Active Learning

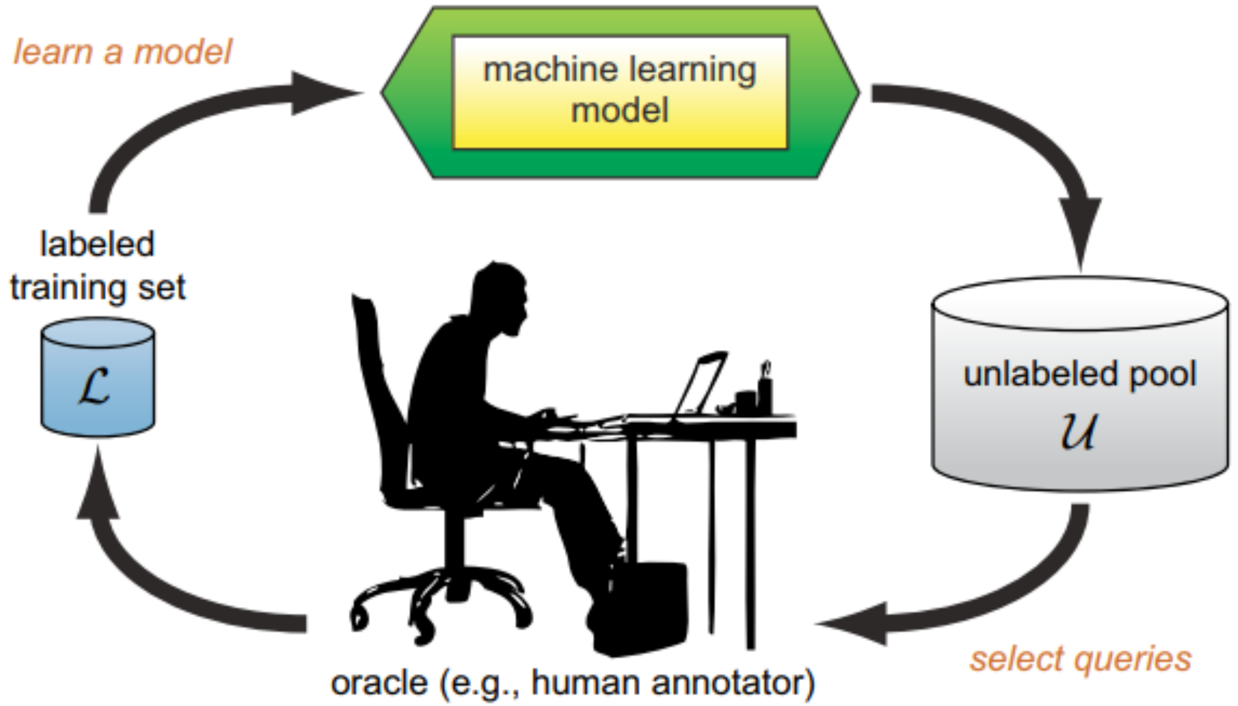
Topics in Trustworthy AI

Daksh Mittal

Spring 2025

Active Learning

- **Ground truth labels are costly:** Learn efficiently with less labels.
- Randomly querying X would be expensive.
- Adaptively choose which X to query next.

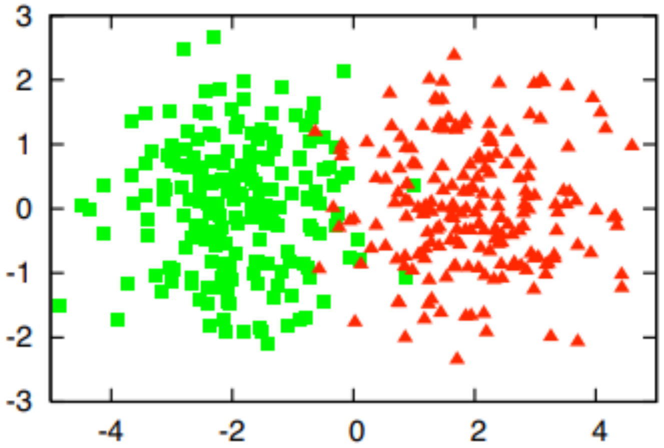


Efficient data collection for improving ML models

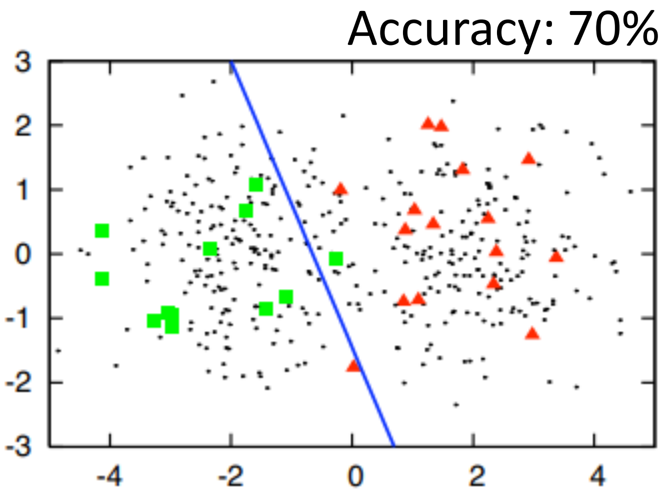
Active Learning: Decision problem under uncertainty

- Active Learning - Adaptive data collection for improving AI/ML models
  - ✓Uncertainty Quantification as a prerequisite
  - ✓How to select queries? Ideas closely linked to Multi-armed Bandits, and Bayesian optimization

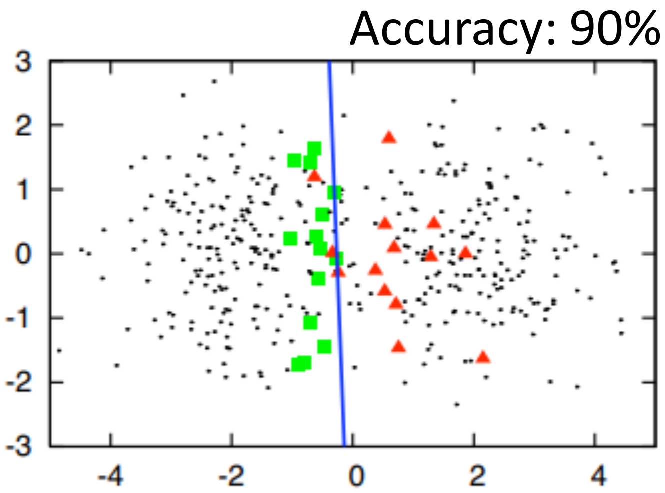
An Example of Active Learning



Dataset
(Two class Gaussians)



Randomly query
30 points



Actively query
30 points

Source: https://burrsettles.com/pub/settles.activelearning.pdf

Different Active Learning Scenarios

- **Query synthesis**
  - ✓ Generate a query (X) to be labeled
  - ✓ Potential issue - labeling arbitrary inputs
- **Stream-based selective sampling**
  - ✓ Sample an instance from input space – decide to query or discard
- **Pool-based sampling**
  - ✓ Sample a large pool of instances from input space – select the best query

Querying strategies

1. **Uncertainty based strategies:** Sample from the regions of the space with highest uncertainty.

2. **Representativeness-based strategies:** Query data that is representative of the underlying population – diversity and density based criteria.

3. **Performance-based strategies:** Sample to directly optimize the performance of the model.

Uncertainty based strategies

Uncertainty Sampling

1) **Least confident sampling:** Query the instance whose prediction is **least confident**. Let $\hat{Y} = \mathrm{argmax}_Y\, p_M(Y|X, D_{train})$,

$$X^* = \mathrm{argmax}_X \left(1 - p_M\left(\hat{Y}\middle|X, D_{train}\right)\right)$$

2) **Margin sampling:** Query the instance with **least difference** between probabilities of top two classes -

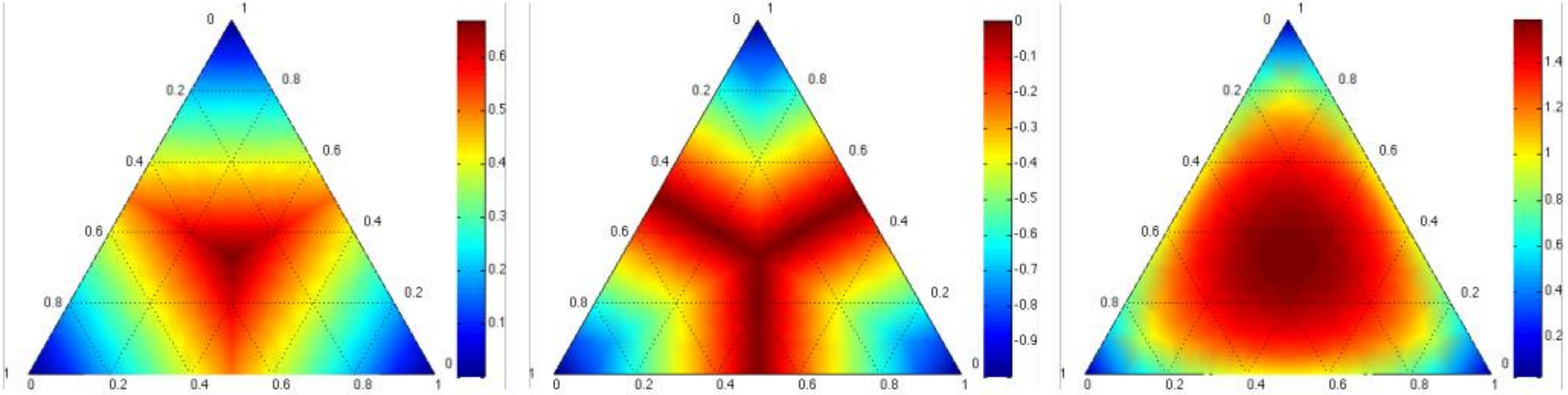
$$X^* = \mathrm{argmin}_X\left(p_M(\hat{Y}_1|X, D_{train}) - p_M(\hat{Y}_2|X, D_{train})\right)$$

3) **Shannon entropy:** Query the instance with **highest entropy**

$$X^* = \mathrm{argmax}_X\left(-\sum_c p_M\left(\hat{Y} = c\middle|X, D_{train}\right)\log\left(p_M\left(\hat{Y} = c\middle|X, D_{train}\right)\right)\right)$$

No differentiation between epistemic and aleatoric uncertainty

Visualizing uncertainty sampling



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

Source: https://burrsettles.com/pub/settles.activelearning.pdf

Query by committee

Maintain a committee $\boldsymbol{\mathcal{C}} = \{\theta_1, \theta_2, \cdots, \theta_N\}$ of models – represent competing hypotheses.

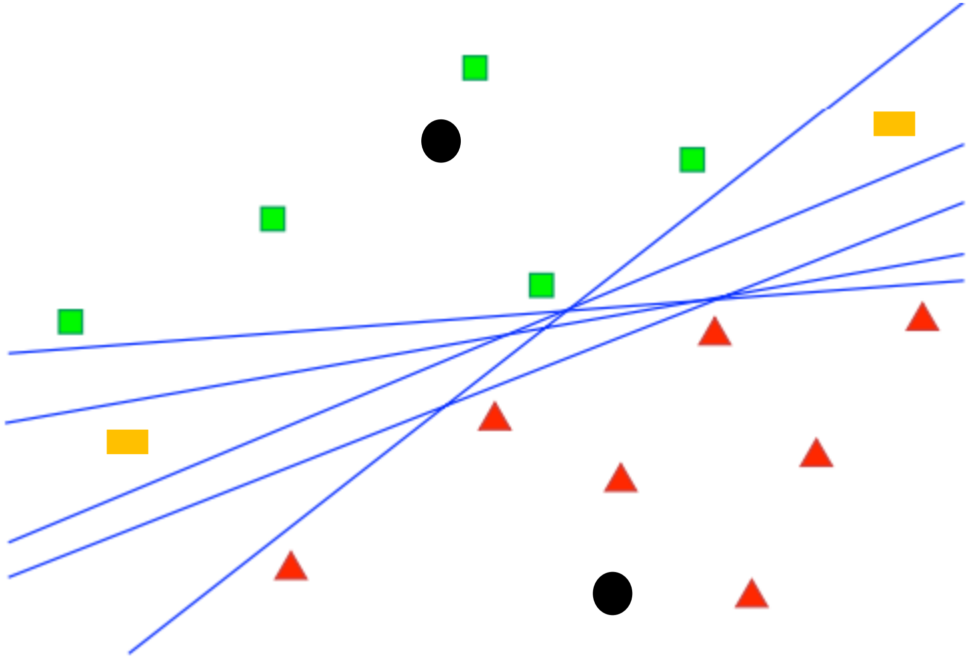
- ✓Most informative query - instance about which the models most disagree.
- ✓How to get different models? Ensembles, Explicit prior and posterior distributions of model parameters etc.
- ✓How to measure disagreement? Vote entropy, Average KL divergence

$$\alpha_{VE}(X) = -\sum_{c} \frac{V(c)}{N} \log\left(\frac{V(c)}{N}\right)$$

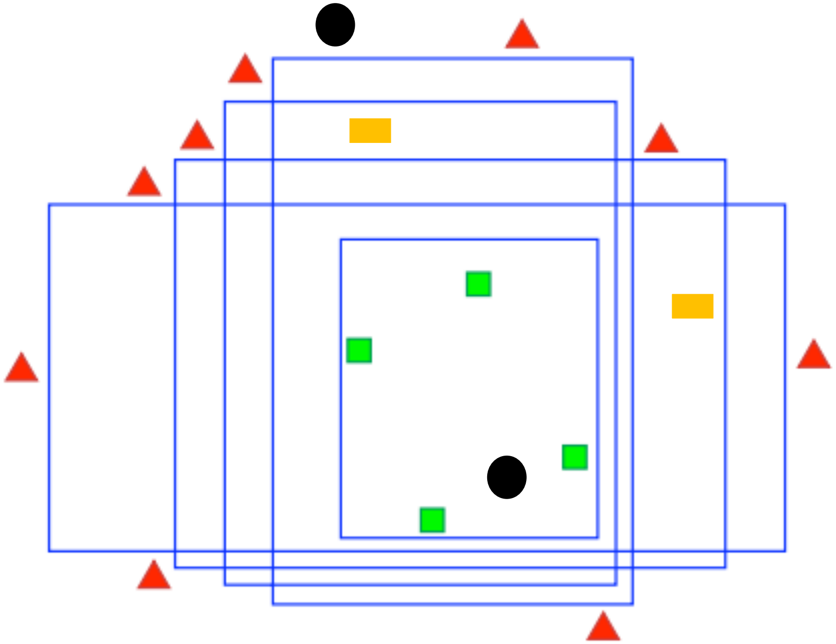
Trying to capture some notion of epistemic uncertainty!

Visualizing query by committee

Models agree
Models disagree
Class 1
Class 2



Linear model class



Axis parallel box model class

Bayesian Active Learning by Disagreement (BALD)

Maximize the mutual information between the model predictions and model parameters

$$\alpha_{BALD}(X) = I(Y, \theta | X, D_{train}) = \mathrm{H}(\mathrm{Y}|\mathrm{X}, D_{train}) - \mathbb{E}_{\theta\sim p(\theta|D_{train})}\left(H(Y|X, \theta)\right)$$

✓ **Term 1 :** X with high entropy in average output $\left[P(Y|X, D_{train}) = \mathbb{E}_{\theta\sim p(\theta|D_{train})}\left(P(Y|X, \theta)\right)\right]$

✓ **Term 2 :** Penalizes X, where many models are not confident – Inputs X on which models disagree.

✓ **Larger difference indicates :** Labeling X would provide more information about model's parameters.

Uncertainty based strategies – summary

- Query most uncertain input X
  - ✓ notion of uncertainty differs among methods
- **Potential issue:** Prone to querying outliers
  - ✓ Might not improve the performance of the model

Representativeness based strategies

- **Informative instances:** Uncertainty +“representative” of the underlying distribution

- **Density weighted methods:** average similarity to other instances

$$\text{(Uncertainty based metric)}\left(\frac{1}{U}\sum_{u\in[1,U]} sim(X, X^u)\right)^{\beta}$$

- **Other methods:** K-means clustering, Core set

  - ✓ **Main idea:** Least similar to labeled set + Most similar to unlabeled set

Performance-based strategies

- Expected error reduction:

Probability that label $Y = c$ for input X

$$Minimize_X \sum_c p(Y = c|X, \theta) \left(\sum_{u \in U} (1 - p_{\theta^{+(X,c)}}(\hat{Y}|X^u)) \right)$$

Error on the unlabeled pool U, under the updated model $[\theta^{+(X,c)}]$

- Expected log-loss:

$$Minimize_X \sum_c p(Y = c|X, \theta) \left(- \sum_{u \in U} \sum_{\hat{c}} p_{\theta^{+(X,c)}}(\hat{c}|X^u) \log \left(p_{\theta^{+(X,c)}}(\hat{c}|X^u)\right) \right)$$

Log loss on the unlabeled pool U, under the updated model $[\theta^{+(X,c)}]$

Other considerations!

- Batch mode Active learning
- Variable labeling costs
- Multi-task active learning – single query labeled for multiple tasks

Critical components and Limitations

- **Uncertainty Quantification**: Requires methodologies that effectively differentiate epistemic and aleatoric uncertainty

- **Explicitly optimize for the objective under consideration:** model improvement over target population

- Heuristics try to balance trade-offs between uncertainty, representation - **How to approach in a principled manner?**

- Mostly caters to classification tasks and suffers highly under batching**.**