Active Learning

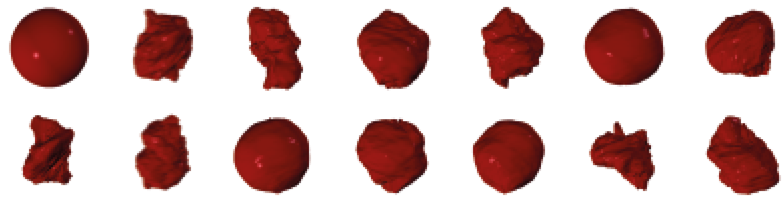
SPiNCOM reading group

Sep. 30th , 2016

Dimitris Berberidis

A toy example: Alien fruits

- Consider alien fruits of various shapes



- Train classifier to distinguish safe fruits from dangerous ones

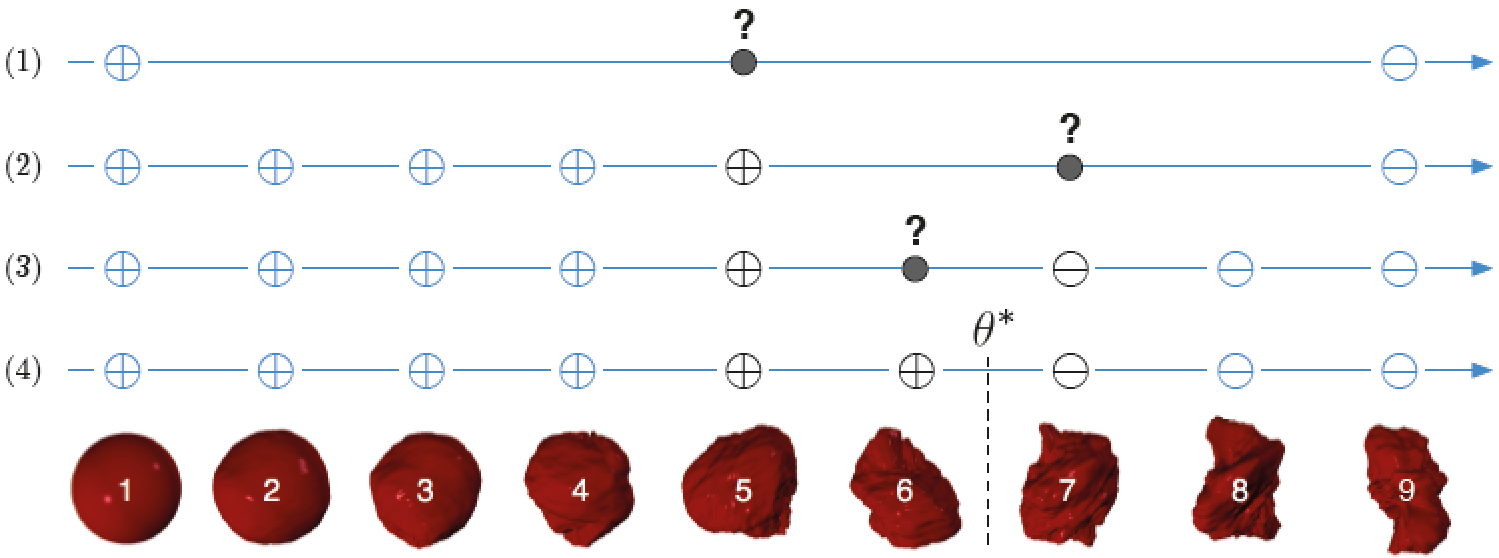
$$\mathcal{X} = \{\text{shape of fruits}\}$$
$$\mathcal{Y} = \{\text{safe, noxious}\}$$

$$h(x;\theta) = \begin{cases} \oplus \text{ safe} & \text{if } x < \theta \text{, and} \\ \ominus \text{ noxious} & \text{otherwise} \end{cases}$$

- Passive learning: Training data are given by uniform sampling and labeling
- Our setting
  - Obtaining labels costly
  - Unlabeled instances easily available

A toy example: alien fruits

- What if we sample fruits smartly instead of randomly?



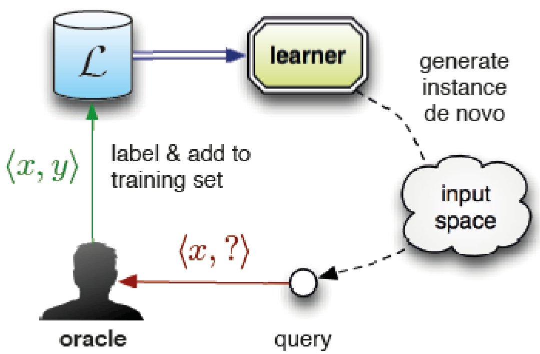
- θ^* can be identified with using far fewer samples

Active learning

General Goal: For a given budget of labeled training data, maximize learner's accuracy by actively selecting which instances (feature vectors) to label ("query").

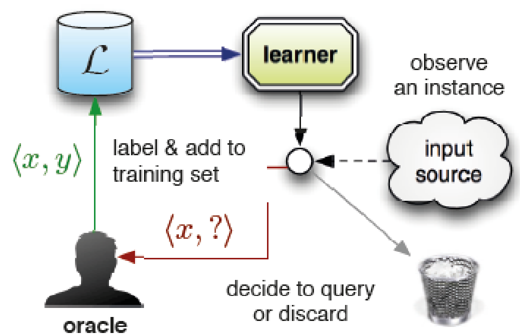
- Active learning (AL) scenarios considered

Query synthesis



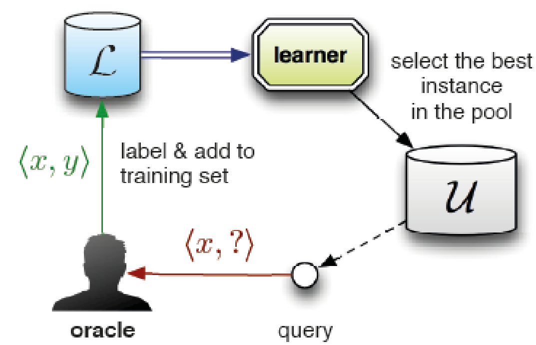
First to be considered, often not applicable

Selective sampling



Ideal for online settings with streaming data

Pool-based sampling



More general, OUR FOCUS

Roadmap

- ❑ Uncertainty sampling
- ❑ Searching the hypothesis space
    - Query by disagreement
    - Query by committee
- ❑ Expected error minimization
    - Expected error reduction
    - Variance reduction
    - Batch queries and submodularity
- ❑ Cluster-based AL
- ❑ AL + semi-supervised learning
- ❑ A unified view
- ❑ Conclusions

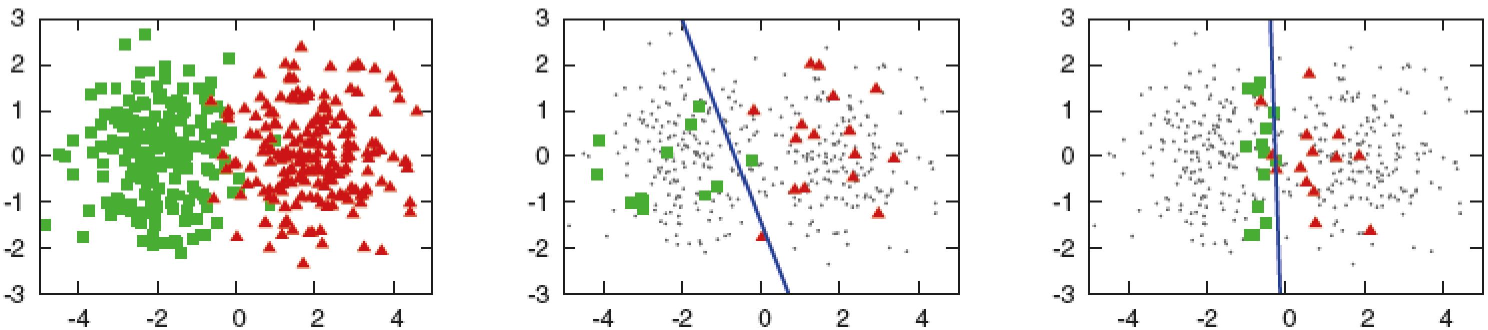
Burr Settles, “Active Learning”, *Synthesis lectures on AI and ML*, 2012.

Uncertainty sampling

- Most popular AL method: Intuitive, easy to implement

$\mathcal{U}$ = a pool of unlabeled instances $\{x^{(u)}\}_{u=1}^{U}$
$\mathcal{L}$ = set of initial labeled instances $\{\langle x, y\rangle^{(l)}\}_{l=1}^{L}$
for $t = 1, 2, \ldots$ **do**
    $\theta = \textbf{train}(\mathcal{L})$
    select $x^* \in \mathcal{U}$, the most uncertain instance according to model θ — the key
    query the oracle to obtain label y^*
    add $\langle x^*, y^*\rangle$ to $\mathcal{L}$
    remove x^* from $\mathcal{U}$
end for

- Support vector classifier: uncertain about points close to decision boundary



a 2D toy data set   random sampling   uncertainty sampling

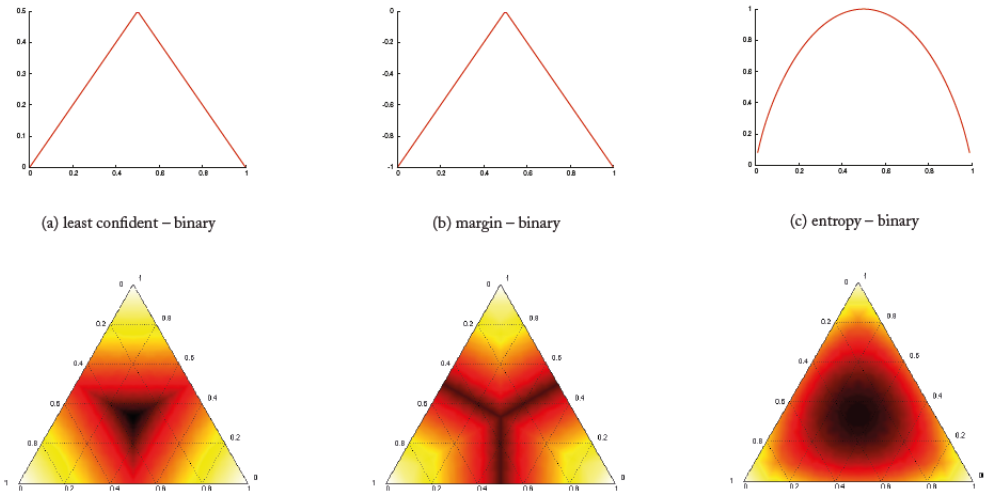
Measures of uncertainty

- Uncertainty of label as modeled by $P_\theta(y|x)$ (e.g. $P_\theta(y=1|x) = (1+e^{-\theta^T x})^{-1}$ for l.r.)

Least confident: $x^*_{LC} = \underset{x}{\operatorname{argmin}}\ P_\theta(\hat{y}|x) = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x)$ where $\hat{y} = \operatorname{argmax}_y P_\theta(y|x)$

Least margin: $x^*_M = \underset{x}{\operatorname{argmin}} \left[P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)\right] = \underset{x}{\operatorname{argmax}} \left[P_\theta(\hat{y}_2|x) - P_\theta(\hat{y}_1|x)\right]$

Highest entropy: $x^*_H = \underset{x}{\operatorname{argmax}}\ H_\theta(Y|x) = \underset{x}{\operatorname{argmax}}\ -\sum_y P_\theta(y|x) \log P_\theta(y|x)$

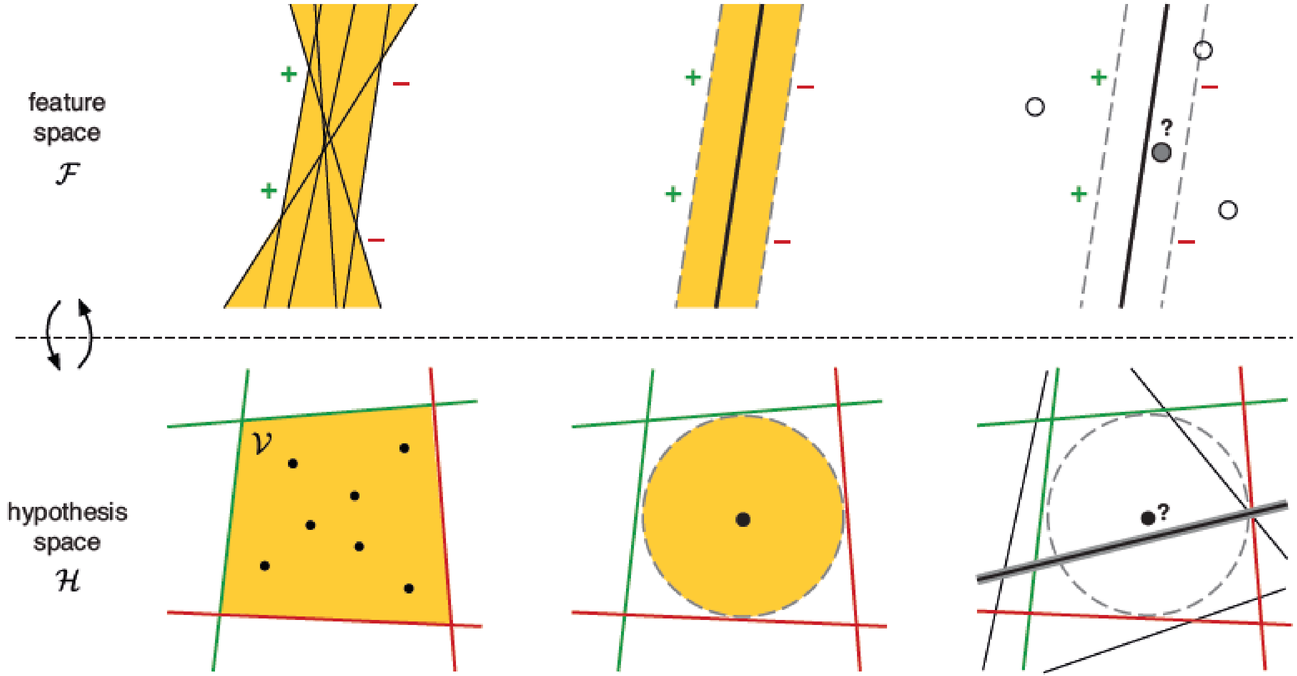


(a) least confident – binary (b) margin – binary (c) entropy – binary

(d) least confident – ternary (e) margin – ternary (f) entropy – ternary

- Limitation: Utility scores based on output of single (possibly bad) hypothesis.

Searching through the hypothesis space

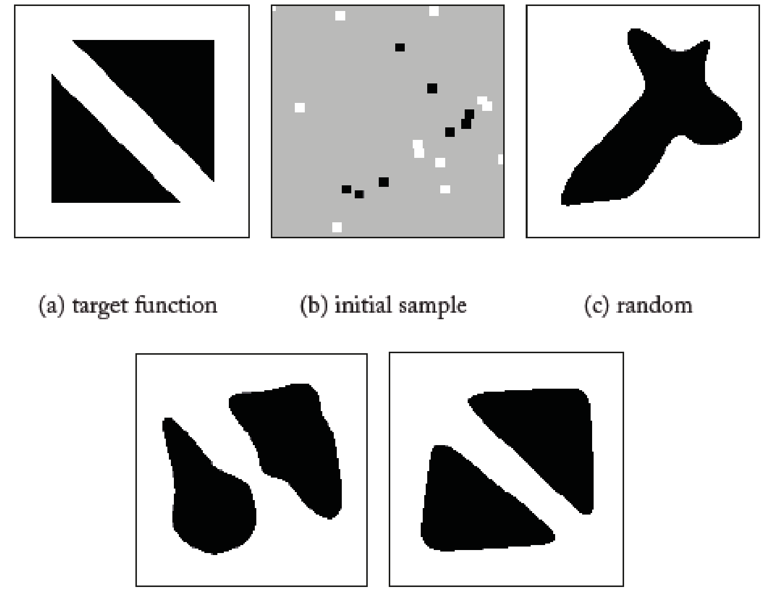


- Instance points in $\mathcal{H}$ correspond to hyperplanes in $\mathcal{F}$
- Version space $\mathcal{V} \subseteq \mathcal{H}$: Subset of all hypotheses consistent with tr. data
  - Max. margin methods (e.g. SVMs) lead to hypotheses in center of $\mathcal{V}$
  - Labeling instances close to decision hyperplane approx. bisects $\mathcal{V}$
  - Instances that greatly reduce the volume of $\mathcal{V}$ are of interest.

Query by disagreement

- One of the oldest AL algorithms [Cohn et al., ‘94]

$\mathcal{V} \subseteq \mathcal{H}$ is the set of all “legal” hypotheses
for $t = 1, 2, \ldots$ **do**
  receive instance $x \backsim \mathcal{D}_X$
  if $h_1(x) \neq h_2(x)$ for any $h_1, h_2 \in \mathcal{V}$ **then**
    query label y for instance x
    $\mathcal{L} = \mathcal{L} \cup \langle x, y \rangle$
    $\mathcal{V} = \{h : h(x') = y' \text{ for all } \langle x', y' \rangle \in \mathcal{L}\}$
  else
    do nothing; discard x
  end if
end for
return the labeled set $\mathcal{L}$ for training



(a) target function (b) initial sample (c) random
(d) uncertainty (e) disagreement

- “Store” version space implicitly with following trick

$$h_1 = \textbf{train}(\mathcal{L} \cup \langle x, \oplus \rangle) \quad \text{and} \quad h_2 = \textbf{train}(\mathcal{L} \cup \langle x, \ominus \rangle)$$

- Limitations: Too complex, all controversial instances treated equally

Query by committee

- Independently train a committee $\mathcal{C}$ of $|\mathcal{C}|$ hypotheses.
- Label instance most controversial among committee members

Vote entropy: $$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_y \frac{\text{vote}_{\mathcal{C}}(y,x)}{|\mathcal{C}|} \log \frac{\text{vote}_{\mathcal{C}}(y,x)}{|\mathcal{C}|}$$

$$\text{vote}_{\mathcal{C}}(y,x) = \sum_{\theta \in \mathcal{C}} \mathbf{1}_{\{h_\theta(x)=y\}}$$

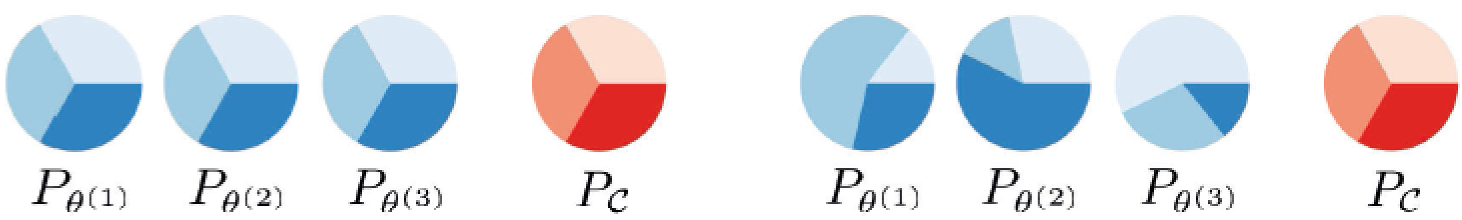
Soft vote entropy: $$x^*_{SVE} = \underset{x}{\operatorname{argmax}} - \sum_y P_{\mathcal{C}}(y|x) \log P_{\mathcal{C}}(y|x)$$

$$P_{\mathcal{C}}(y|x) = \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} P_\theta(y|x)$$

KL divergence: $$x^*_{KL} = \underset{x}{\operatorname{argmax}} \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} KL(\, P_\theta(Y|x) \parallel P_{\mathcal{C}}(Y|x)\,)$$

$$KL(\, P_\theta(Y|x) \parallel P_{\mathcal{C}}(Y|x)\,) = \sum_y P_\theta(y|x) \log \frac{P_\theta(y|x)}{P_{\mathcal{C}}(y|x)}$$

- Key difference: VE cannot distinguish between case (a) and (b)



(a) uncertain but in agreement (b) uncertain and in disagreement

Information theoretic interpretation

- Ideally maximize information between label r.v. Y and $\mathcal{V}$

$$I(Y;\mathcal{V}) \;=\; H(\mathcal{V}) - H(\mathcal{V}|Y) \;=\; H(\mathcal{V}) - \mathbb{E}_Y\big[H(\mathcal{V}|y)\big]$$

- Problem can be reformulated in more convenient form

$$I(Y;\mathcal{V}) \;=\; H(Y) - H(Y|\mathcal{V}) \;=\; H(Y) - \boxed{\mathbb{E}_{\theta\in\mathcal{V}}\big[H_\theta(Y)\big]}$$

Measures disagreement

- Uncertainty sampling focuses on maximizing $H(Y)$
  - QBC approximates second term with $\mathcal{C} \approx \mathcal{V}$ and $p(\theta) = \frac{1}{|C|}$

- Another alternative formulation (recall KL-based QBC)

$$I(Y;\mathcal{V}) \;=\; KL\big(\,P(Y,\mathcal{V}) \,\|\, P(Y)P(\mathcal{V})\,\big) \;=\; \mathbb{E}_{\theta\in\mathcal{V}}\Big[\,KL\big(\,P_\theta(Y) \,\|\, P(Y)\,\big)\,\Big]$$

QBC approximates: $P(Y) \approx P_C(Y)$

Bound on label complexity

- Label complexity for passive learning (assume $\mathcal{L} \backsim \mathcal{D}_{XY}$)

To achieve $P\left(\mathrm{err}(h_t) \leq \epsilon\right) \geq 1-\delta$, one needs $L_{\mathrm{PASS}} \leq O\left(\frac{1}{\epsilon}\left(d\log\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right) = \tilde{O}\left(\frac{d}{\epsilon}\right)$

where $\mathrm{err}(h_t)$ is expected error rate and VC dimension d measures complexity of $\mathcal{H}$

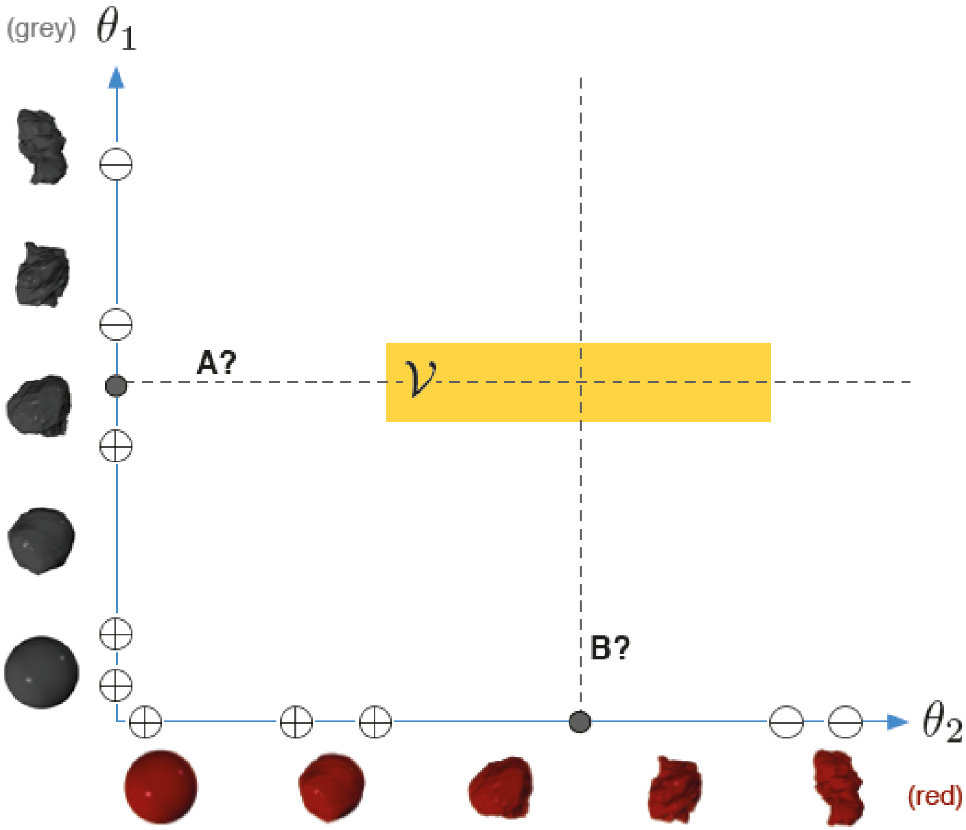
- Dis. coef. ξ : Quantifies how fast the reg. of disagreement shrinks

$$\left.\begin{aligned}
\Delta(h_1, h_2) &= P_{\mathcal{D}}(h_1(x) \neq h_2(x)) \\
B(h^*, r) &= \{h \in \mathcal{H} \mid \Delta(h^*, h) \leq r\} \\
\mathrm{DIS}(\mathcal{V}) &= \{x \in \mathcal{X} \mid \exists\, h_1, h_2 \in \mathcal{V} : h_1(x) \neq h_2(x)\}
\end{aligned}\right\} \quad \xi = \sup_{r>0} \frac{P_{\mathcal{D}}(\mathrm{DIS}(B(h^*, r)))}{r}$$

- QBD achieves logarithmically lower label complexity (if ξ does not explode)

$$L_{\mathrm{QBD}} \leq O\left(\xi\left(d\log\xi+\log\frac{\log 1/\epsilon}{\delta}\right)\log\frac{1}{\epsilon}\right) = \tilde{O}\left(\xi d\log\frac{1}{\epsilon}\right)$$

Alien fruit example: A problematic case



- Candidate queries A and B both bisect $\mathcal{V}$ (appear equally informative)
  - However, generalization error depends on the (ignored) distribution of input

- Generally: Both unc. sampling and QBD may suffer high generalization error

Expected error reduction

- Ideally select query by minimizing expected generalization error

Retrained model using (x, y)

$$x^*_{ER} = \underset{x}{\operatorname{argmin}} \; \mathbb{E}_{Y|\theta,x}\left[\sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta^+,x'}[y \neq \hat{y}]\right] = \underset{x}{\operatorname{argmin}} \sum_{y} P_\theta(y|x)\left[\sum_{x' \in \mathcal{U}} 1 - p_{\theta^+}(\hat{y}|x')\right]$$

- Less stringent objective: Expected log-loss

$$\begin{aligned} x^*_{LL} &= \underset{x}{\operatorname{argmin}} \; \mathbb{E}_{Y|\theta,x}\left[\sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta^+,x'}[-\log p_{\theta^+}(y|x')]\right] \\ &= \underset{x}{\operatorname{argmin}} \sum_{y} P_\theta(y|x)\left[\sum_{x' \in \mathcal{U}} -\sum_{y'} p_{\theta^+}(y'|x') \log p_{\theta^+}(y'|x')\right] \\ &= \underset{x}{\operatorname{argmin}} \sum_{y} P_\theta(y|x) \sum_{x' \in \mathcal{U}} H_{\theta^+}(Y|x') \end{aligned}$$

- (Extremely) high complexity required to retrain model for each candidate

Variance reduction

- Learners expected error can be decomposed

$$\mathbb{E}\left[(\hat{y}-y)^2|x\right] = \mathbb{E}_{Y|x}\left[\underbrace{(y-\mathbb{E}_{Y|x}[y|x])^2}_{\text{Noise}}\right] + \underbrace{\left(\mathbb{E}_{\mathcal{L}}[\hat{y}]-\mathbb{E}_{Y|x}[y|x]\right)^2}_{\text{Bias}} + \underbrace{\mathbb{E}_{\mathcal{L}}\left[(\hat{y}-\mathbb{E}_{\mathcal{L}}[\hat{y}])^2\right]}_{\text{Variance}}$$

- Noise is ind. of training data and bias is due to model class (e.g. linear model)

- Focus on minimizing variance of predictions of unlabeled data

$$x^*_{VR} = \underset{x}{\operatorname{argmin}} \sum_{x' \in \mathcal{U}} \mathrm{Var}_{\theta+}(Y|x')$$

- Question: Can we minimize variance <u>without retraining</u>?
  - Design of experiments approach (typically for regression)

Optimal experimental design

- Fisher information matrix (FIM) of model

Fisher score

$$F = \mathbb{E}_X\left[\left(\frac{\partial}{\partial\theta}\log P_\theta(Y|x)\right)^2\right] = \mathbb{E}_X\left[\frac{\partial^2}{\partial\theta^2}\log P_\theta(Y|x)\right] = \sum_x \mathbb{E}\left[\nabla x \nabla x^T\right]$$

- Covariance of parameter estimates lower bounded by F^{-1}
- A-optimal design: $x^*_A = \arg\min_x \operatorname{tr}\left(\left[\underbrace{F_{\mathcal{L}} + \mathbb{E}[\nabla x \nabla x^T]}_{\text{Additive property of FIM}}\right]^{-1}\right)$
- Can easily be adapted to minimize variance of predictions

$$\begin{aligned} X^*_{\text{FIR}} &= \arg\min_x \sum_{x' \in U} \mathbf{Var}_{\theta^+}(Y|x') \\ &= \arg\min_x \sum_{x' \in U} \mathbf{tr}(A_{x'}\left[F_{\mathcal{L}} + \mathbb{E}\left[\nabla x \nabla x^\top\right]\right]) \\ &= \arg\min_x \mathbf{tr}(F_U\left[\left[F_{\mathcal{L}} + \mathbb{E}\left[\nabla x \nabla x^\top\right]\right]\right]^{-1}) \end{aligned}$$

Fisher information ratio

- FIM can be efficiently updated using the Woodberry matrix identity

Batch queries and submodularity

- Query a batch $\mathcal{Q}$ of instances
  - Not necessarily the $|\mathcal{Q}|$ individually best
  - Key is to avoid correlated instances
- Submodularity property for functions over sets ($\mathcal{A} \subseteq \mathcal{A}'$)

$$s(\mathcal{A} \cup \{x\}) - s(\mathcal{A}) \geq s(\mathcal{A}' \cup \{x\}) - s(\mathcal{A}')$$

$$s(\mathcal{A}) + s(\mathcal{B}) \geq s(\mathcal{A} \cup \mathcal{B}) + s(\mathcal{A} \cap \mathcal{B})$$

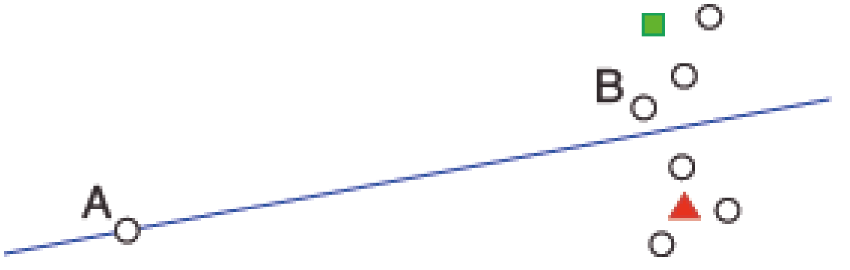
- Greedy approach on submodular function guarantees: $(1 - \frac{1}{e}) \times s(\mathcal{A}_N^*)$
- Maximizing the variance difference can be submodular

$$\begin{aligned} s(\mathcal{Q}) &= \sum_{x \in \mathcal{U}} \mathrm{Var}_\theta(Y|x) - \mathrm{Var}_{\theta+\mathcal{Q}}(Y|x) \\ &= \mathbf{tr}(F_\mathcal{U} F_\mathcal{L}^{-1}) - \mathbf{tr}(F_\mathcal{U}[F_\mathcal{L} + F_\mathcal{Q}]^{-1}) \end{aligned}$$

- For linear regression FIM is ind. of θ (offline computation !)

Density-weighted methods

- Back to classification
- Pathological case: Least confident (most uncertain) instance is an outlier
  - B in fact more informative than A



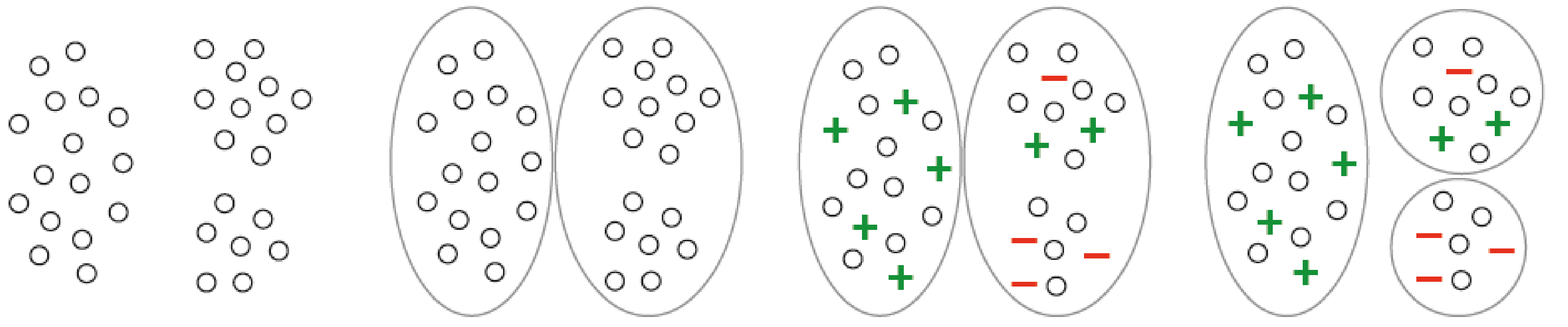
- Error and variance reduction less sensitive to outliers but costly
- Information density heuristic
  - Instances more representative of input distribution are promoted

$$x^*_{ID} = \underset{x}{\operatorname{argmax}} \; \phi_A(x) \times \left(\frac{1}{U} \sum_{x' \in \mathcal{U}} \operatorname{sim}(x, x') \right)^{\beta}$$

Information utility score (e.g. entropy)

Similarity measure (e.g. Eucledian distance)

Hierarchical cluster-based AL



(a) data (b) clustering (c) queries (d) refinement

- Assist AL by clustering the input space
  - Obtain data and find initial coarse clustering
  - Query instances from different clusters
  - Iteratively refine clusters so that they become more “pure”
  - Focus querying on more impure clusters

- Working assumption: Cluster structure is correlated with label structure
  - If not, above algorithm degrades to random sampling

Active and semi-supervised learning

- Two approaches are complementary
  - AL minimizes labeling effort by querying most informative instances
  - Semi-sup. learning exploits latent structure (unlabeled) to improve accuracy

- *Self training is* complementary to *uncertainty sampling* [Yarowsky, '95]

- *Co-training* complementary to *QBD* [Blum and Mitchel, '98]

- *Entropy regularization* complementary to *error reduction w. log-loss*

$$\ell_\theta(\mathcal{L},\mathcal{U}) = \sum_{\langle x,y\rangle\in\mathcal{L}} \log P_\theta(y|x) \; - \sum_k \frac{\theta_k^2}{2\sigma^2} \; - \sum_{x'\in\mathcal{U}} H_\theta(Y|x')$$

Unified view (I)

- Ideal: Maximize total gain in information

$$x^* = \underset{x}{\operatorname{argmax}} \sum_{x' \in \mathcal{U}} \Big(H_\theta(Y|x') - H_{\theta+}(Y|x') \Big)$$

- Since true label is unknown, one resorts to

$$x^* = \underset{x}{\operatorname{argmax}} \sum_{x' \in \mathcal{U}} \Big(H_\theta(Y|x') - \mathbb{E}_{Y|\theta,x}\big[H_{\theta+}(Y|x')\big] \Big)$$

- Approximations lead to uncertainty sampling heuristic

$$\begin{aligned} x^* &= \underset{x}{\operatorname{argmax}} \sum_{x' \in \mathcal{U}} \Big(H_\theta(Y|x') - \mathbb{E}_{Y|\theta,x}\big[H_{\theta+}(Y|x')\big] \Big) \\ &\approx \underset{x}{\operatorname{argmax}}\ H_\theta(Y|x) - \mathbb{E}_{Y|\theta,x}\big[H_{\theta+}(Y|x)\big] \\ &\approx \underset{x}{\operatorname{argmax}}\ H_\theta(Y|x) \end{aligned}$$

Uncertainty sampling

Unified view (II)

- A different approximation

$$
\begin{aligned}
x^* &= \operatorname*{argmax}_x \sum_{x' \in \mathcal{U}} \Big(H_\theta(Y|x') - \mathbb{E}_{Y|\theta,x}\big[H_{\theta^+}(Y|x')\big] \Big) \\
&= \operatorname*{argmax}_x \sum_{x' \in \mathcal{U}} H_\theta(Y|x') - \sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta,x}\big[H_{\theta^+}(Y|x')\big] \\
&= \operatorname*{argmin}_x \sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta,x}\big[H_{\theta^+}(Y|x')\big]
\end{aligned}
$$

ınged for all queries

- Log-loss minimization and variance-reduction target the above measure

- Approximation given by density weighted methods

$$
\begin{aligned}
x^* &= \operatorname*{argmax}_x \sum_{x' \in \mathcal{U}} \Big(H_\theta(Y|x') - \mathbb{E}_{Y|\theta,x}\big[H_{\theta^+}(Y|x')\big] \Big) \\
&\approx \operatorname*{argmax}_x \sum_{x' \in \mathcal{U}} \Big(\mathrm{sim}(x, x') \times H_\theta(Y|x) \Big)
\end{aligned}
$$

Overview

Query Strategy	Advantages	Disadvantages
uncertainty sampling	simplest approach, very fast, easy to implement, usable with any probabilistic model, justifiable for max-margin classifiers	myopic, runs the risk of becoming overly confident about incorrect predictions
QBC and disagreement-based methods	reasonably simple, usable with almost any base learning algorithm, theoretical guarantees under some conditions	can be difficult to train/maintain multiple hypotheses, still myopic in terms of reducing generalization error
error/variance reduction	directly optimizes the objective of interest, empirically successful, natural extension to batch queries with some guarantees	computationally expensive, difficult to implement, limited to pool-based or synthesis scenarios, VR limited to regression models
density weighting	simple, inherits advantages of the base heuristic while making it less myopic in terms of the input distribution, can be made fast	input distribution or cluster structure may have no relationship to the labels
hierarchical sampling	exploits cluster structure, degrades gracefully if clusters are not correlated with the labels, theoretical guarantees	requires a *hierarchical* clustering of the data, which can be slow and expensive in practice, limited to pool-based scenario
active + semi-supervised	exploits latent structure in the data, aims to make good use of data through both active and semi-supervised methods	not a single algorithm/framework but a suite of approaches, inherits the pros and cons of the base algorithms

Practical considerations

- Real labeling costs

$$x^*_{VOI} = \underset{x}{\operatorname{argmin}} \; \underbrace{c(x)}_{\text{Cost of annotating specific query}} + \mathbb{E}_{Y|\theta,x}\left[\sum_{x' \in \mathcal{U}} \underbrace{\kappa_{\theta^+}(\hat{y}, x')}_{\text{Cost of prediction}} \right]$$

- Multi-task AL (multiple labels per instance)

$$\begin{aligned} x^*_{MT} &= \underset{x}{\operatorname{argmax}} \; H_\theta(Y_1, Y_2|x) \\ &= \underset{x}{\operatorname{argmax}} \; H_\theta(Y_1|x) + H_\theta(Y_2|Y_1, x) \\ &= \underset{x}{\operatorname{argmax}} \; H_\theta(Y_1|x) + \mathbb{E}_{Y_1|x}\left[H_\theta(Y_2|y_1, x)\right] \end{aligned}$$

- Skewed label distributions (class imbalance)
- Unreliable oracles (e.g. labels given by human experts)
- When AL is used training data are biased to model class
  - If unsure about model, random sampling may be preferable

Conclusions

- AL allows for sample (label) complexity reduction
  - Simple heuristics: Uncertainty sampling, QBD,QBC, cluster-based AL
  - High complexity near-optimal methods: Expected error/variance reduction
  - Encompasses optimal experimental design
  - Linked to semi-supervised learning
  - Information-theoretic interpretations

- Possible research directions
  - Use of AL methods in learning over graphs (GSP, classification over graphs)
  - Use o MCMC and IS to approx. posterior in complex models (e.g. BMRF)

Thank You