Data Mining & Machine Learning

CS37300
Purdue University

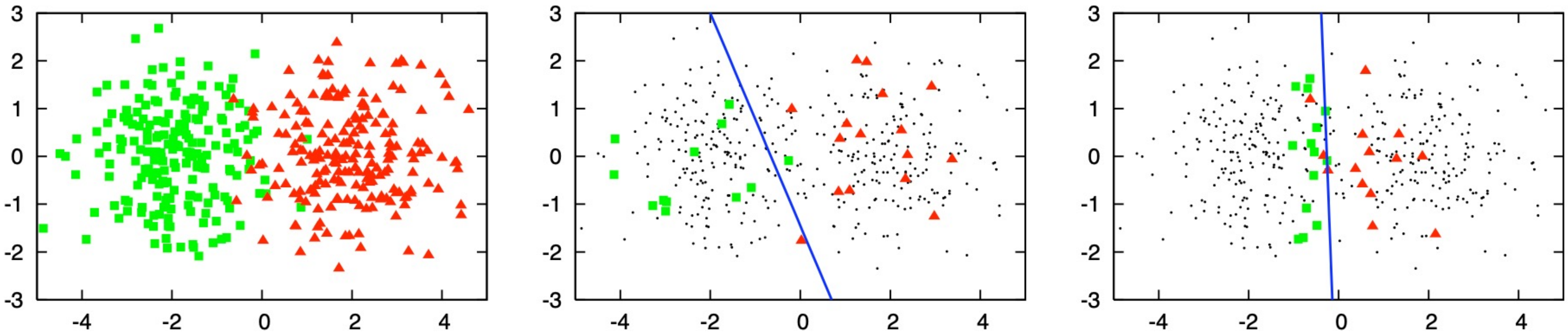
Oct 16, 2023

Active learning

- Setup: Limited data available -- x_i, *labelled* y_i
- Premise: Learner allowed to choose which data to learn from
  - Query for more data
  - Limited budget for getting new data
- Goal: Reach greater accuracy with few labelled data points

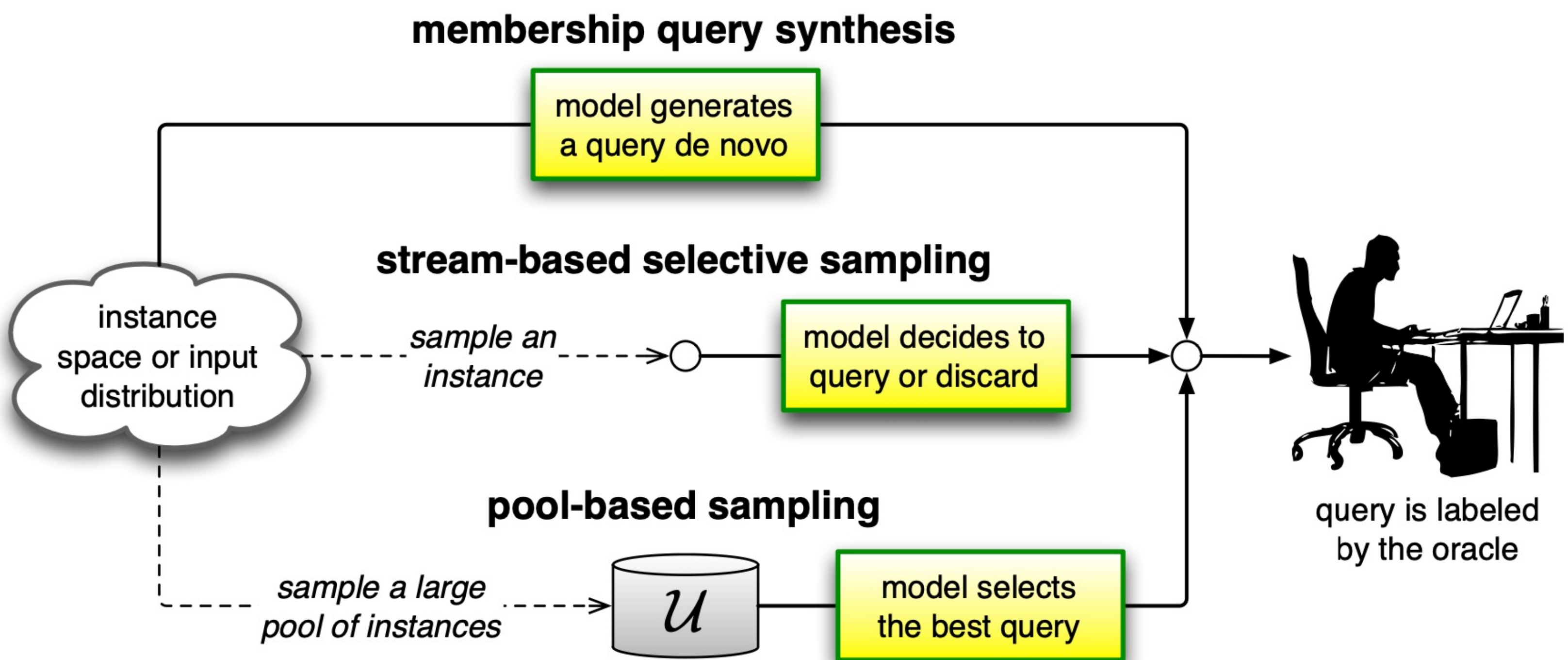
Example

- This lecture draws heavily from “Active Learning Literature Survery” by Burr Settles



(a) (b) (c)

Labelling models



Stream-based selective sampling

- New unlabelled data points drawn from the data distribution and presented to the model one-at-a-time in a stream
- The model decides to query for the label or discard
- Information-based measures are often used
- Applications: part-of-speech tagging, learning ranking functions, word sense disambiguation

Pool-based sampling

- Available: labelled data x_i, y_i and pool of unlabelled data points x_j
- Query from the pool for a suitable x_j
- Applications include image-classification, video-classification, medical diagnosis
- Both stream and pool based techniques assume some underlying distribution from which the data is drawn
- Model could generate a query “de novo” to form a new x_j

Query strategies

- Uncertainty sampling
- Query-by-committee
- Expected model change
- Expected error reduction
- Variance reduction

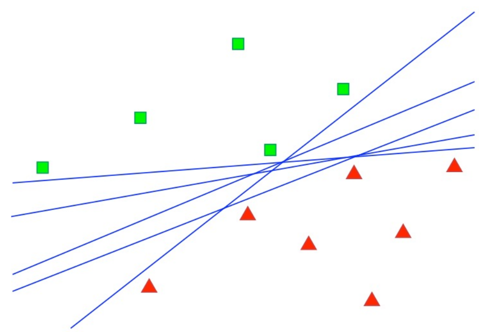
Uncertainty sampling

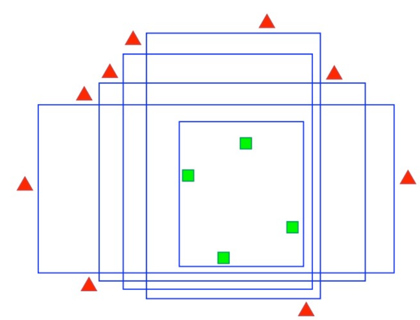
- Query the instance from the pool that the model is the least certain about
- Consider a multi-class problem with predicted label $\widehat{y_x} = \text{argmax}_y P(y|x)$: uncertainty could be quantified as $\text{argmin}_x\, P(\widehat{y_x}\,|x)$
  - Throws away any information about other classes
- Margin-based uncertainty $\text{argmin}_x\, P\left(\widehat{y_x}^{(1)}\middle|x\right) - P(\widehat{y_x}^{(2)}|x)$
- Entropy-based uncertainty $\text{argmax}_x\, \sum_i P(\widehat{y_x}^{(i)}\,|x) \log P\left(\widehat{y_x}^{(i)}\middle|x\right)$

Query-by-committee

- Train several different models on the same data
- Query for the data point they disagree on

$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}$$





Expected model change

- Which data point if we knew the label of would lead to the biggest change in the model parameters ?
- Measure the change with the gradient.
- Say $\mathcal{L}$ be the set of labelled data points

$$x^*_{EGL} = \underset{x}{\operatorname{argmax}} \sum_i P_\theta(y_i|x) \left\| \nabla \ell_\theta(\boldsymbol{\mathcal{L}} \cup \langle x, y_i \rangle) \right\|$$

Expected Error reduction

- Which data point to label to minimize the *expected* loss?
- For a loss function $\ell(y, P_\theta(\hat{y}|x))$, we could use

- $\text{Argmin}_x \; \sum_i P_\theta\left(y^{(i)}\middle|x\right) \sum_u (\ell(y^{(i)}, P_{\theta+(x,y^{(i)})}(\hat{y}|x_u)))$
- Here $P_\theta\left(y^{(i)}\middle|x\right)$ is the posterior without including any new data point
- $P_{\theta+(x,y^{(i)})}$ is the posterior with including a new data point $x, y^{(i)}$

Variance reduction

- Make use of geometry of the loss function to estimate expected variance in output of each unlabelled data point
    - Also makes use of gradient, as well as the Hessian (2^{nd} order information)

Extensions

- Structured outputs e.g. labelled sequences
- Variable labelling costs
- Active feature collection or active data completion
- Active class selection i.e. “For which class should the model query for a new data point?”
- Semi-supervised learning