- Knowledge transfer in an abstract level
- More intelligent selection of knowledge to learn: by asking queries

Knowledge Transfer Between Learner and Oracle

One-way knowledge transfer in low-level abstraction forms

- Using lower-levels of abstractions about concepts

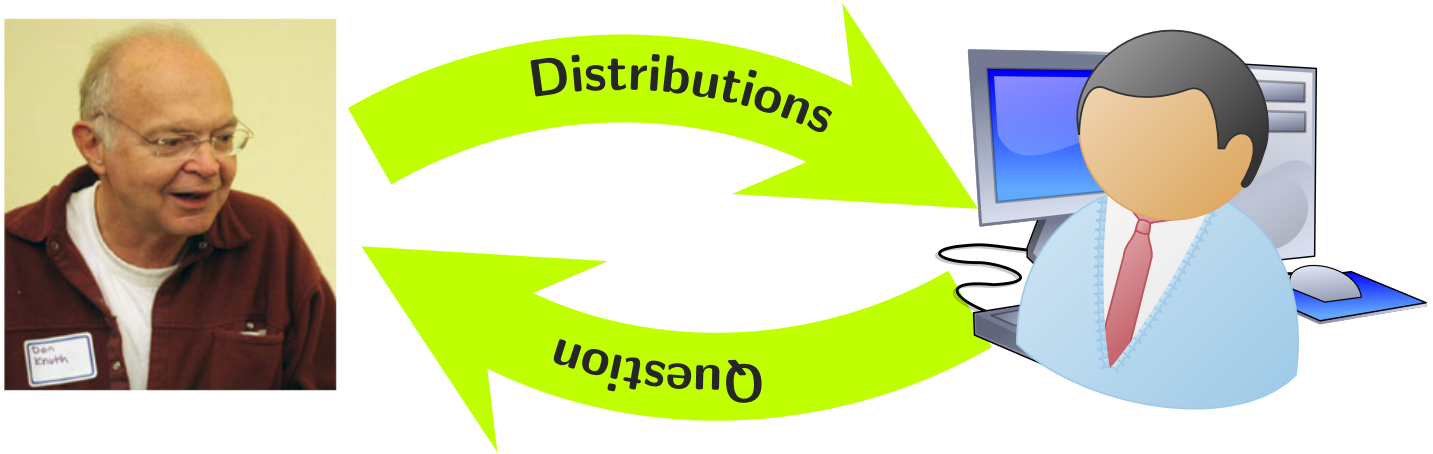
Motivation: Improving knowledge transfer between learner and oracle



To improve knowledge transfer from human to machines

- Active Learning- Questions from Oracle and accepting answers- in higher levels of abstractions.
  - Structured data
  - Distributional data

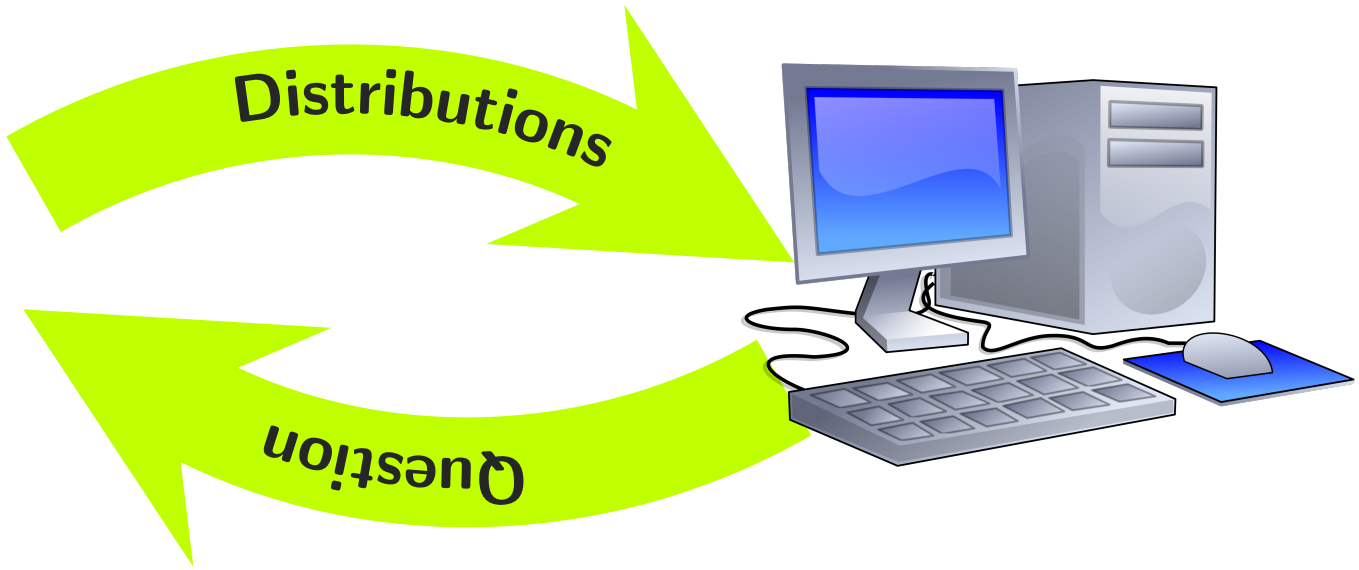
Everything starts with questions!



Querying results in

- Learner's knowledge biases toward its own queries
  - Learner's knowledge is not independent of its question!
  - Learner's knowledge does not cover all concepts proportional to their importance!

Active learning as a Game



Active learning is a game in which

- Learner wants to learn with minimum number of queries.
- Oracle wants to teach the most difficult hypothesis class to the learner.
- Active learning is a min-max problem.

Research goal

Research goal: Learning on Distributions.

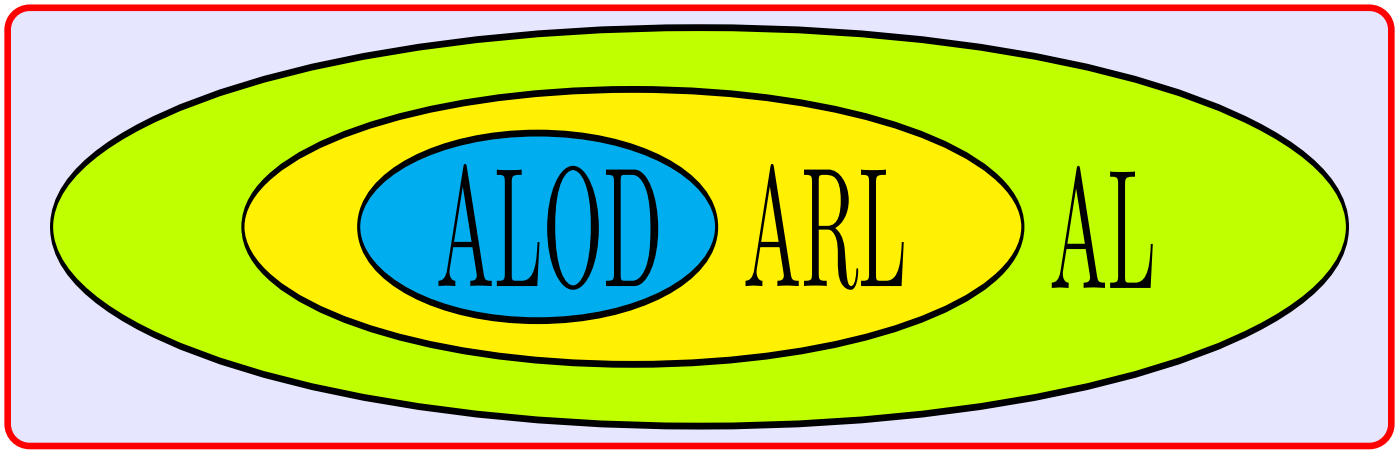
- Active Learning (AL): Proposing a consistent and principled active learning algorithm

Research goal

Research goal: Learning on Distributions.

- Active Learning (AL): Proposing a consistent and principled active learning algorithm
- Active Robust Learning (ARL): Capable of solving the challenge of inexact examples

Research goal

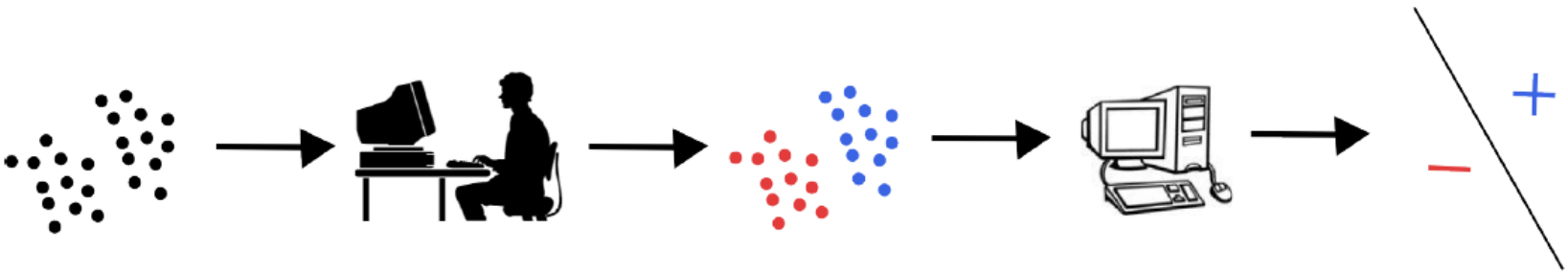


Research goal: Learning on Distributions.

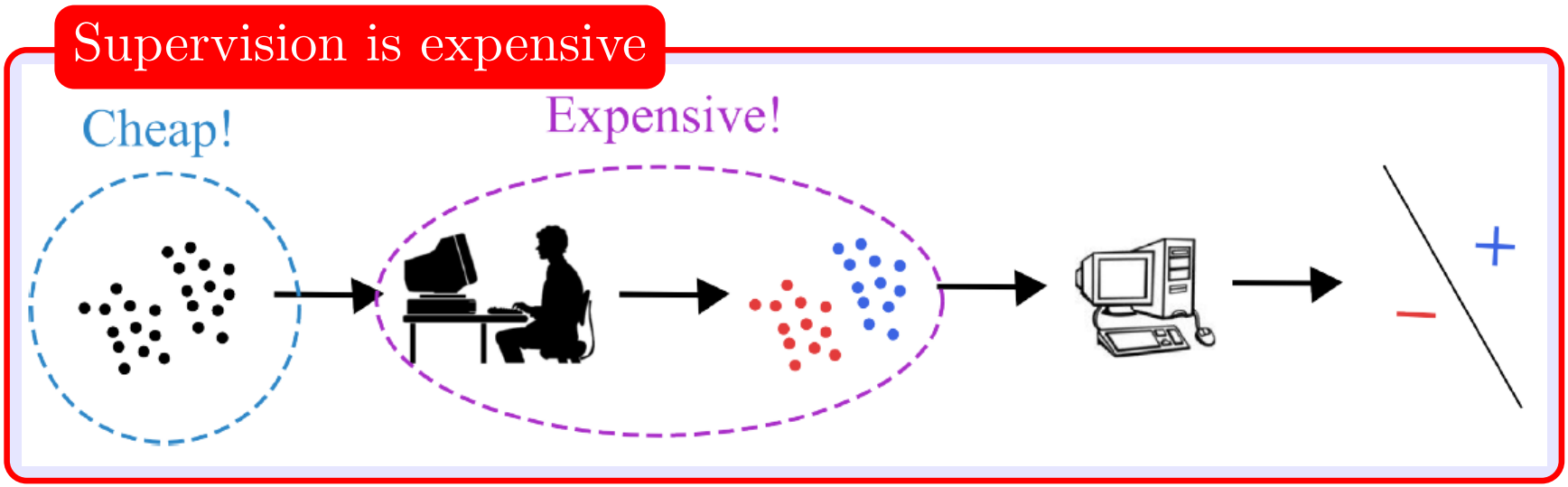
- Active Learning (AL): Proposing a consistent and principled active learning algorithm
- Active Robust Learning (ARL): Capable of solving the challenge of inexact examples
- Active Learning On Distributions (ALOD): Solving the challenge of lack of access to examples in active learning on distributions.

Supervised learning

Supervised Learning Process

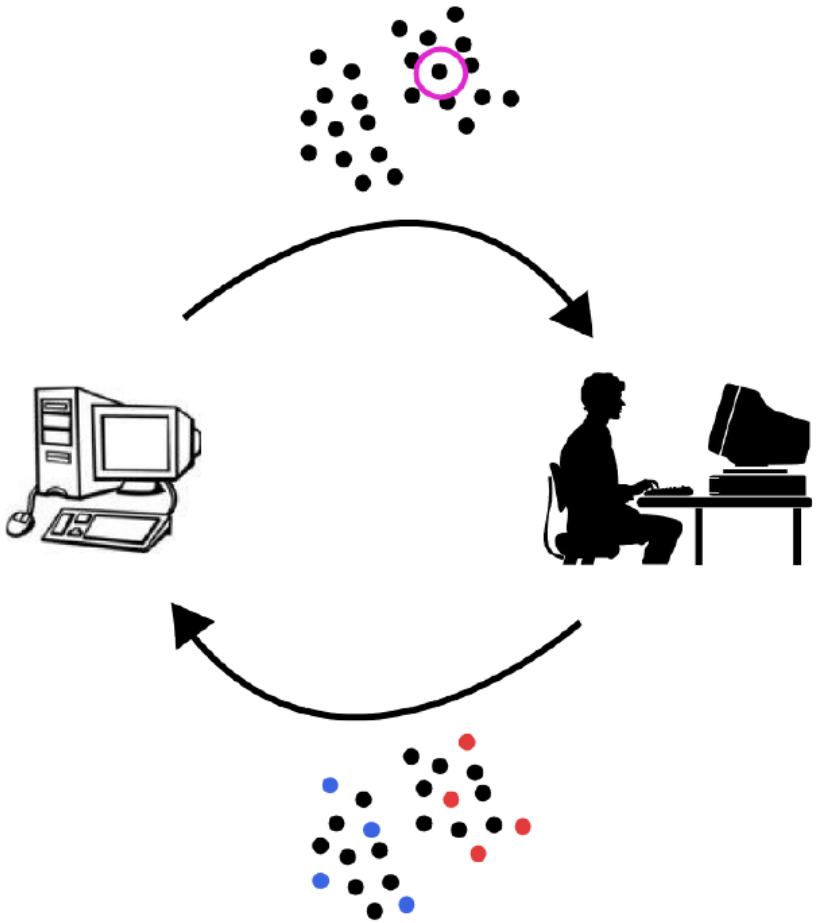


Supervised learning



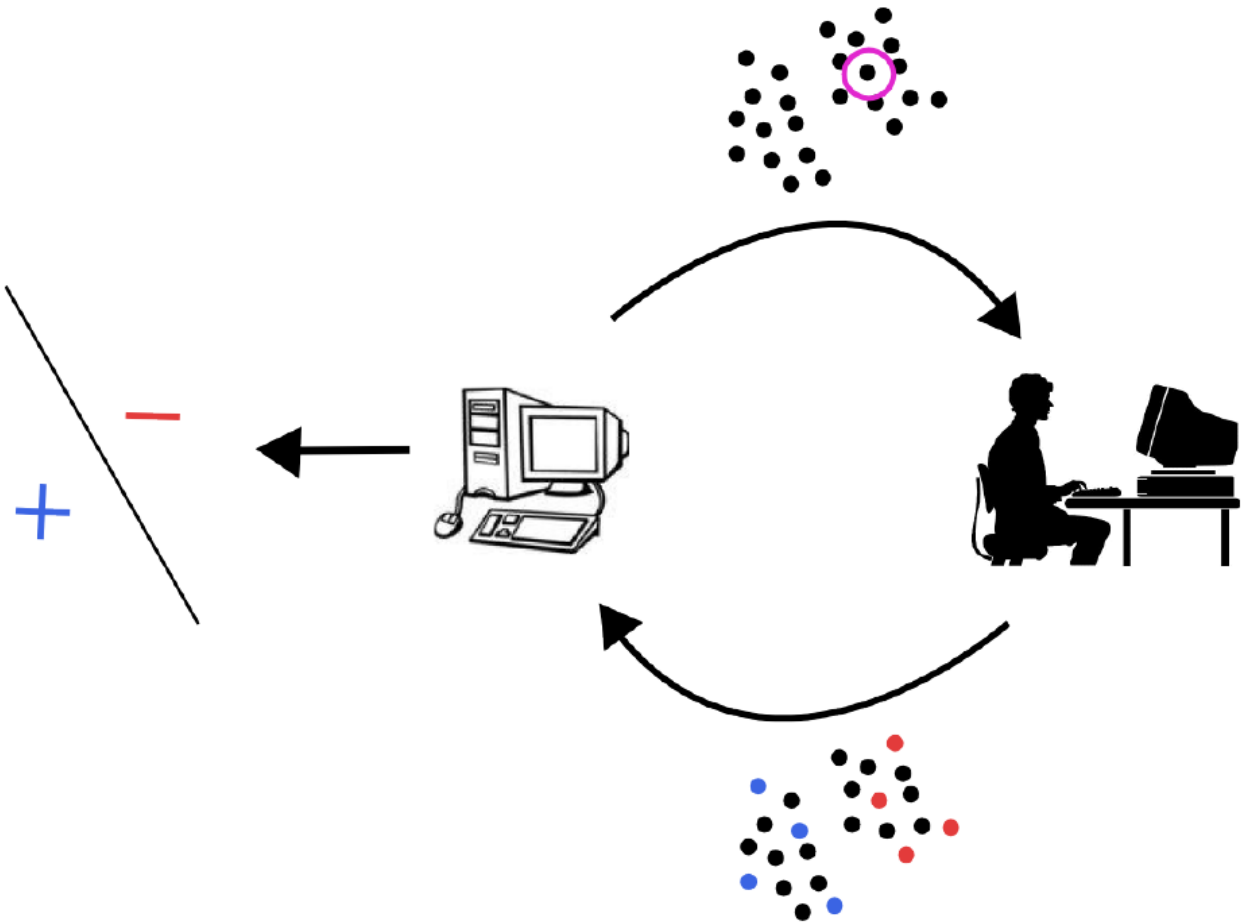
Active Learning: All examples have not been created equal

Supervision is expensive



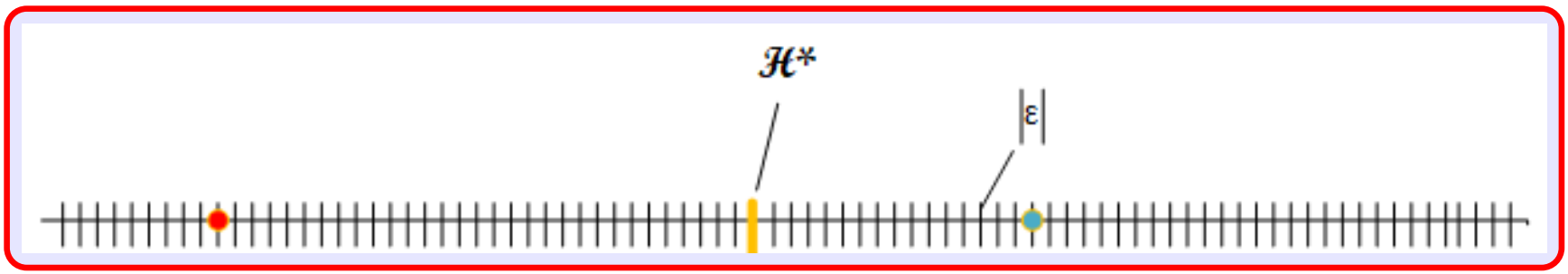
Active Learning: Concentrate on better ones

Supervision is expensive



Active Learning: A one dimensional example

$$H(x) = \begin{cases} 0 & x \leq x^* \\ 1 & x > x^* \end{cases} \tag{1}$$



For a maximum error of ϵ, supervised learning, needs $O(\frac{1}{\epsilon})$ examples.
Active learning can achieve the sampe error rate, using $O(log(\frac{1}{\epsilon}))$ examples using a simple binary search algorithm.

Active Learning: Poses new challenges

Samples are no longer Independent and Identically Distributed (i.i.d)

Almost all of machine learning algorithms, rely on samples being i.i.d, therefore, their promises are not valid given their assumptions are not.

Sampling Bias

Learning algorithms become bias to the labeled samples

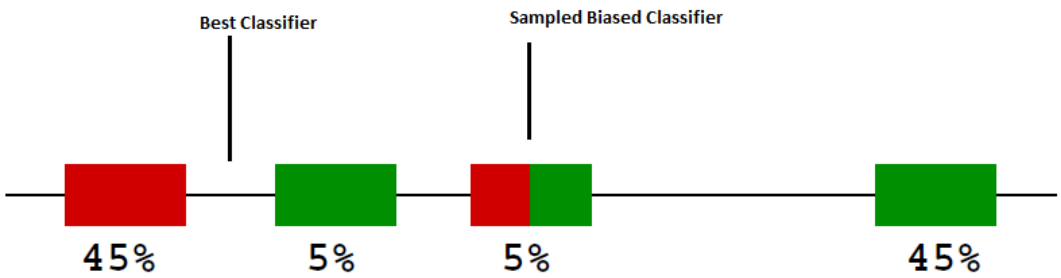
Active Learning: Sample Bias even with infinite amount of labeled data

Start with a pool of unlabeled data

Pick a few points at random and get their labels

Repeat

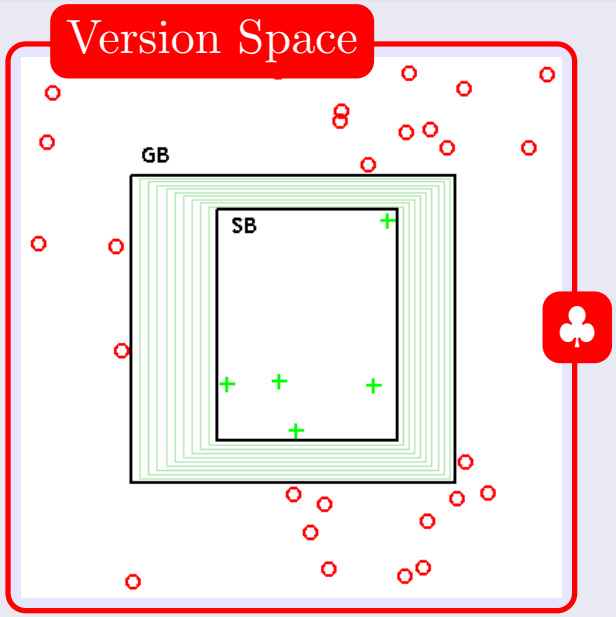
Fit a classifier to the labels seen so far
Query the unlabeled point that is closest to the boundary (or most uncertain, or most likely to decrease overall uncertainty,...)



Active Learning: Two Important Definitions

Version Space

Set of all classifiers who correctly classify labeled data



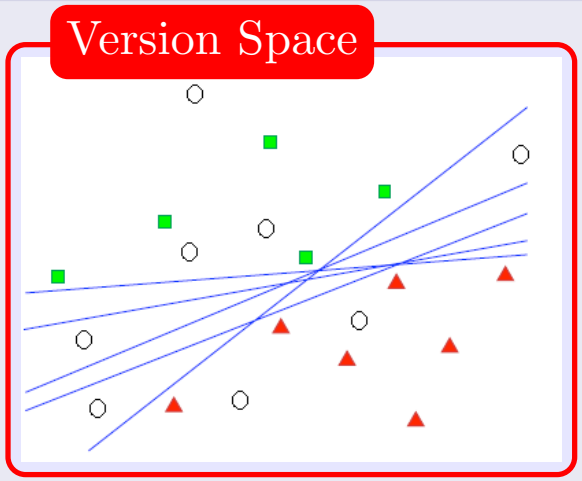
Sample Complexity

Number of labeled samples required to achieve a certain accuracy

Active Learning: Two Important Definitions

Version Space

Set of all classifiers who correctly classify labeled data

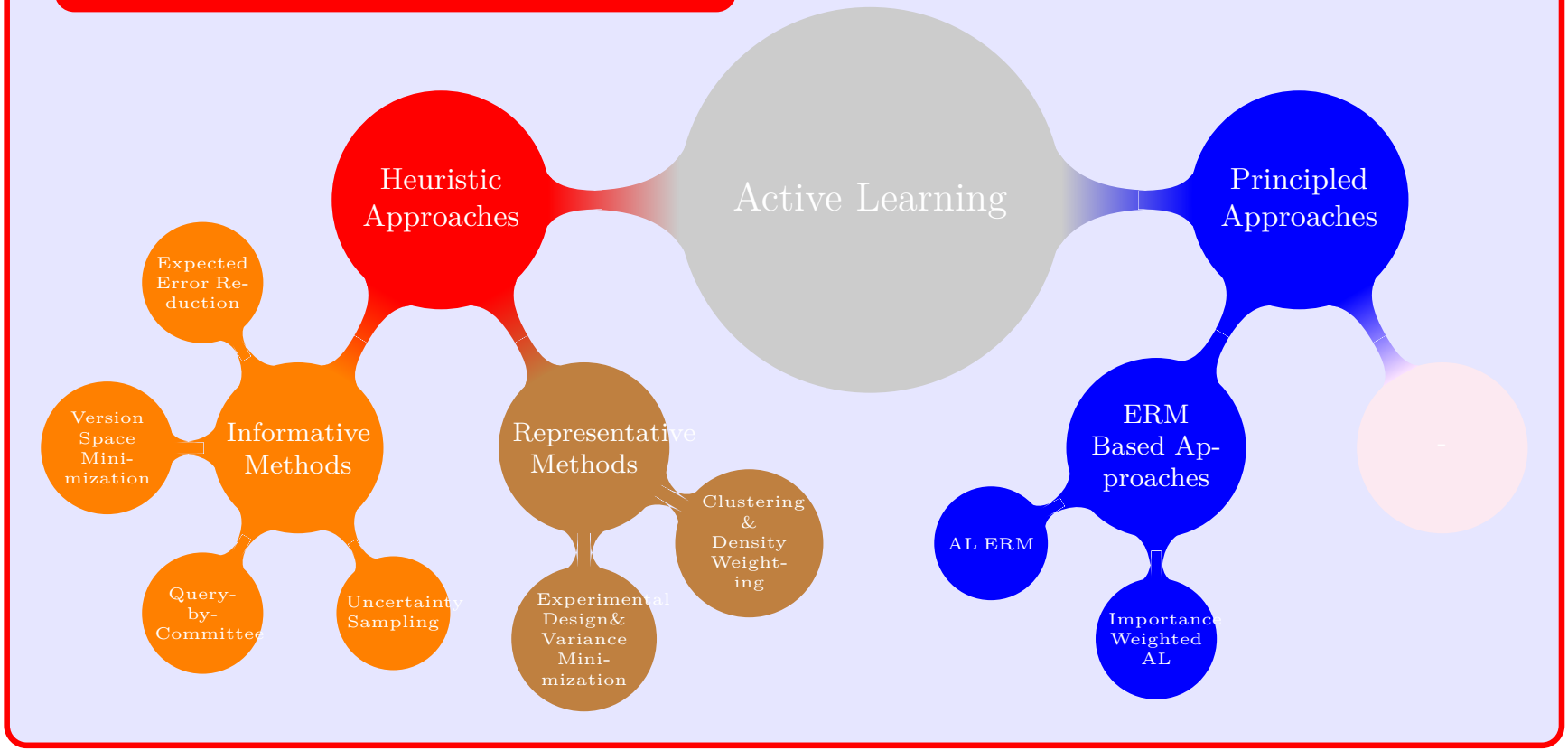


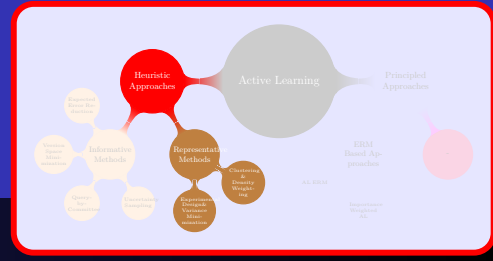
Sample Complexity

Number of labeled samples required to achieve a certain accuracy

Active Learning Approaches

Active Learning Taxonomy





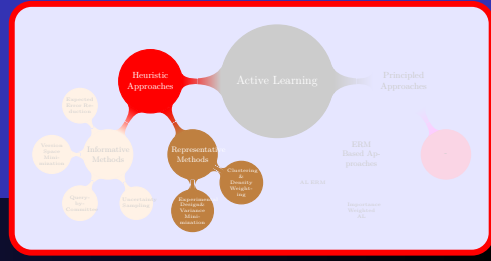
Representative Active Learning Methods

These group of methods assume that there exists a "structure" in the data which determines labels.

- Cluster assumption: A cluster in data, consists of only one label.
- Low Density assumption: In areas where there are points from different classes, density is low.
- Manifold assumption: There exist some manifolds in space, where in those manifolds, points of different classes can be easily separated.

Data Production Process

In the representative methods, the concentration is on the "process that produces the data".



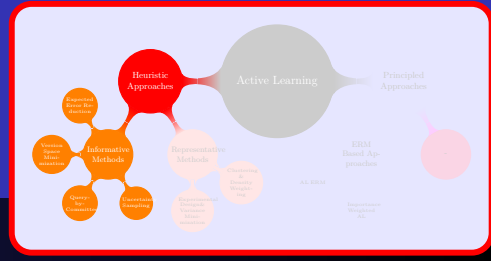
Representative Active Learning Methods

- Cluster-based Active learning Approaches
- Optimal Experiment Design, Variance Minimization Approaches

Fisher Information Based Methods

Disadvantages of Representative Active Learning Methods

- Often cluster, low-density, and manifold assumptions are not valid in real-world data.
- Representative Methods don't pay attention to "Label production process".
- They don't consider classifier risk into account.



Informative Active Learning methods

This approach concentrates on the effect of selecting an example for query on classifier and its uncertainty.

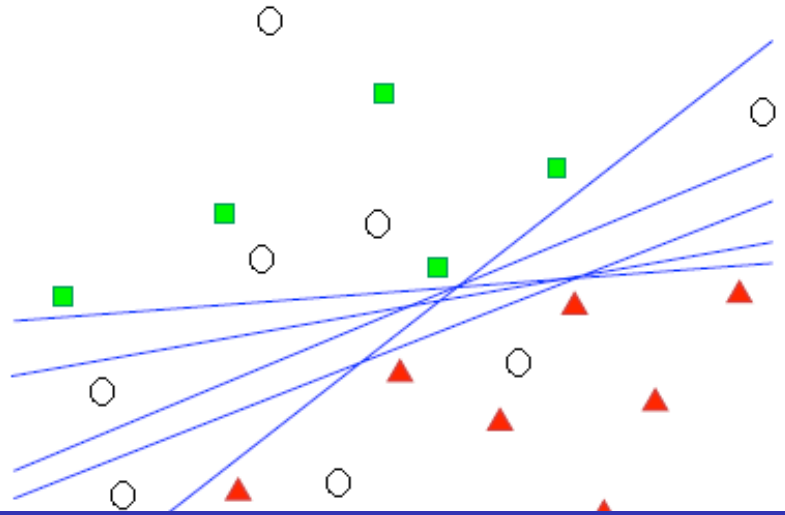
- Committee-based active learning
  Query examples with maximum disagreement between a set of classifiers.
- Mimizing version space for support vector machine.

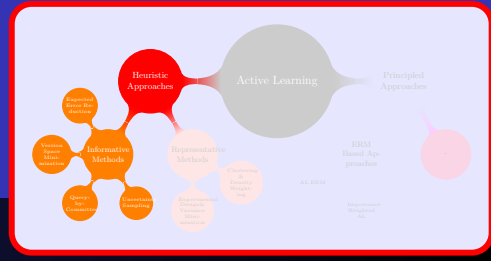
$$x^* = arg\min_x |f(x)|$$

$$x^* = arg\min_x \max_y l(y, f(x))$$

$$x^* = arg \min_{x_q} \min_f \max_{y_q} \frac{\lambda}{2}\|f\|^2 + \sum_{x_i} l(y_i, f(x_i)) \qquad (2)$$

$$= arg \min_{x_q} \min_f \frac{\lambda}{2}\|f\|^2 + |f(x_q)|$$





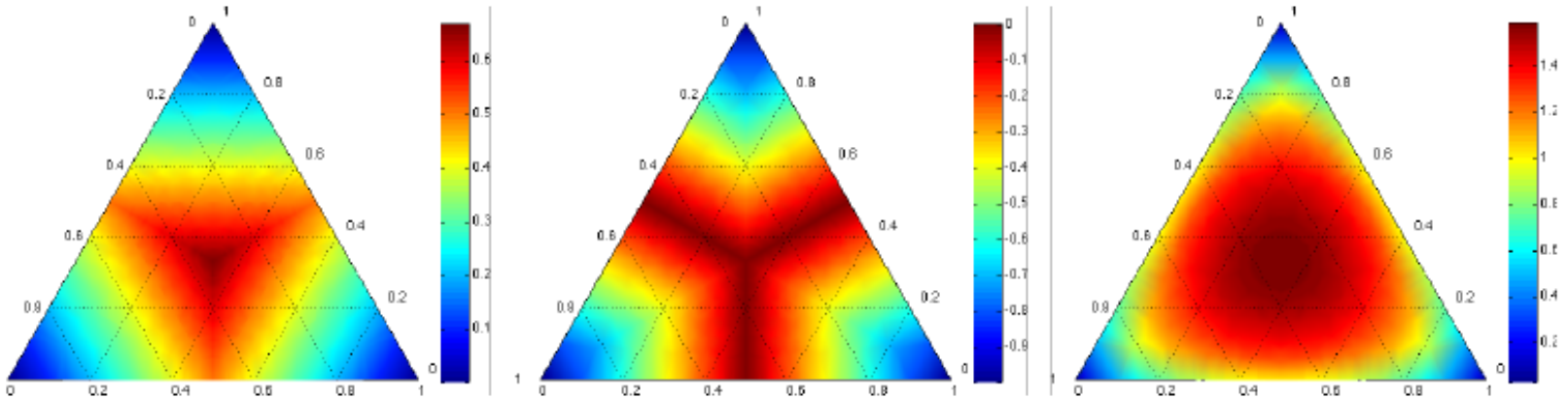
Informative Active Learning methods

- Uncertainty Sampling: Query examples with the maximum uncertainty

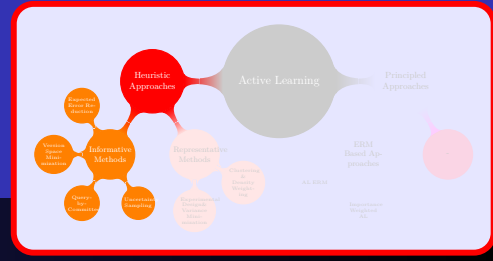
$$x_{LC} = arg\max_x 1 - P(\widehat{y}|x)$$

$$x_M = arg\max_x P(\widehat{y}_1|x) - P(\widehat{y}_2|x)$$

$$x_H = arg\max_x -\sum_i P(y_i|x) log P(y_i|x)$$



(a) least confident (b) margin (c) entropy



Informative Active Learning methods

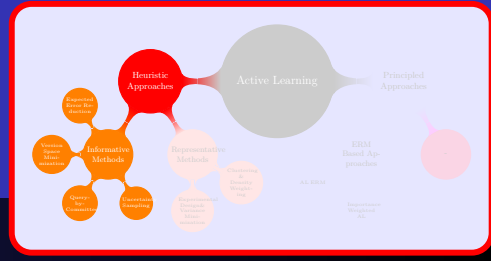
- Exprected Error Minimization: Selecting example with maximum amount of reduction in classifier error.
  - Estimation of Expected Error on labeled plus query data.

$$L = E_{X,Y}[l(g(x), y)] = E_{Y|X} E_X[l(g(x), y)]$$

  - Selecting example which minimizes the error
  - Estimation of classifier error on 0-1 loss with sum of uncertainty on classifier on unlabeled data.

$$x^* = arg\min_x \sum_i P_\theta(y_i|x) \Big(\sum_{u=1}^{U} 1 - P_{\theta^{+<x,y_i>}}(\hat{y}|x_u) \Big)$$

in which $\hat{y} = arg\max_y P_{\theta^{+<x,y_i>}}(y|x_u)$
In 2016, it has been shown that the above estimation is not correct.



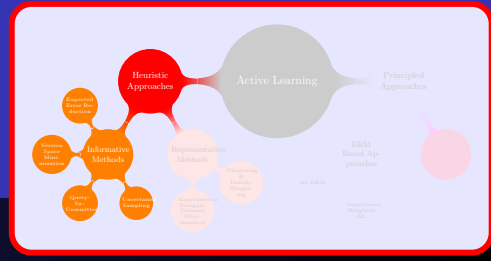
Informative Active Learning methods

Label production process

In informative methods, concentration is on **label production process**.

Disadvantages of Informative methods

1. Not paying enough attention to data structure.



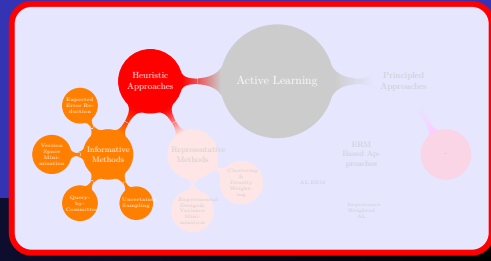
Informative Active Learning methods

Label production process

In informative methods, concentration is on **label production process**.

Disadvantages of Informative methods

1. Not paying enough attention to data structure.
2. Often becomes bias to labeled data.



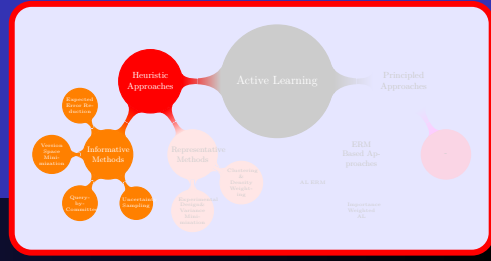
Informative Active Learning methods

Label production process

In informative methods, concentration is on **label production process**.

Disadvantages of Informative methods

1. Not paying enough attention to data structure.
2. Often becomes bias to labeled data.
3. No attention to risk of the classifier.



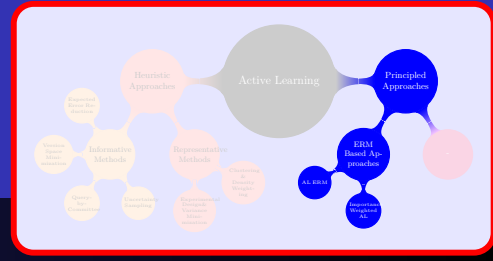
Informative Active Learning methods

Label production process

In informative methods, concentration is on **label production process**.

Disadvantages of Informative methods

1. Not paying enough attention to data structure.
2. Often becomes bias to labeled data.
3. No attention to risk of the classifier.
4. Although in active learning, we can't do cross validation easily, most of these approaches have a **number of parameters**.

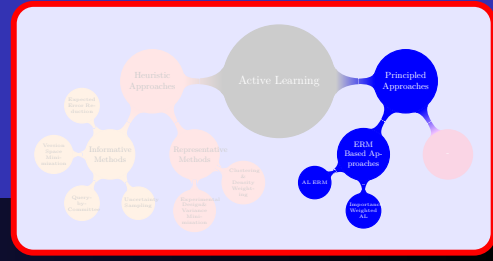


Active Learning: Principled Approaches

- In active learning, data is not **independent and identically distributed** (iid).
- In machine learning, almost always all algorithms are based on iid assumption.
- Learning algorithms becomes bias to labeled data.
- To eliminate bias in active learning, we must consider risk of the classifier in selection of examples for query.

Sampling Bias

No i.i.d selection of examples, creates **Sampling Bias**, therefore, training based on **Empirical Risk Minimization (ERM)** on this data, will not result in a good classifier for *all* of data. This further creates problems for selection of next examples in active learning process.

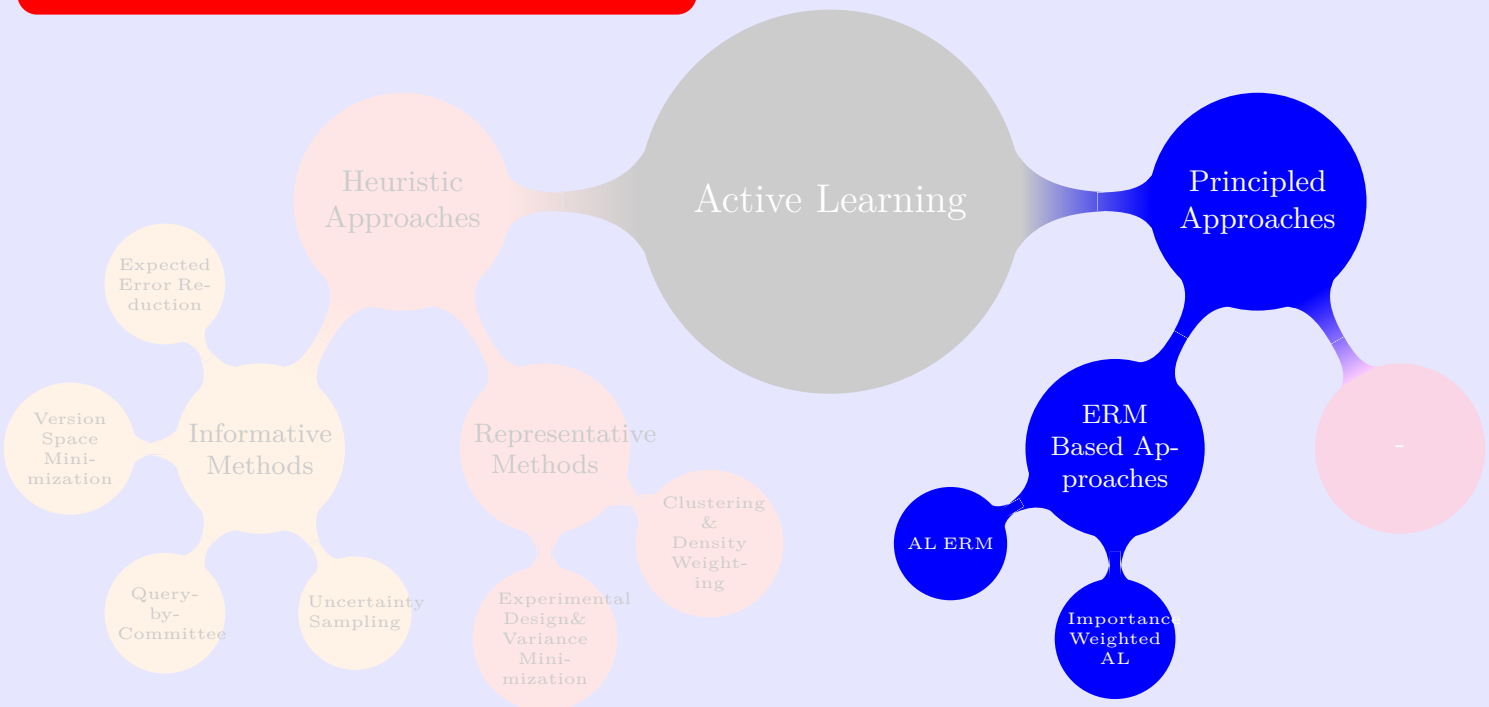


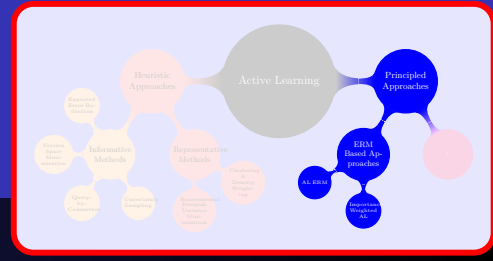
Active learning: Facing Sampling Bias

Two principled approaches that consider sampling bias:

- Empirical Risk Minimization for Active Learning (ERM AL)
- Importance Weighted Active Learning (IWAL)

Active Learning Taxonomy





Active learning: Principles approaches

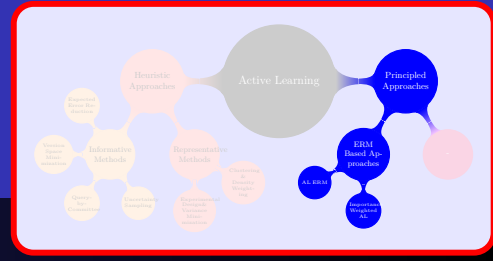
- ERM AL: Active learning based on Empirical Risk Minimization
  With probability $1-\delta$, we have

$$\mathbb{E}_{\mathcal{D}}[l(f(x),y)] \leq \hat{\mathbb{E}}_{Q}[l(f(x),y)] + MMD[\mathcal{C},p(x),q(x)] + 2R_q(\mathcal{L}) + \sqrt{\frac{ln(\frac{1}{\delta})}{q}},$$

where

$$R_n(\mathcal{L}) = \mathbb{E}_S\Big[\mathbb{E}_\sigma\big[\sup_{l\in\mathcal{L}}\big(\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(x_i,y_i)\big)\big]\Big]$$

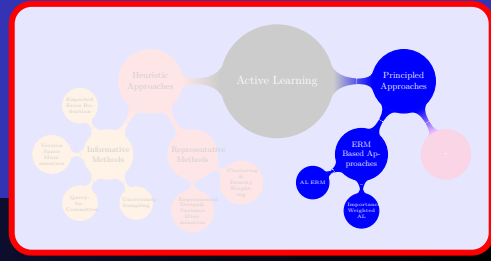
$$MMD(\mathcal{C},p(x),q(x)) = \sup_{g\in\mathcal{C}}\Big\{\mathbb{E}_{x\sim p(x)}[g(x)] - \mathbb{E}_{x\sim q(x)}[g(x)]\Big\}$$



Active learning: Principles approaches

- Importance Weighted Active Learning:
  Correcting bias with weighting examples
  weight example t with p_t:
  $p_t = \max_{f,g \in \mathcal{H}_t, y \in Y} l(f(x), y) - l(g(x), y)$
  $\mathcal{H}_t$ t-th Hypothesis class $Q_i \in \{0, 1\}$ and $\mathbb{E}[Q_i] = p_i$.

$$L(h) = \mathbb{E}[l(h(x), y)], \quad L_n(h) = \frac{1}{n} \sum_{i=1}^{n} \frac{Q_i}{p_i} l(h(x_i), y_i) \qquad (4)$$



Active learning: Principles approaches

- Importance Weighted Active Learning:

Theorem

For each distribution P and any classes of finite functions H, and for any $\delta > 0$ if constant $p_{min} > 0$ exists that $p_i \geq p_{min}, \forall i$ then we have

$$\mathbf{P}\Big[\max_{h\in\mathcal{H}} |L_n(h) - L(h)| > \frac{\sqrt{2}}{p_{min}}\sqrt{\frac{ln|H| + ln\frac{2}{\delta}}{n}}\Big] < \delta \qquad (5)$$

Definitions

Definition

Bayes Decision function and Bayes Risk
Let $\eta^*(x, y)$ be (unknown) distribution of labels given examples, that is

$$\eta^*(x, y) = \mathbf{P}(Y = y | X = x).$$

The optimal classifier (Optimal Bayes decision function) $g^*(x)$ and Bayes risk, respectively are

$$g^*(x) = arg \max_{y \in \mathcal{C}} \{\eta^*(x, y)\}, \qquad L^* = \mathbf{P}(g^*(X) \neq Y) \tag{6}$$

Probabilistic Minimax Active Learning (PMAL): Definition

Definition

Plug-in Classifier and its risk

Let

$$\eta(x, y; \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{D}}) \equiv \eta^{\mathcal{D}}(x, y) := \mathbf{P}\{Y = y | \mathcal{D}, X = x\}, \qquad (7)$$

classifier function $g^{\mathcal{D}}(x)$, plug-in classifier, is defined:

$$g^{\mathcal{D}}(x) = arg\max_{y \in \mathcal{C}} \{\eta(x, y; \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{D}})\}$$

Let y be label of x, risk of this classifier is

$$L(g^{\mathcal{D}}) = \mathbf{P}\{g^{\mathcal{D}}(X) \neq Y | \mathcal{D}\}$$

Plug-in vs Empirical Risk Minimization (ERM) classifiers

Previously, researchers thought that it is not possible to have plug-in classifier with the learning rate of $\mathcal{O}(\frac{1}{n})$.
Tsybakov proved that we can have a plug-in classifier with this rate of learning.

Back

If label assignment process doesn't have any constraint, active learning is not possible. For active learning, we must relate the data production process to the label assignment process in order to use the existence of unlabeled data.
Tsybakov used the following assumptions which is correct for various estimators and various distributions:

Remark

Let $\eta^{\mathcal{D}}$ be an estimator of η^* using the data $\mathcal{D}$ and $\mathcal{P}$ is class of distributions on (X, Y) such that with constants $C_1 > 0, C_2 > 0$ and a positive sequence like a_n, we have for all x with respect to approximately:

$$\sup_{P \in \mathcal{P}} P\{\mathcal{A}(x, \mathcal{D}) \geq \delta\} \leq C_1 exp(-C_2 a_n \delta^2)$$

where, $\mathcal{A}(x, \mathcal{D}) := \max_{y \in \mathcal{C}}\{\eta^*(x, y) - \eta(x, y; \mathbf{x}_{\mathcal{D}}, \mathbf{y}_{\mathcal{D}})\}$

Probabilistic Minimax Active Learning (PMAL)

Let $\mathcal{A}(x, \mathcal{D}) := \max_{y \in \mathcal{C}} \{\eta^*(x, y) - \eta(x, y; \mathbf{x}_\mathcal{D}, \mathbf{y}_\mathcal{D})\}$

Lemma (Distance of regression function to optimal regression by adding unlabeled data)

Assume that unlabeled data helps in reducing risk of classifier using Tsybakov's assumption (33) on $\mathcal{P}$, then $\exists c_\mathcal{U} > 1$ such that with probability $1 - C_1 exp(-C_2 a_n \delta^2)$, we have

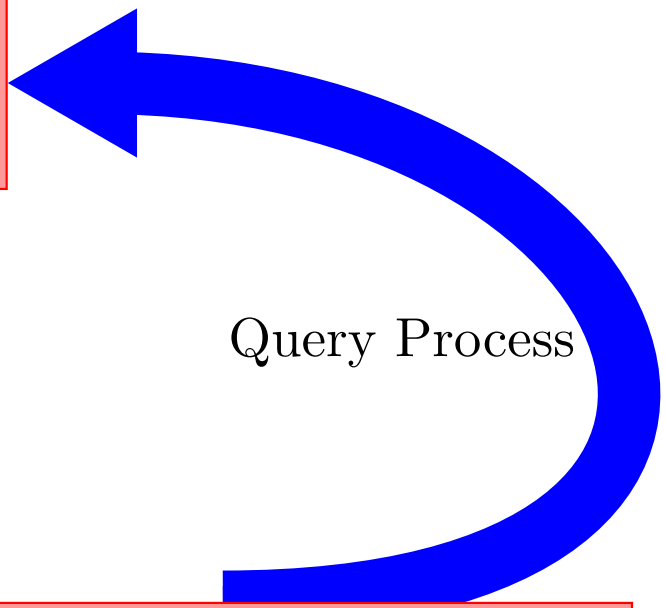
$$\int_{\mathbb{R}^d} \mathcal{A}(x, \mathcal{D}) \mu(dx) \leq c_\mathcal{U}^{-1} \int_{\mathbb{R}^d} \mathcal{A}(x, \mathcal{L}) \mu(dx) + 2\delta$$

Details

How does risk change if

we add a set Q of label data while we don't know their labels?

Risk for data & query:
$L(g^{\mathcal{D}\cup\mathcal{Q}})$



Risk for data:
$L(g^{\mathcal{D}})$

Upper bound of the risk of plug-in classifier

Theorem (Upper bound of the Excess Risk for multi-class plug-in classifier)

Let $\mathcal{T}^{\mathcal{D}}(x,y) = \eta^*(x,y) - \eta^{\mathcal{D}}(x,y)$. *Then we have*

$$L(g^{\mathcal{D}}) - L^* \leq \sum_{i=1}^{|\mathcal{C}|} E|\mathcal{T}^{\mathcal{D}}(X, c_i)| \leq |\mathcal{C}| \max_{y \in \mathcal{C}} E|\mathcal{T}^{\mathcal{D}}(X,y)| \qquad (8)$$

Risk of classifier without *knowing label's of examples*

Theorem (Upper bound of the risk of classifier after adding labeled data without knowing their labels)

Let assumptions of lemma 4 holds, and let $E_\eta = \mathbf{E}(\max_{y\in\mathcal{C}} \eta(X, y; \mathbf{x}_\mathcal{L}, \mathbf{y}_\mathcal{L}))$,

$$\Delta_\mathcal{L} \equiv \zeta \int_{\mathbb{R}^d} \max_{y\in\mathcal{C}} \{\eta^*(x, y) - \eta(x, y; \mathbf{x}_\mathcal{L}, \mathbf{y}_\mathcal{L})\}\mu(dx)$$

, and $\beta = 1 - c_\mathcal{U}^{-1}$. *Then with probability* $1 - C_1 exp(-C_2 a_n \delta^2)$ *upper bound of the risk of classifier trained using labeled data and query data* $\mathcal{Q}$, $L(g^{\mathcal{L}\mathcal{Q}})$, *is*

$$L(g^{\mathcal{L}\mathcal{Q}}) - L^* \leq \beta\Delta_\mathcal{L} + 2\zeta\delta + \zeta\alpha\big(1 - c_0 \min_{\mathbf{y}_\mathcal{Q}} \max_{\mathbf{y}_\mathcal{U}} \mathbf{P}(\mathbf{y}|\mathbf{x})\big) E_\eta$$

where $\alpha > 0$, $c_0 > 1, \zeta \geq 2$ *are constants independent of data.*

Probabilistic Minimax Active Learning

How to select examples to query their label that causes the most improvement of classifier accuracy?

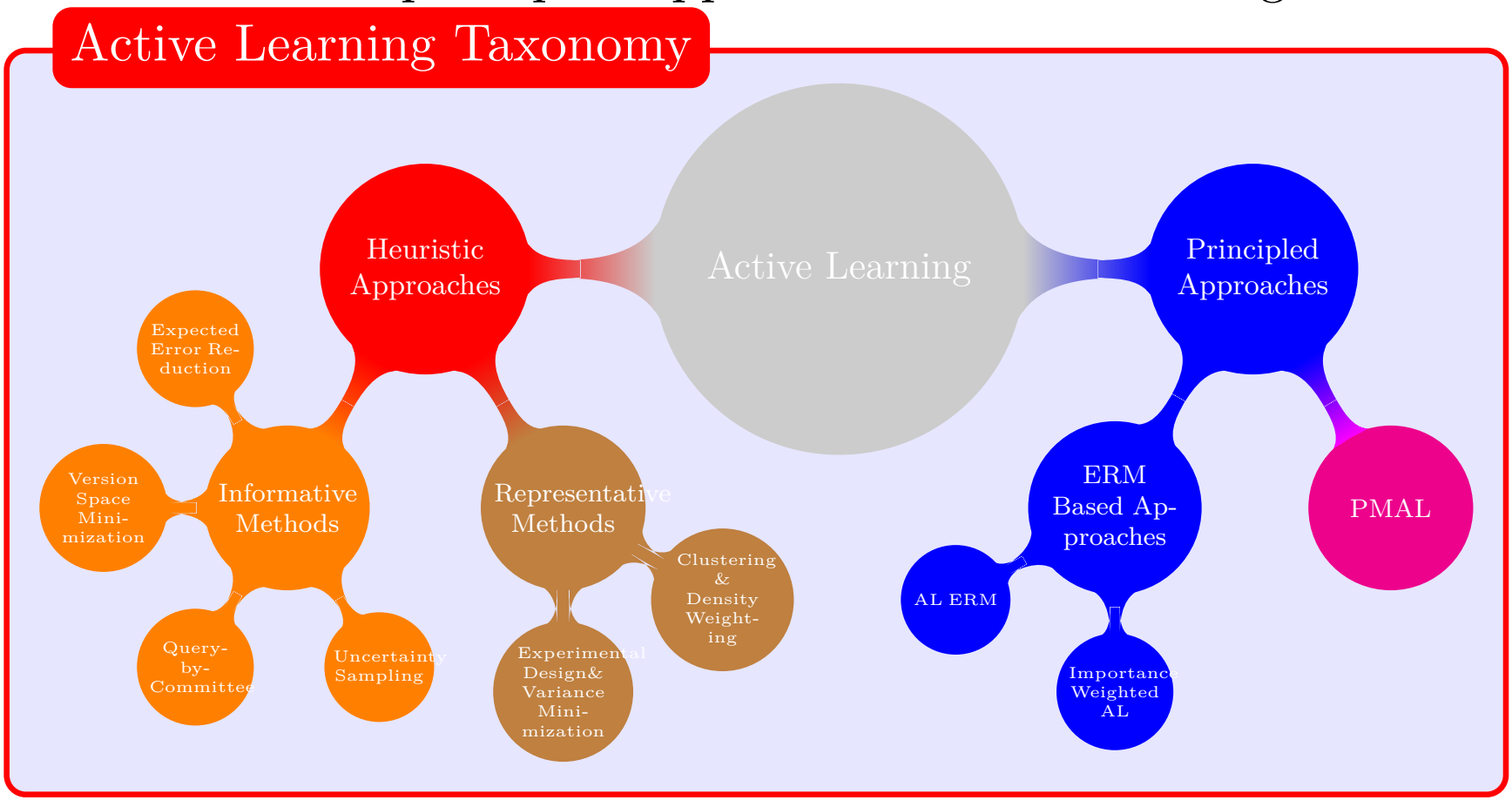
Corollary

Based on theorem 37, the best query which minimizes the risk of classifier is

$$\mathcal{Q} = arg \max_{\{\mathcal{Q},|\mathcal{Q}|=b\}} \min_{\mathbf{y}_{\mathcal{Q}}} \max_{\mathbf{y}_{\mathcal{U}}} \mathbf{P}(\mathbf{y}|\mathbf{x}) \tag{9}$$

Active learning Taxonomy

PMAL: A principled approach for active learning

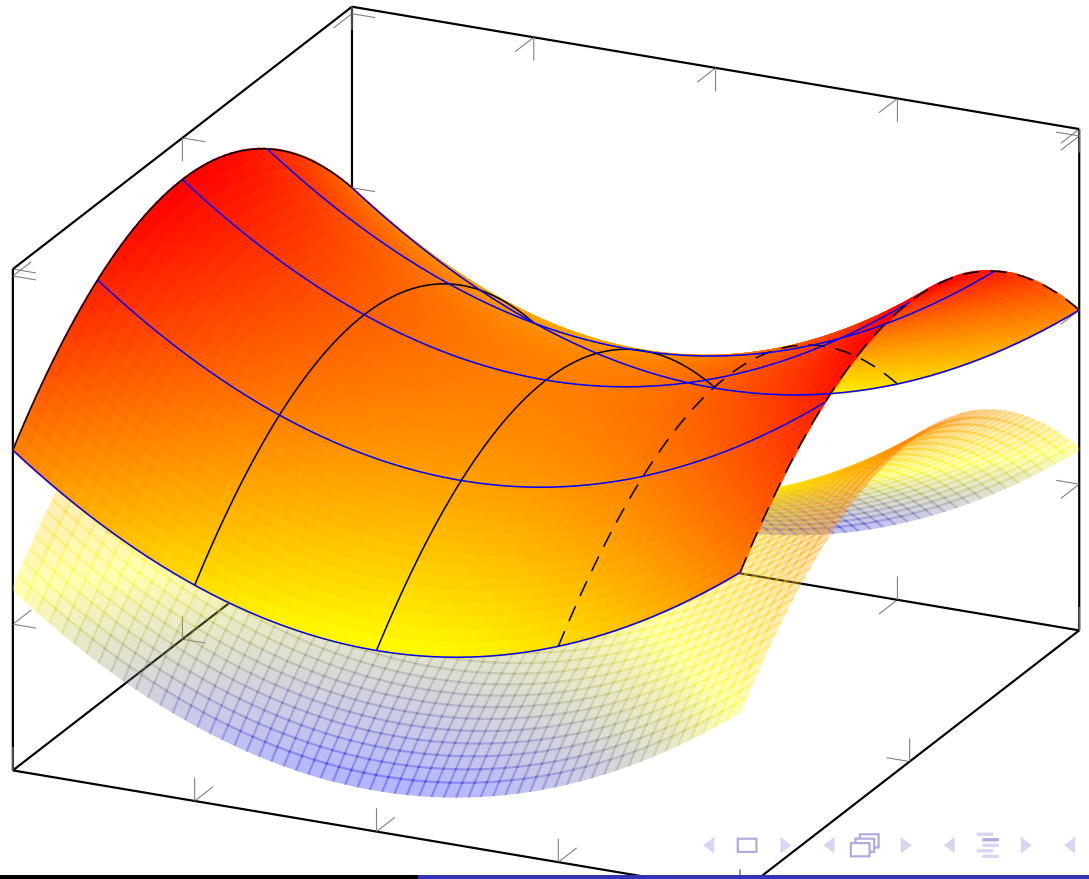


Interpretation

Based on min-max theorem, we have:

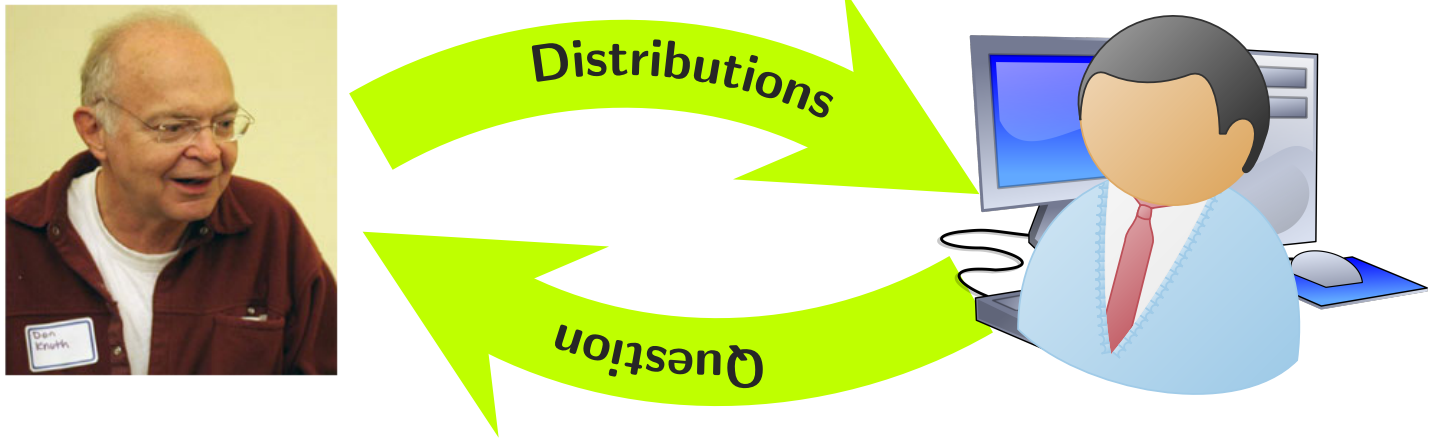
$$\forall \mathbf{y}_{\mathcal{U}}, \quad \min_{\mathbf{y}_{\mathcal{Q}}} \mathbf{P}(\mathbf{y}|\mathbf{x}) \leq \min_{\mathbf{y}_{\mathcal{Q}}} \max_{\mathbf{y}_{\mathcal{U}}} \mathbf{P}(\mathbf{y}|\mathbf{x}) \leq \max_{\mathbf{y}_{\mathcal{U}}} \mathbf{P}(\mathbf{y}|\mathbf{x}), \forall \mathbf{y}_{\mathcal{Q}} \quad (10)$$

PMAL selects the query set, such that the probability of the most wrong label for them in the current model be



How should a student ask questions? Learning theory answers:

$$\mathcal{Q} = arg\max_{\{\mathcal{Q},|\mathcal{Q}|=b\}} \min_{\mathbf{y}_{\mathcal{Q}}} \max_{\mathbf{y}_{\mathcal{U}}} \mathbf{P}(\mathbf{y}|\mathbf{x})$$



How should we ask questions such that our knowledge doesn't biases to the current available information?

- Assuming the correct answer of the model to all other questions, we must ask questions with the maximum probability of the most wrong answer.

Functional Gradient Approach to Probabilistic Minimax Active Learning

- Assumption: $P(y_i|x_i, \mathbf{y}, \mathbf{x}, \Psi)$ in which Ψ is a function
- In each iteration m estimates Ψ_m from Ψ, then $\Psi_m = \Psi_0 + \sum_{t=1}^{m} g_t$
- Each $g_t, t \in [1, m]$ is a gradient step in functional space and is equal to

$$g_m^* = E_{x,y}[\bigtriangledown_\Psi logP(y|x, \mathbf{y}, \mathbf{x}, \Psi)|_{\Psi_{m-1}}] \tag{11}$$

- Functional gradient in data points are equal to

$$g_m(y_i, x_i) = \bigtriangledown_\Psi logP(y_i|x_i, \mathbf{y}, \mathbf{x}, \Psi)|_{\Psi_{m-1}} \tag{12}$$

- Let:

$$\{a_m^*, \beta_m\} = arg \min_{a,\beta} \max_{y_i, i \in \mathcal{D}_u} \sum_{i=1}^{n} [g_m(x_i, y_i) - \beta h(x_i; a)]^2 \tag{13}$$

Functional Gradient Approach to Probabilistic Minimax Active Learning

Algorithm 1 General Functional Gradient Probabilistic Active Learning

procedure GFGP Active Learning(K, y_l, Ψ_{m_0}, m_l)
  $m \leftarrow m_0 + 1$
  while $m \leq m_0 + m_l$ or $\|\Psi_m - \Psi_{m-1}\| > \delta$ **do**
    $g_m(y_i, x_i) = \bigtriangledown_\Psi logp(y_i|x_i, \mathbf{y}, \mathbf{x}, \Psi)|_{\Psi_{m-1}}, \forall i \in [1, n]$
    Compute a_m^*, β_m^* using (13) .
    $\Psi_m = \Psi_{m-1} + \beta_m^* h(., .; a_m^*)$
    $m \leftarrow m + 1$
  $q \leftarrow arg\max_q \min_{y_q} \max_{y_{\bar{u}}} \prod_{i=1}^{n} P(y_i|x_i, \mathbf{y}, \mathbf{x}, \Psi_m)$
  return $q, \Psi_m(x)$ ▷ Query instance x_q

PMAL for robust learning

Let

$$x_i^*|x_i \sim \mathcal{N}(0, \sigma_i^2), \quad y_i = x_i^{*\mathsf{T}} w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{15}$$

, we have:

$$\begin{aligned} \mathbf{P}(y_i|x_i, w) &= \int \mathbf{P}(y_i|w, x_i^*)\mathbf{P}(x_i^*|x_i)dx_i^* \\ &= \mathcal{N}(y_i; x_i^{\mathsf{T}} w, \sigma^2 + \sigma_i^2\|w\|^2) \end{aligned} \tag{16}$$

Furthermore, let $\sigma^i(w) = \sigma^2 + \sigma_i^2\|w\|^2$. To compute $\mathbf{P}(\mathbf{y}|\mathbf{x})$, we have:

$$\mathbf{P}(\mathbf{y}|\mathbf{x}) = c \int \prod_i \sigma^i(w)^{-\frac{1}{2}} exp(-\frac{1}{2}\sum_i \frac{(y_i - w^{\mathsf{T}} x_i)^2}{\sigma^i(w)})\mathbf{P}(w)dw \tag{17}$$

Unfortunately, this integral is intractable. Moreover, $\mathbf{y}_{\mathcal{U}}$ is unknown

$\mathbf{y}_{\mathcal{U}}$?

Since $\mathbf{y}_{\mathcal{U}}$ is unknown, we cannot estimate this integral numerically.
Consider the following distribution:

$$\tilde{\mathbf{P}}(\mathbf{y}|\mathbf{x}, w) = \prod_{i=1}^{n} \mathcal{N}(y_i; w^{\mathsf{T}} x_i, \sigma^2 + \sigma_i^2 v^2) \tag{18}$$

where v is unknown. Using $\tilde{\mathbf{P}}$ instead of $\mathbf{P}$ solves the challenge of intractability of the integral, but how close these are to $\mathbf{y}_{\mathcal{U}}$?

Analytic solution for objective function

Theorem

Let $w^* = arg\max_w \mathbf{P}(\mathbf{y}, w|\mathbf{x}), \lambda = \frac{1}{\max_i\{\sigma_i^2\}}$, *and*

$$h(w;\mathbf{y},\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n}\{\frac{(y_i - w^\mathsf{T}x_i)^2}{\sigma^i(w)} + log\sigma^i(w)\} + \frac{1}{2\sigma_w^2}\|w\|^2$$

$$g(w,v) = \frac{1}{2}\sum_{i=1}^{n}\{\frac{\sigma_i^2(\|w\|^2 - v^2)(y_i - w^\mathsf{T}x_i)^2}{\sigma^i(v)\sigma^i(w)} + log\frac{\sigma^i(v)}{\sigma^i(w)}\}$$

then we have:

$$KL(\mathbf{P}(\mathbf{y},w|\mathbf{x})||\tilde{\mathbf{P}}(\mathbf{y},w|\mathbf{x})) = c_\lambda\Big\{D(\mathbf{P},\tilde{\mathbf{P}};\mathbf{y},\mathbf{x},\lambda,v) + \epsilon(\lambda)\Big\}$$

where $D(\mathbf{P},\tilde{\mathbf{P}};\mathbf{y},\mathbf{x},\lambda,v) \propto \frac{e^{-h(w^*;\mathbf{y},\mathbf{x}))},g(w^*,v)}{\sqrt{detD^2h(w^*;\mathbf{y},\mathbf{x})}}$, *also* , $c_\lambda = (\frac{2\pi}{\lambda})^{d/2}$, $\epsilon(\lambda) = e^{-h(w^*;\mathbf{y},\mathbf{x}))}\frac{O(1)}{\sqrt{\lambda}}$ *where* $\lambda \to \infty$.

A non-convex objective function

Let $v = \|w^*\|$, $g(w^*, v)$, and $KL(\mathbf{P}, \tilde{\mathbf{P}})$ is nearly zero.
The problem is that estimating w^* is hard specially since h^* is non-convex.
The following lemma proposes an upper bound for h:

Lemma

Let $\sigma_{w*}^{-2} = \sigma^2(\frac{1}{\sigma_w^2} + \sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma^2})$ *and*

$$h'(w; \mathbf{y}, \mathbf{x}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - w^\mathsf{T} x_i)^2}{\sigma^2} + \frac{1}{2} \frac{\sigma_{w^*}^{-2}}{\sigma^2} \|w\|^2 \qquad (19)$$

then we have $h(w; \mathbf{y}, \mathbf{x}) \leq h'(w; \mathbf{y}, \mathbf{x}) + nlog\sigma$

How different are the answers to these two problems? I

Therefore, h' can be used for minimizing h.
Let

$$D_{ij} = \sigma^{-2}(x_j x_j^{\mathsf{T}} + \sigma_j^2 I)(x_i y_i)$$

$$N_{\mathbf{x}} = \frac{1}{\sigma_w^2 \sigma^2} \sum_{i=1}^{n} \|x_i y_i\| \frac{\sigma_i^2}{\sigma^i(w^*)} \leq \frac{1}{\sigma_w^2 \sigma^4} \sum_{i=1}^{n} \|x_i y_i\| \sigma_i^2$$

How different are the answers to these two problems? II

Theorem

If w^* *be a regularized answer to the problem* $arg\min_w h(w; \mathbf{y}, \mathbf{x})$ *and we have* $\tilde{w} = arg\min_w h'(w, \mathbf{y}, \mathbf{x})$, *then:*

$$\frac{\|w^* - \tilde{w}\|}{\|w^*\|} \leq \frac{D_x + N_x}{\sigma_w^{-2}\sigma_{w*}^{-2}}\|w^*\| + \sigma_w^2\Big\{\sum_{i=1}^{n}\frac{\sigma_i^2}{\sigma^2}\Big\} \tag{20}$$

Moreover, $\lim_{\sigma_1,\ldots,\sigma_n\to 0}\|w^* - \tilde{w}\| = 0$. *Furthermore,* $\lim_{\|w^*\|\to\infty} D_{\mathbf{x}}\|w^*\|^2 = \lim_{\|w^*\|\to\infty} N_{\mathbf{x}} = 0$

$$D_{\mathbf{x}} = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\|D_{ij}\||\sigma_j^2 - \sigma_i^2|}{\sigma^i(w^*)\sigma^j(w^*)} \leq \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\|D_{ij}\|}{\sigma^4}|\sigma_j^2 - \sigma_i^2| \tag{21}$$

Active Robust Learning for input space

Theorem

Let

$$\tau = \mathbf{y}_{\mathcal{L}}^{\mathsf{T}}(L_{ll} - L_{lu}L_{uu}^{-1}L_{ul})\mathbf{y}_{\mathcal{L}}$$
$$\kappa_i = (\sigma^2 + \sigma_i^2\tau)^{-1}$$
$$\Upsilon = diag(\kappa)$$
$$\mathcal{S} = \Upsilon - \Upsilon\Big\{\sigma_w^2(K^{-1} + \sigma_w^2\Upsilon)^{-1}\Big\}\Upsilon,$$

we have
$\tilde{\mathbf{P}}(\mathbf{y}|\mathbf{x}) \propto exp(-\frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathcal{S}\mathbf{y})$

Active learning on distributions: Preliminaries I

We have a number of distributions, $P_i, i \in \{1, ..n\}$ as examples.
For each distribution P_i, we have a set of points $x_i = \{x_i^1, x_i^2, ..., x_i^{n_i}\}$ which have been sampled from distribution P_i. Each point in this set, is a vector, that is $x_i^u \in \mathbb{R}^d, \forall u \in [1, n_i]$.
Kernel embedding of P_i using a kernel function k is equal to

$$\mu^*_{P_i} := \int_{\mathcal{Z}} k(x, .) dP_i(z), \tag{22}$$

where kernel function k is finite and positive definite.
Empirical kernel embedding of P_i is

$$\widehat{\mu}_{P_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} k(x_i^j, .) \tag{23}$$

Active learning on distributions: Preliminaries I

Let $\mathcal{H}$, a Hilbert space for kernel function k, then $\mu(x)$ is the evaluation of μ at value x. Specifically, for a given x, we have

$$\widehat{\mu}_{P_i}(x) = \frac{1}{n_i} \sum_{j=1..n_i} k(x_i^j, x) \tag{24}$$

as the evaluation of empirical embedding of distribution P_i at the point x.
We need to compute $p(\mathbf{y}|\widehat{\mu}_{\mathbf{P}})$. For this purpose, we can estimate more exact estimate of this value using Bayesian estimation in a Hilbert space.
Let k_θ be a positive definite kernel function with parameter θ.
In order to be confident that the resulting Bayesian Kernel embedding will be in the Hilbert space associated with the

Active learning on distributions: Preliminaries II

kernel function k, that is $\mu^* \in \mathcal{H}_k$, we propose the prior probability on μ^* as

$$\mu^*|\theta \sim \mathcal{GP}(0, r_\theta(.,.)), \tag{25}$$

$$r_\theta(x,y) = \int k_\theta(x,u)k_\theta(u,y)du, \tag{26}$$

where $\mathcal{GP}$ is a Gaussian process. This selection of r_θ guarantees that $\mu^*_{P_i} \in \mathcal{H}$.

Active learning on distributions: Preliminaries III

We have:

$$\begin{aligned}
&\mathbf{P}(y_i|\widehat{\mu}_{P_i}(x_i),w) \\
&\qquad = \int \mathbf{P}(y_i|\mu^*_{P_i},w)\mathbf{P}(\mu^*_{P_i}|\widehat{\mu}_{P_i}(x_i))d\mu^*_{P_i} \\
&\qquad = \int \mathcal{N}(y_i;w^\mathsf{T}\mu^*_{P_i},\sigma^2)\mathcal{N}(\mu^*_{P_i};\widehat{\mu}_{P_i}(x_i),C_i)d\mu^*_{P_i} \\
&\qquad \propto (\sigma^i(w))^{-\frac{1}{2}} exp(-\frac{1}{2}\frac{(y_i - w^\mathsf{T}B_i\widehat{\mu}_{P_i}(x_i))^2}{\sigma^i(w)})
\end{aligned}$$

Active learning on distribution: Estimation of objective function

Lemma

Let $r_{xx} = r(x, x)$, $R^i = [r(x_i^t, x_i^v)]_{t,v=1}^n$, *and* $R_x^i = [r(x, x_i^1), r(x, x_i^2), ..., r(x, x_i^n)]^\mathsf{T}$. *Furthermore, let*

$$B_i = (R^i + (\rho^2/n_i)I_n)^{-1} R_x^i \tag{27}$$

$$C_i = r_{xx} - {R_x^i}^\mathsf{T} (R^i + (\rho^2/n_i)I_n)^{-1} R_x^i \tag{28}$$

$$\sigma_i(w) = \sigma^2 + w^\mathsf{T} C_i w. \tag{29}$$

Assuming a linear model $y_i = w^\mathsf{T}\mu_{P_i}^* + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$, *we have:*

$$\mathbf{P}(y_i|\widehat{\mu}_{P_i}(x_i), w) \propto \sigma_i(w)^{1/2} exp(-\frac{1}{2\sigma_i(w)}(y_i - w^\mathsf{T} B_i \widehat{\mu}_{P_i}(x_i))^2)$$

Estimation of Active learning Objective function I

For computing $\mathbf{P}(\mathbf{y}|\widehat{\mu}_{\mathbf{P}})$, we have:

$$\mathbf{P}(\mathbf{y}|\widehat{\mu}_{\mathbf{P}}(\mathbf{x})) = \int \mathbf{P}(\mathbf{y}|\widehat{\mu}_{\mathbf{P}}(\mathbf{x}), w)\mathbf{P}(w)dw. \tag{31}$$

This integral is intractable. We need to find an efficient approach for its estimation.
In a simpler situation, we assume that there is no uncertainty in estimating distributions and therefore, each distribution can be determined by their estimate in Hilbert space. When C_i is constant, with considering uncertainty for each distribution, P_i, $w^{\mathsf{T}}C_i w$ will not increase since this quantity itself is a measure of complexity for w. The following theorem, provides a lower bound for $w^{\mathsf{T}}C_i w$.

Estimation of Active learning Objective function II

Theorem (Estimating active learning objective on distributions)

Lower bound $\mathbf{P}(\mathbf{y}|\widehat{\mu}_{\mathbf{P}}(\mathbf{x}))$ *is proportional to*

$$\tilde{\mathbf{P}}(\mathbf{y}|\widehat{\mu}_{\mathbf{P}}(\mathbf{x})) \propto exp(-\frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathcal{S}\mathbf{y}) \tag{32}$$

$$\mathcal{S} = \Upsilon - \sigma_w^2 \Upsilon \Big\{ K_\mu^B - K_\mu^B \{\sigma_w^{-2}\Upsilon^{-1} + K_\mu^B\}^{-1} K_\mu^B \Big\} \tag{33}$$

Let $\mu_{R_i} = (R^i + (\tau^2/n_i)I_n)^{-1}\widehat{\mu}_{P_i}(x_i)$, *then* K_μ^B *is a matrix with elements*

$$K_\mu^B(i,j) = [\mu_{R_i}^{\mathsf{T}} R^{ij} \mu_{R_j}] \tag{34}$$

Estimation of Active learning Objective function III

Theorem (Estimating active learning objective on distributions)

$$\Upsilon = \Upsilon = diag(\kappa), \quad \kappa_i^{-1} = \sigma^2 + \tau_i, \tag{35}$$

where

$$L^t = (K_\mu^B + \sigma_w^{-2} I)^{-1} E^t (K_\mu^B + \sigma_w^{-2} I)^{-1} \tag{36}$$

$$E^t = K_\mu^B - P_t^\mathsf{T} (R^t + (\tau^2/n_t) I_n)^{-1} P_t \tag{37}$$

$$P_t = [R^{ti} \mu_{R_i}]_{i=1}^n \in \mathbb{R}^{n_t \times n}. \tag{38}$$

We also, have $\tau_i = \mathbf{y}_l^\mathsf{T} \{L_{ll}^i - {L_{ul}^i}^\mathsf{T} (L_{uu}^i)^{-1} L_{ul}^i\} \mathbf{y}_l$. *This lower bound is tight that is there exists label* $\mathbf{y}_\mathcal{U} \in [-1, 1]^{n_u}$, *where the value of* $\mathbf{P}(\mathbf{y}|\widehat{\mu}_\mathbf{P}(\mathbf{x}))$ *is equal to this lower bound.*

Conclusion

- Active learning is a fundamental problem in machine learning which is **irreducible** to other problems.
- Solving **optimal Bayesian active learning** is an important tool for active robust learning and active learning on distributions.
- Active learning on Distributions which we didn't cover in this short presentation is an important step toward learning on more abstract/structured forms of data.

Papers

In these slides, we concentrated on some parts of the following paper:

- **Prepare for the worst, hope for the best: Active Robust Learning on Distributions**,
  IEEE TRANSACTIONS ON CYBERNETICS, VOL. 52, NO. 6, JUNE 2022

Papers

Probabilistic Minimax approach to active learning and learning on distributions, further have been explored in the following:

- **Local variational Probabilistic Minimax Active Learning**,
  Elsevier, Expert Systems with Applications: An International Journal Volume 211 Issue C Jan 2023
- **Prepare for the worst, hope for the best: Active Robust Learning on Distributions**,
  IEEE TRANSACTIONS ON CYBERNETICS, VOL. 52, NO. 6, JUNE 2022
- **Functional Gradient Approach to Probabilistic Minimax Active Learning**,
  Elsevier, Engineering Applications of Artificial Intelligence
- **Identifying Crisis-related Informative Tweets Using Learning On Distributions**
  Elsevier, Information Processing and Management
- **Active Robust Learning**,
  Arxive Preprint, 2016

Questions

Representative Active learning algorithms

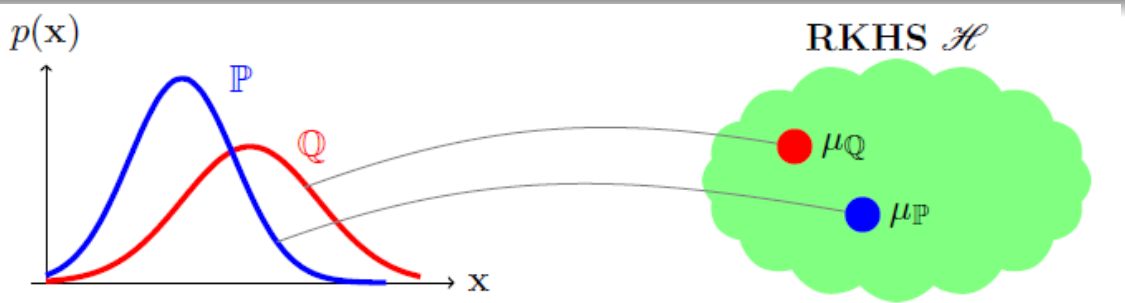
- Algorithms based on Minimization of variance and Optimal Experiment Design
  Minimizing a feature of the Fisher Information matrix:

$$\mathcal{I}(\theta) = N \int_x P(x) \int_y P_\theta(y|x) \frac{\partial^2}{\partial \theta^2} log P_\theta(y|x)$$

  - A-Optimal: Minimizes the trace of Fisher information matrix.
  - D-Optimal: Minimizes the determinant of Fisher information matrix.
  - E-Optimal: Minimizes the largest eigen value of Fisher information matrix.

BACK

Challenges of learning on distributions using a sample



- The mean embedding of a characteristics kernel preserves all features of a distribution.

$$\mu(\mathbb{P}) = \mathbb{E}_{x\sim\mathbb{P}}[k(.,x)] = \int_x k(.,x)d\mathbb{P}(x)$$

- Therefore,

$$\mu(\mathbb{P}) = \mu(\mathbb{Q}) \Rightarrow \mathbb{P} = \mathbb{Q}$$

- However, the size of sample is finite and any estimate of distribution is inexact.

BACK