Active Learning Beyond Label Feedback

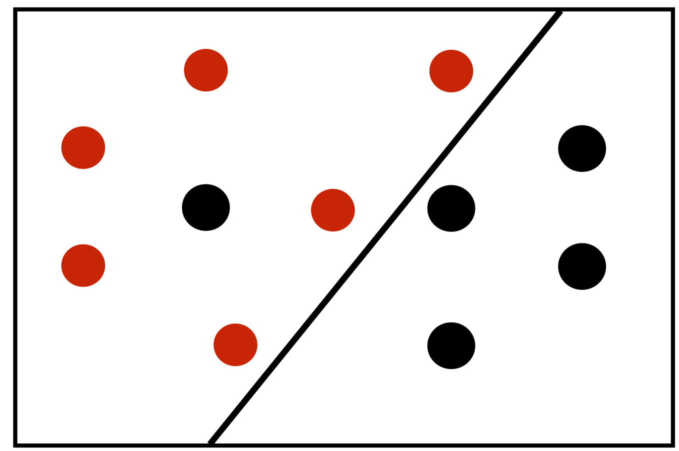
Kamalika Chaudhuri
University of California, San Diego

Joint work with **Chicheng Zhang, Tara Javidi** and **Songbai Yan**

Classification

Given: (x_i, y_i)

Vector of features — Discrete Labels



Find: Prediction rule in a class to predict y from x

Challenge: Acquiring Labeled Data

Unlabeled data
is cheap

Labels are
expensive

Active Learning

Given: $(x_i, \quad y_i)$

Find: Prediction rule to predict y from x

Active Learning

Given: $(x_i, \cancel{y_i})$

Interactive
Label Queries

Find: Prediction rule to predict y from x

Active Learning

Given: $(x_{i}, \cancel{y_i)}$

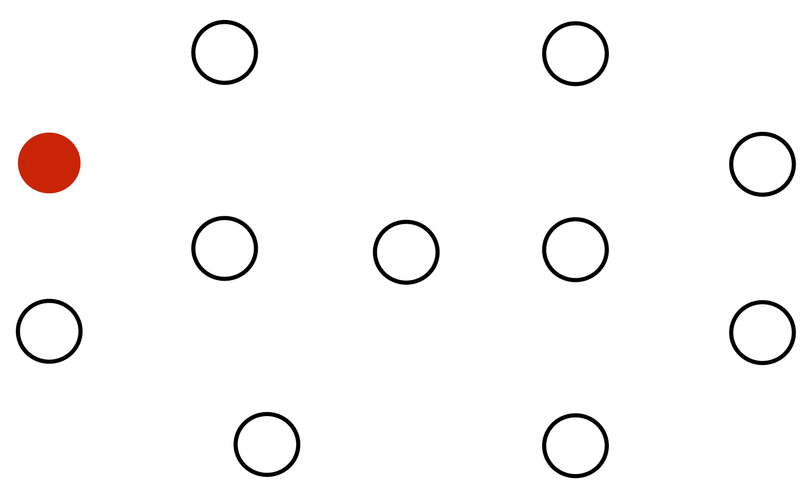
Interactive
Label Queries

Find: Prediction rule to predict y from x
using few label queries

Why Active Learning Helps?

Given: Unlabeled data, interactive label queries

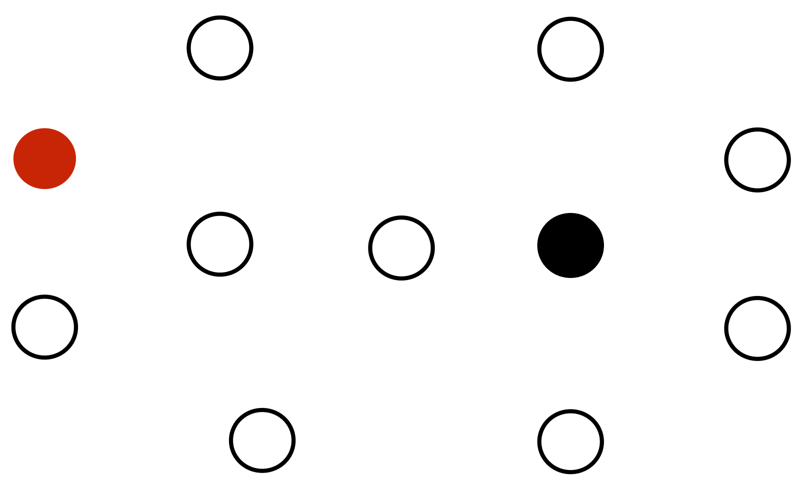
Find: Good prediction rule using few label queries



Why Active Learning Helps?

Given: Unlabeled data, interactive label queries

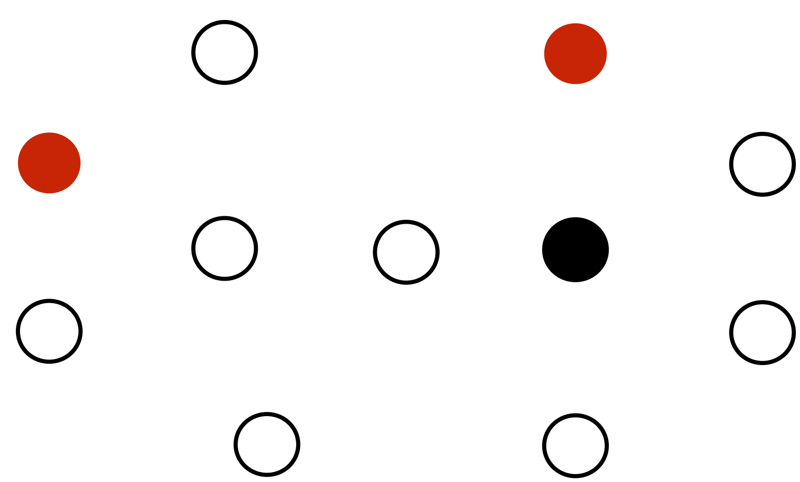
Find: Good prediction rule using few label queries



Why Active Learning Helps?

Given: Unlabeled data, interactive label queries

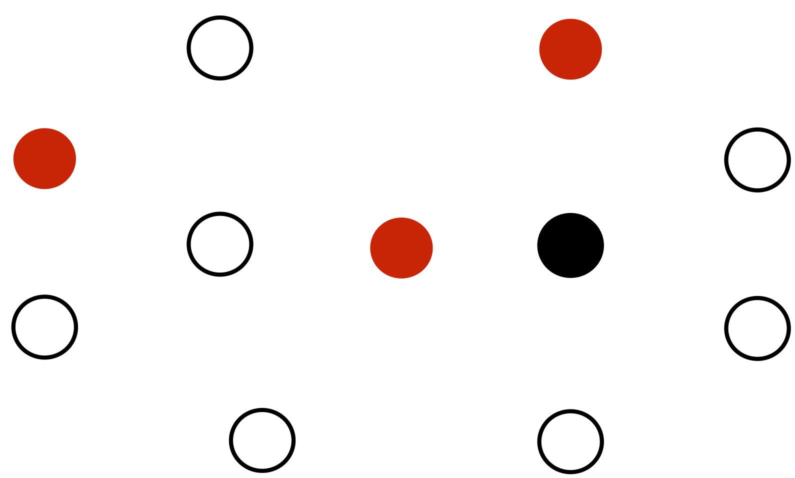
Find: Good prediction rule using few label queries



Why Active Learning Helps?

Given: Unlabeled data, interactive label queries

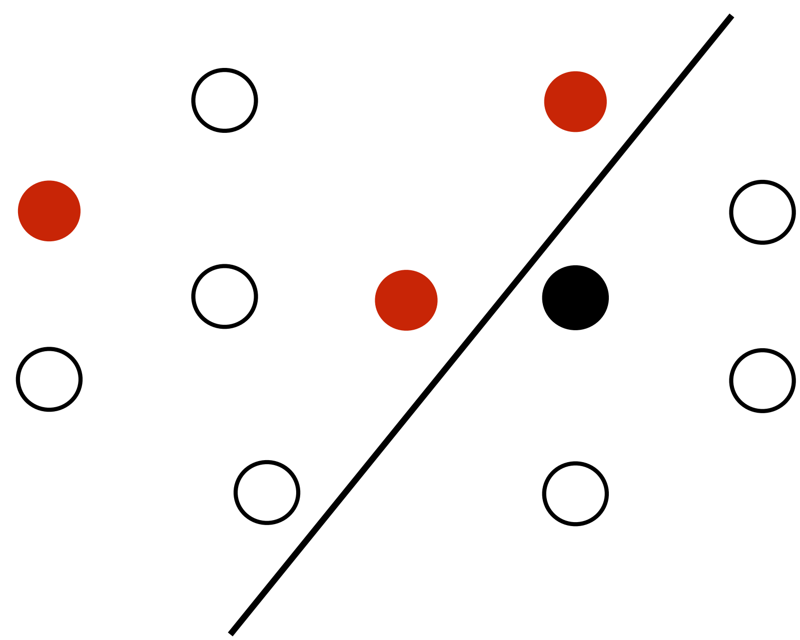
Find: Good prediction rule using few label queries



Why Active Learning Helps?

Given: Unlabeled data, interactive label queries

Find: Good prediction rule using few label queries



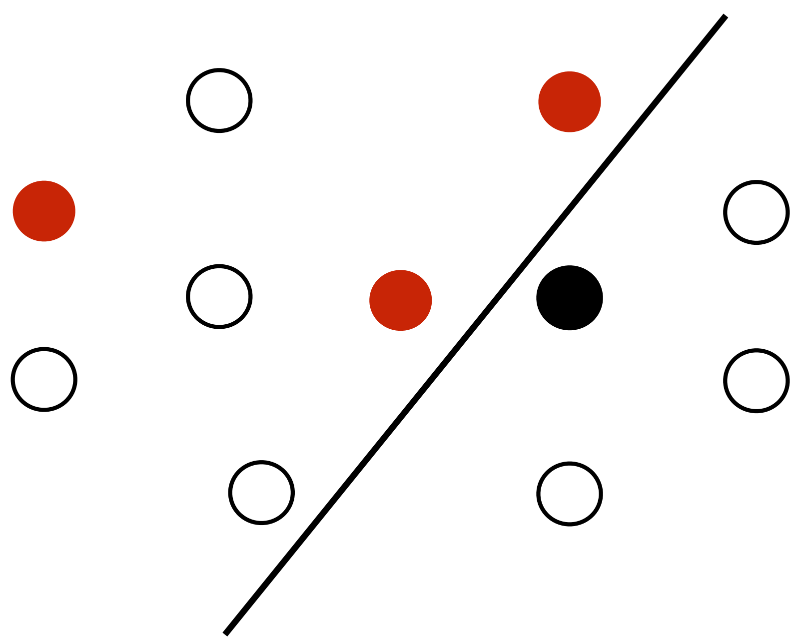
Challenge: “Incorrect” Responses

What makes Active Learning Hard?

Given: Unlabeled data, interactive label queries

No assumptions on data distribution

Find: Good prediction rule using few label queries

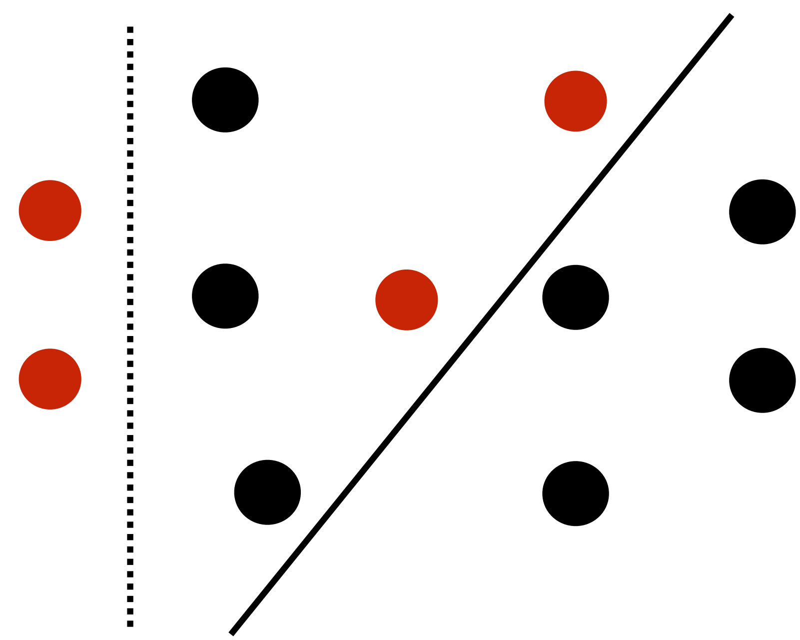


What makes Active Learning Hard?

Given: Unlabeled data, interactive label queries

No assumptions on data distribution

Find: Good prediction rule using few label queries



Statistically inconsistent!

Talk Agenda

Can other kinds of queries help active learning?

This talk:

1. Weak and strong labelers
2. Abstaining labelers

Talk Outline

1. Weak and Strong Labelers

   - the model

Probably Approx. Correct (PAC) Model

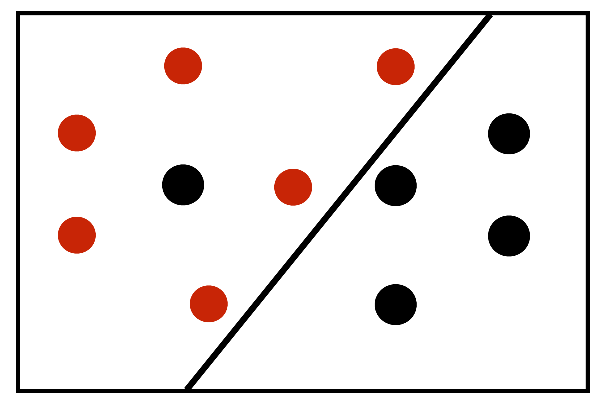
Given: Concept class C

Samples (x_i, y_i) from data distribution D

Example: C = linear classifiers

Find: c in C with low

$$\Pr_{(x,y)\sim D}(c(x) \neq y)$$



Probably Approx. Correct (PAC) Model

Given: Concept class C

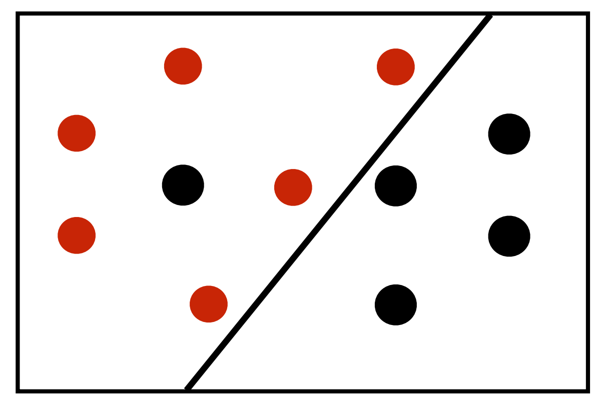
Samples (x_i, y_i) from data distribution D

Example: C = linear classifiers

Find: c in C with low

$$\Pr_{(x,y)\sim D}(c(x) \neq y)$$

Error

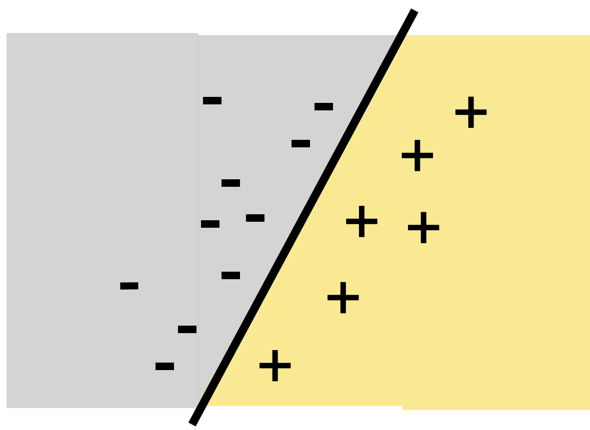


PAC Model: Realizable vs. Agnostic

Given: Concept class C, Samples (x_i, y_i) from D

Find: c in C with low error

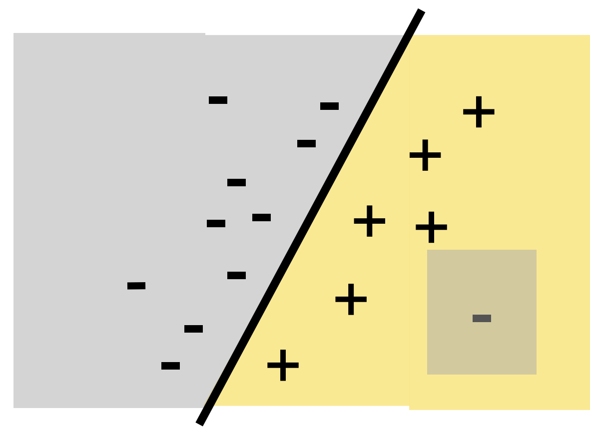
Realizable



$\exists c^* \in C$ such that

$c^*(x) = y, \forall (x, y) \sim D$

Agnostic



No Assumptions on D

Agnostic Active Learning

Given: Concept class C (best c in C has error ν^*)

$(x_i, \cancel{y_i})$ drawn from D

Interactive
Label Queries

Find: c in C with error $\leq \nu^* + \epsilon$
using few label queries
with no assumptions on D

Methods for Agnostic Active Learning

- Disagreement-based Active Learning [CAL94, BBL06, H07, DHM07, many others]
- Margin/Confidence-based Active Learning [BZ07, BL13, ABL14, ZC14]
- Clustering-based Active Learning [DH08, UWB13]

This work: based on disagreement-based active learning

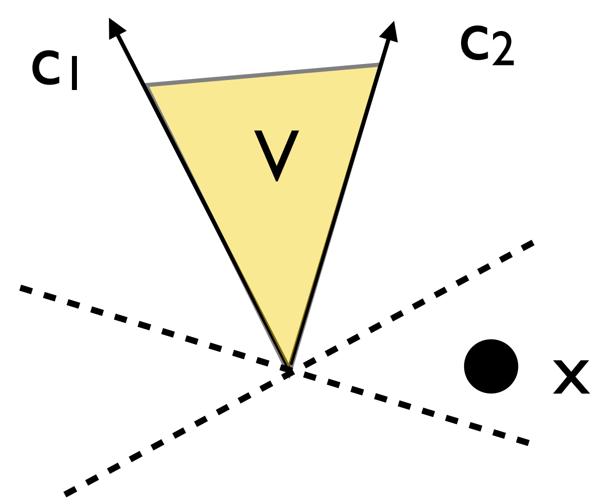
Disagreement-based Active Learning

1. Maintain candidate set V (that contains best c in C)
2. For unlabeled x, if there exist c_1, c_2 in V s.t

$$c_1(x) \neq c_2(x)$$

then, x is in **disagreement region** of V

Query label of x, and update V



[CAL94, BBL06, DHM07, H07, BDL09, BHLZ10, …]

What if we have auxiliary information?

..as an extra oracle

Oracle and Weak Labeler

Oracle:
expensive but correct

Weak labeler:
cheap, sometimes wrong

The Model

Given: $(x_i, \cancel{y_i})$

Interactive
Label Queries

The Model

Given: $(x_i, \cancel{y_i})$

Interactive Label Queries to Oracle O or Weak Labeler W

The Model

Given: $(x_i, \cancel{y_i)}$

Interactive Label Queries to Oracle O or Weak Labeler W

Find: Prediction rule to predict y from x using few label queries to O

Formal Model

Given: Concept class C (best c has error ν^* wrt O)

(x_i, ~~y_i)~~ drawn from D

Formal Model

Given: Concept class C (best c has error ν^* wrt O)

$(x_i, \cancel{y_i)}$ drawn from D

Oracle O and Weak labeler W

Formal Model

Given: Concept class C (best c has error ν^* wrt O)

(x_i, ~~y_i)~~ drawn from D

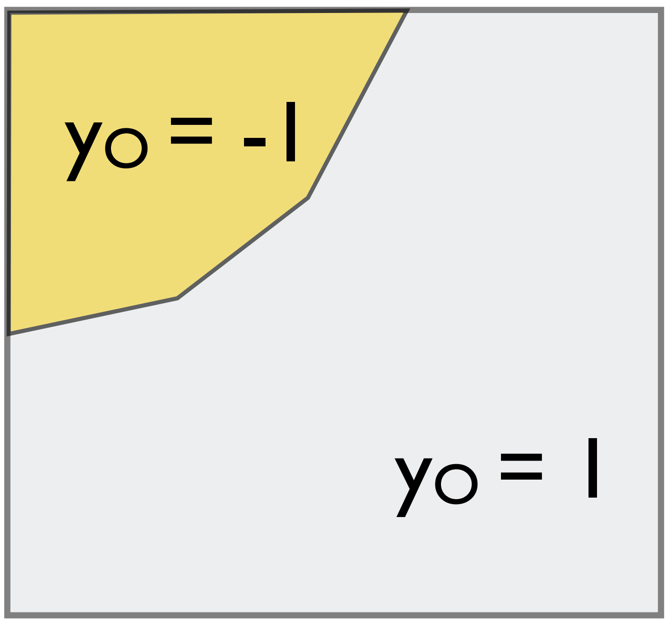
Oracle O and Weak labeler W

Find: c in C with error $\leq \nu^* + \epsilon$ wrt O

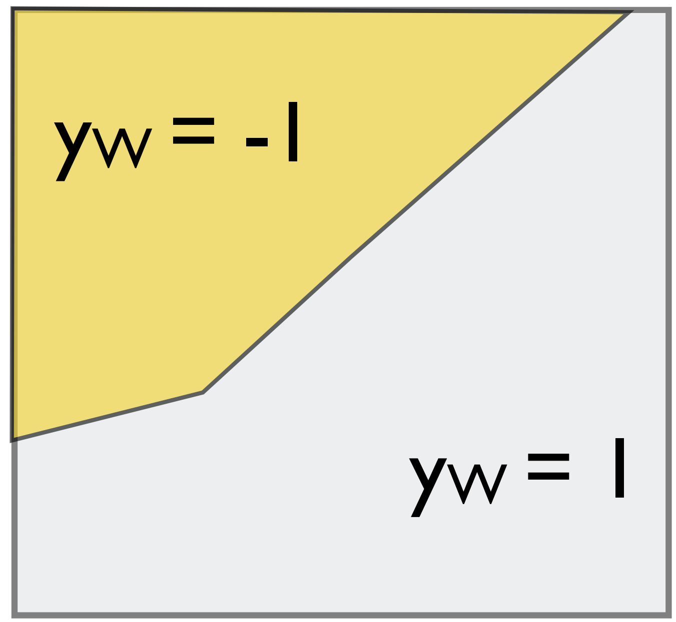
using minimum label queries to O

Formal Model Implications

Weak labeler W may be biased



Labels by O



Labels by W

Previous Work

[UBS12] Explicit assumptions on where W and O differ (close to decision boundaries)

[MCR14] No explicit assumptions, but applies to online selective classification and robust regression

This talk: General learning strategy from W and O with no explicit assumptions

Talk Outline

1. Weak and Strong Labelers
    - the model
    - algorithm

How to learn in this model?

Main Ideas:

Learn a **difference classifier** h to predict when O and W differ

How to learn in this model?

Main Ideas:

Learn a **difference classifier** h to predict when O and W differ

Use h with standard active learning to decide if we should query O or W

Algorithm Outline

1. Draw $x_1,..,x_m$. For each x_i, query O and W. Set:

$$y_{i,D} = 1 \quad \text{if} \qquad y_{i,O} \neq y_{i,W}$$

Algorithm Outline

1. Draw $x_1,..,x_m$. For each x_i, query O and W. Set:

   $$y_{i,D} = 1 \quad \text{if} \qquad y_{i,O} \neq y_{i,W}$$

2. Train **difference classifier** h in H on $\{ (x_i, y_{i,D}) \}$

Algorithm Outline

1. Draw $x_1,..,x_m$. For each x_i, query O and W. Set:

$$y_{i,D} = 1 \quad \text{if} \quad y_{i,O} \neq y_{i,W}$$

2. Train **difference classifier** h in H on $\{ (x_i, y_{i,D}) \}$

3. Run standard disagreement based active learning algorithm A. If A queries the label of x then:

if $h(x) = 1$, query O, else query W

Algorithm Outline

1. Draw $x_1,..,x_m$. For each x_i, query O and W. Set:

$$y_{i,D} = 1 \quad \text{if} \quad y_{i,O} \neq y_{i,W}$$

2. Train **difference classifier** h in H on $\{ (x_i, y_{i,D}) \}$

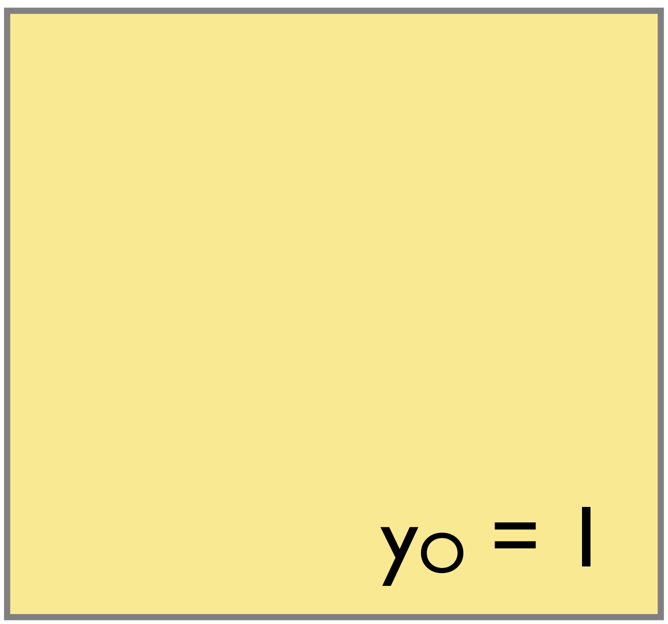
3. Run standard disagreement based active learning algorithm A. If A queries the label of x then:

   if $h(x) = 1$, query O, else query W

Is this **statistically consistent?**

Key Observation 1

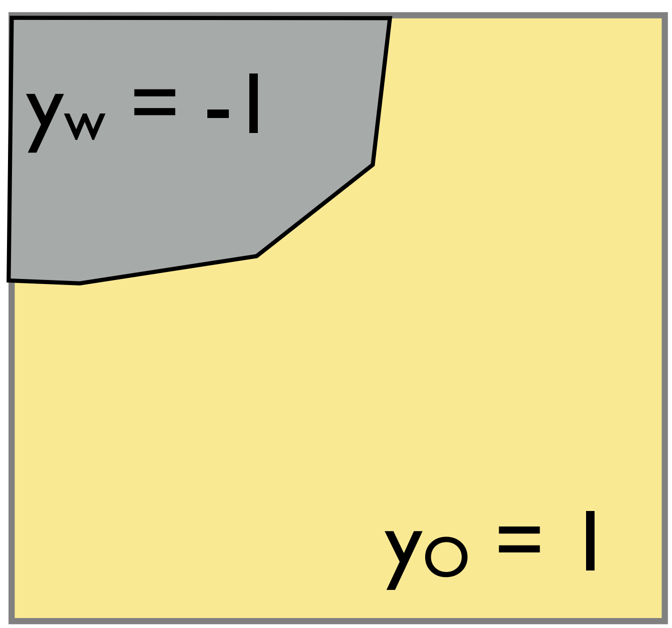
Directly learning difference classifier may lead to inconsistent annotation on target task



Actual Labels

Key Observation 1

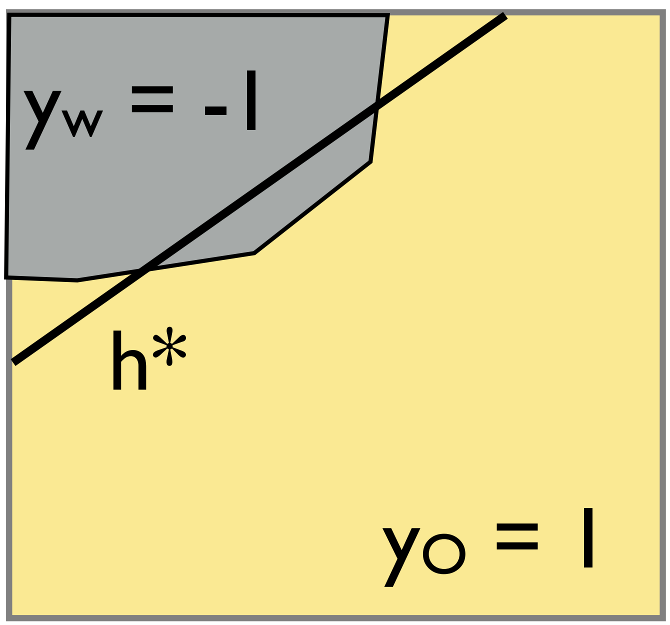
Directly learning difference classifier may lead to inconsistent annotation on target task



Actual Labels

Key Observation 1

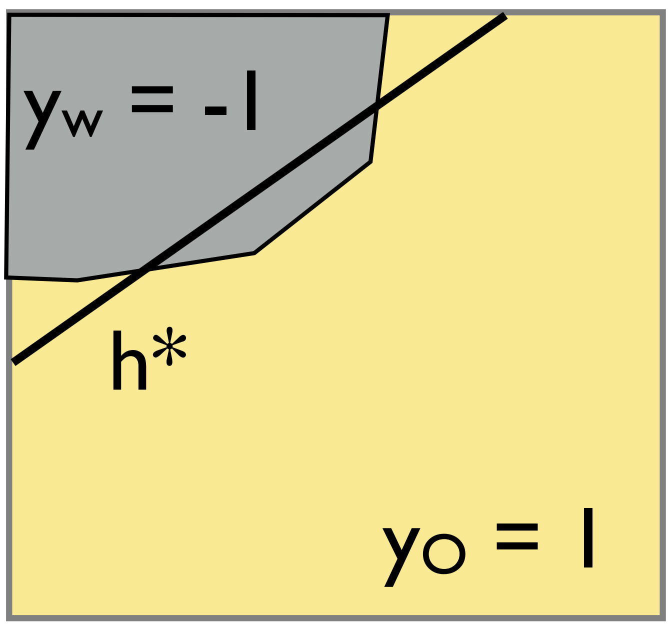
Directly learning difference classifier may lead to inconsistent annotation on target task



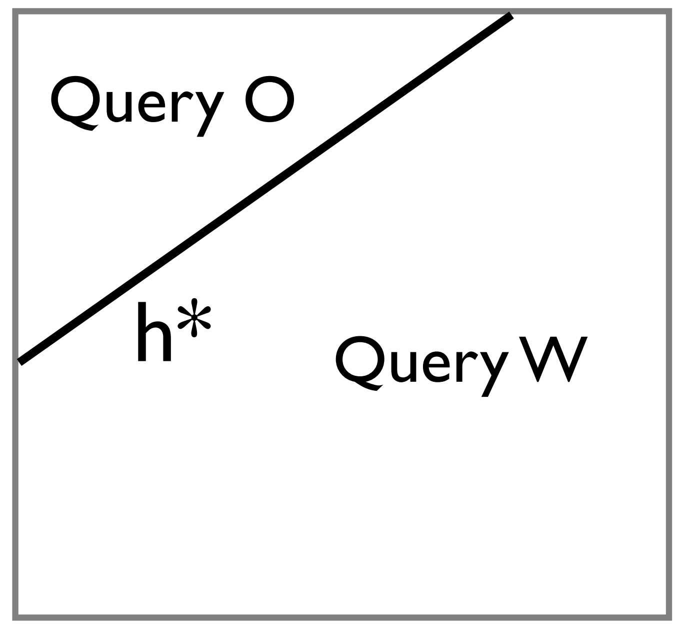
Actual Labels

Key Observation 1

Directly learning difference classifier may lead to inconsistent annotation on target task



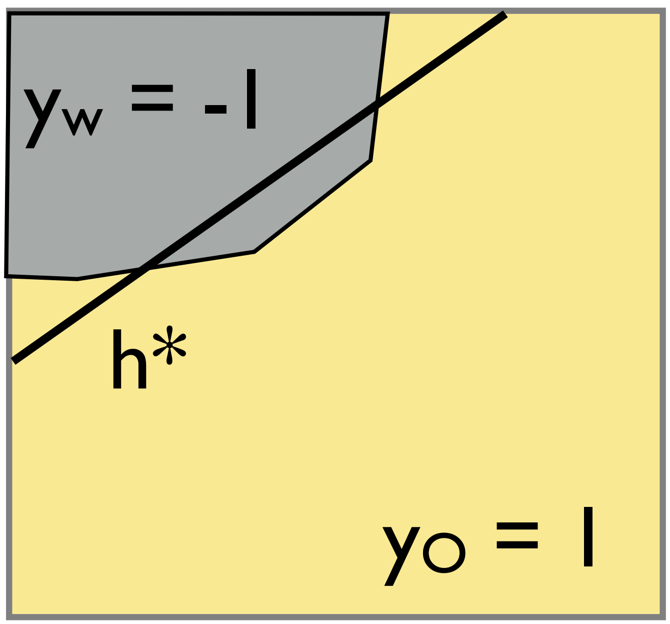
Actual Labels



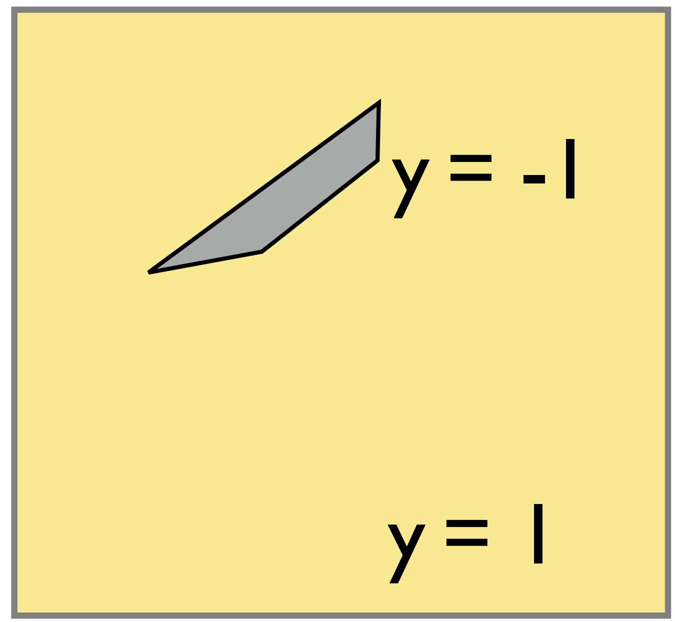
Annotation using h* as difference classifier

Key Observation 1

Directly learning difference classifier may lead to inconsistent annotation on target task



Actual Labels

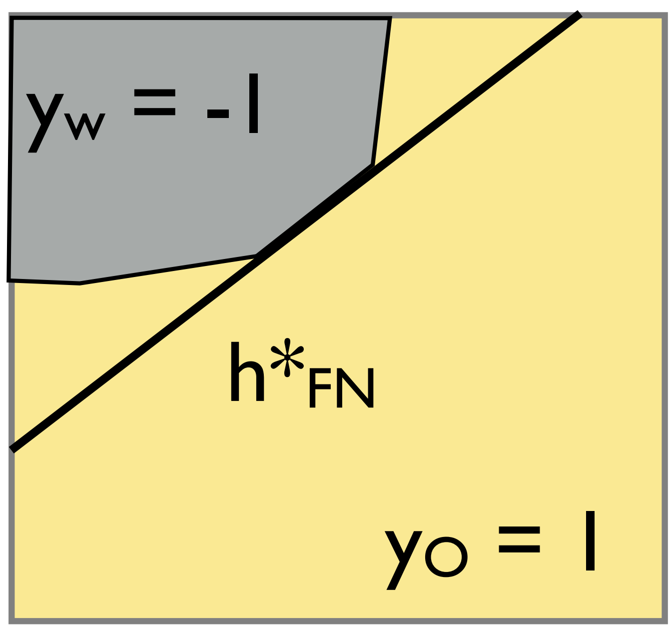


Annotation using h* as difference classifier

Solution

Train a **cost-sensitive** difference classifier

Constrain False Negative (FN) rate as very low

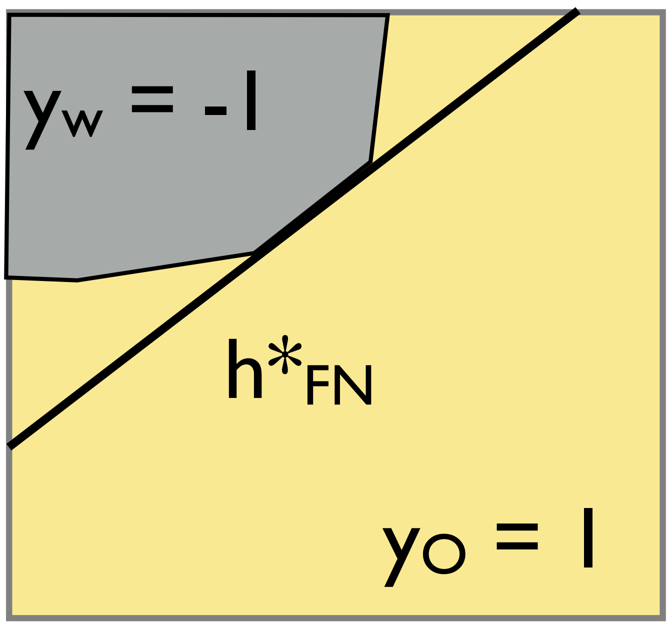


Actual Labels

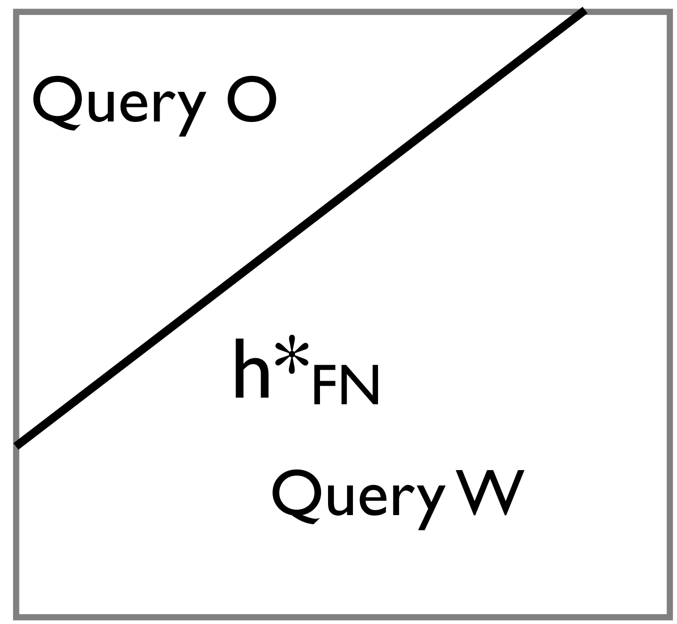
Solution

Train a **cost-sensitive** difference classifier

Constrain False Negative (FN) rate as very low



Actual Labels

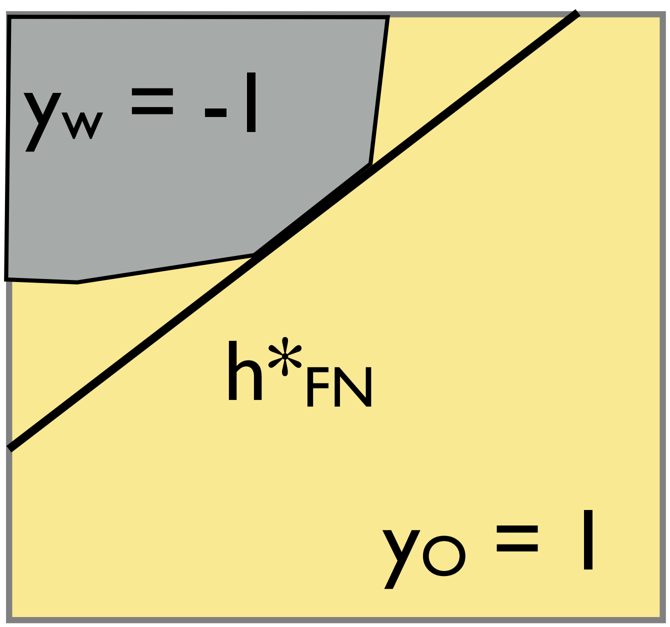


Annotation using h^*_{FN} as difference classifier

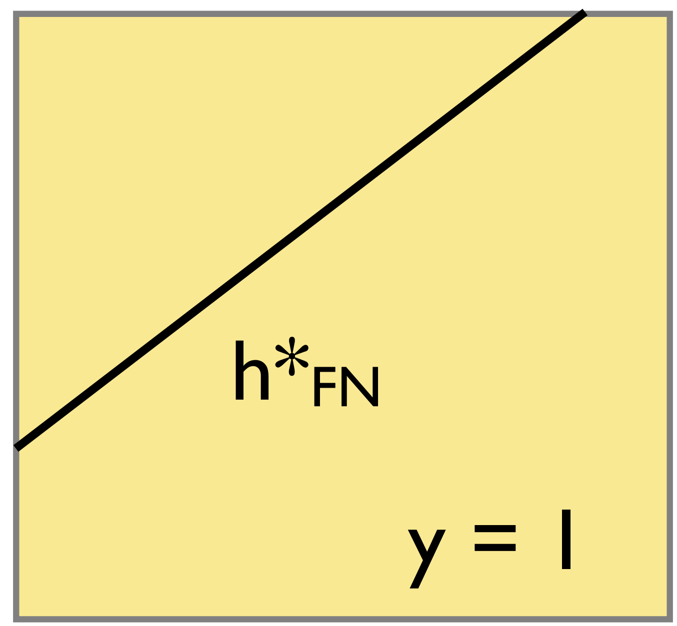
Solution

Train a **cost-sensitive** difference classifier

Constrain False Negative (FN) rate as very low



Actual Labels



Annotation using h^*_{FN} as difference classifier

Algorithm Outline

1. Draw $x_1,..,x_m$. For each x_i, query O and W. Set:

$$y_{i,D} = 1 \quad \text{if} \quad y_{i,O} \neq y_{i,W}$$

2. Train difference classifier h in H on $\{ (x_i, y_{i,D}) \}$ with false negative (FN) rate $\leq \epsilon$

3. Run standard disagreement based active learning algorithm A. If A queries the label of x then:

   if $h(x) = 1$, query O, else query W

Theorem: This is statistically consistent

Talk Outline

1. Weak and Strong Labelers
    - the model
    - algorithm
    - analysis

What about **label complexity**?

Label complexity = #label queries to O

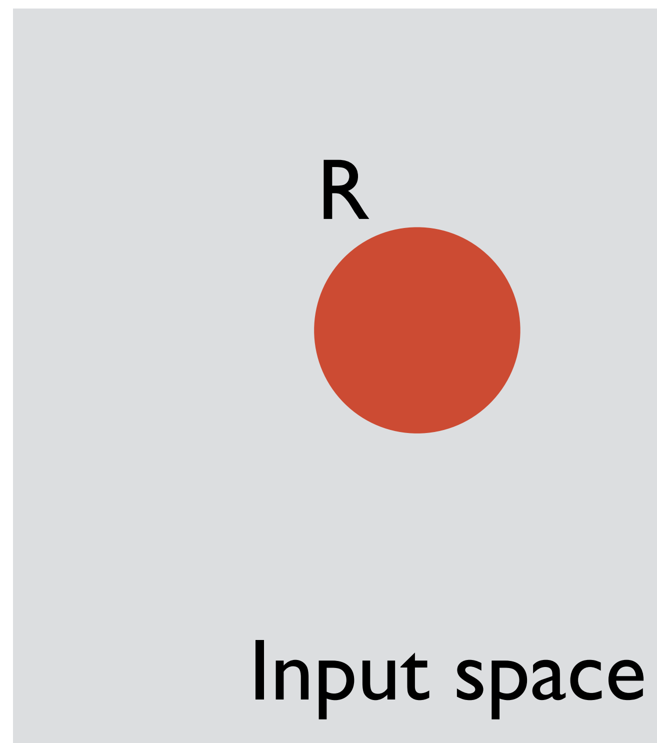
#labels to train difference classifier $\approx \tilde{O}\left(\frac{d'}{\epsilon}\right)$

(d’ = VCdim(H), ϵ = target excess error)

Can we do better?

Key Observation 2

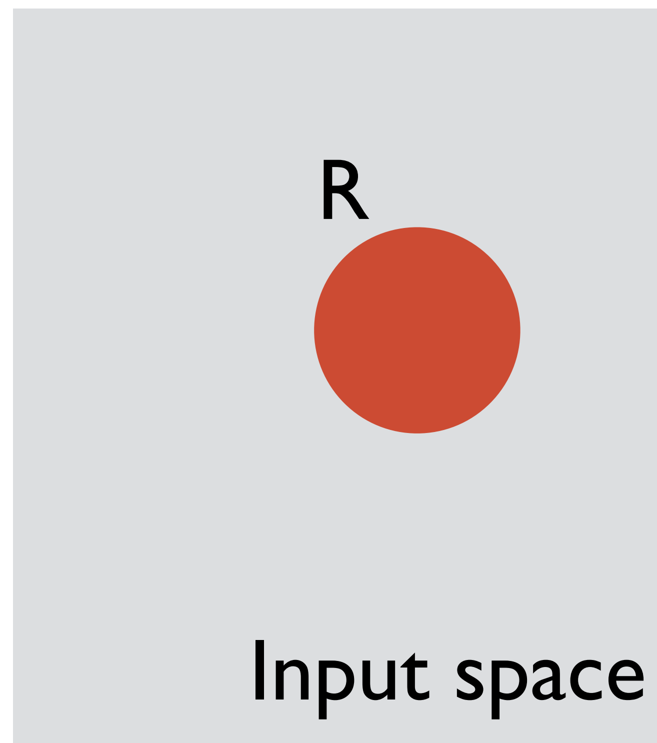
R = disagreement region of current confidence set



Key Observation 2

R = disagreement region of current confidence set

Need to learn difference classifier with FN rate $\leq \epsilon / \Pr(R)$ over R

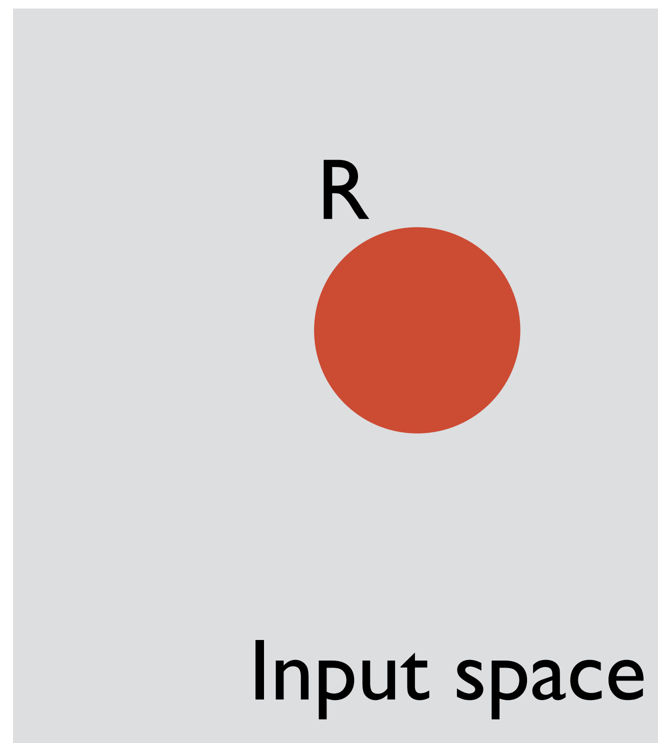


Key Observation 2

R = disagreement region of current confidence set

Need to learn difference classifier with FN rate $\leq \epsilon / \Pr(R)$ over R

Need $\approx \tilde{O}\left(\frac{d' \Pr(R)}{\epsilon}\right)$ labels



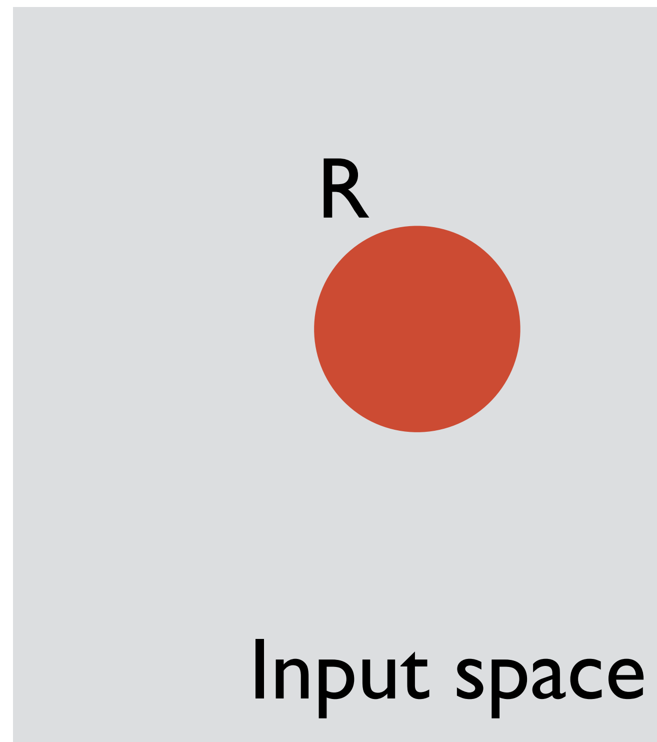
Key Observation 2

R = disagreement region of current confidence set

Need to learn difference classifier with FN rate $\leq \epsilon / \Pr(R)$ over R

Need $\approx \tilde{O}\left(\frac{d' \Pr(R)}{\epsilon}\right)$ labels

Problem: R keeps changing, so have to retrain



Full Algorithm

H = difference concept class, d' = VCdim(H)

For **epochs** 1, 2, 3, ….

Full Algorithm

H = difference concept class, d' = VCdim(H)

For **epochs** 1, 2, 3, ….

Epoch k: target excess error $\epsilon_k \approx 1/2^k$

Confidence set V_k, with disagreement region DIS(V_k)

Full Algorithm

H = difference concept class, d' = VCdim(H)

For **epochs** 1, 2, 3, $\ldots$.

Epoch k: target excess error $\epsilon_k \approx 1/2^k$

Confidence set V_k, with disagreement region DIS(V_k)

Draw $\tilde{O}(d' \Pr(DIS(V_k))/\epsilon_k)$ samples $x_1,..,x_m$ from DIS(V_k).

Query O and W for each x_i and train a difference classifier h.

Full Algorithm

H = difference concept class, d' = VCdim(H)

For **epochs** 1, 2, 3, ….

Epoch k: target excess error $\epsilon_k \approx 1/2^k$

Confidence set V_k, with disagreement region $DIS(V_k)$

Draw $\tilde{O}(d' \Pr(DIS(V_k))/\epsilon_k)$ samples $x_1,..,x_m$ from $DIS(V_k)$.

Query O and W for each x_i and train a difference classifier h.

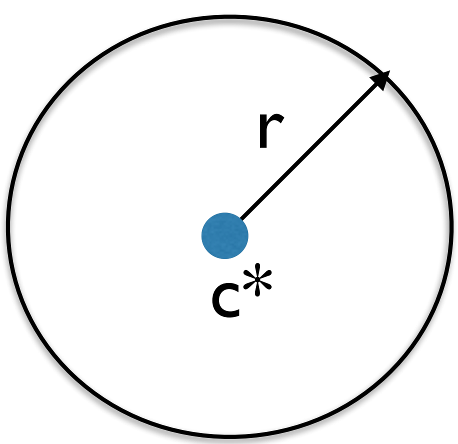
Run disagreement based active learning algorithm A to target excess error ϵ_k. If A queries the label of x then:

if h(x) = 1, query O, else query W

Label Complexity: Definitions

Disagreement Region DIS(V) of a set V:

All x such that there exist c_1 and c_2 in V s.t. $c_1(x) \neq c_2(x)$



$B_D(c^*, r)$

Disagreement Coefficient:

$$\theta(r) = \sup_{r' \geq r} \frac{\Pr(DIS(B_D(c^*, r')))}{r'}$$

(Rate of change of disagreement region as r changes)

Label Complexity

Total #labels to train difference classifier $\approx \tilde{O}\left(\frac{d'\theta(\nu^* + \epsilon)}{\epsilon}\right)$

How many labels for the rest of active learning?

Label Complexity: Assumptions

For any r, t, there is a h in H such that:

$$\Pr(h(x) = -1, x \in DIS(B(c^*, r), y_O \neq y_W) \leq t$$

(Low FN over disagreemt region)

$$\Pr(h(x) = 1, x \in DIS(B(c^*, r)) \leq \alpha(r, t)$$

(Low positives)

Note: $\alpha(r, t) \leq \Pr(DIS(B(c^*, r))$

Label Complexity

#labels to train difference classifier $\approx \tilde{O}\left(\frac{d'\theta(\nu^* + \epsilon)}{\epsilon}\right)$

#labels for active learning $\approx \tilde{O}\left(\frac{d\sigma(\nu^*)^2}{\epsilon^2}\right)$

where: $\sigma \approx \frac{\alpha(2\nu^* + \epsilon, O(\epsilon))}{2\nu^* + \epsilon} \leq \theta$

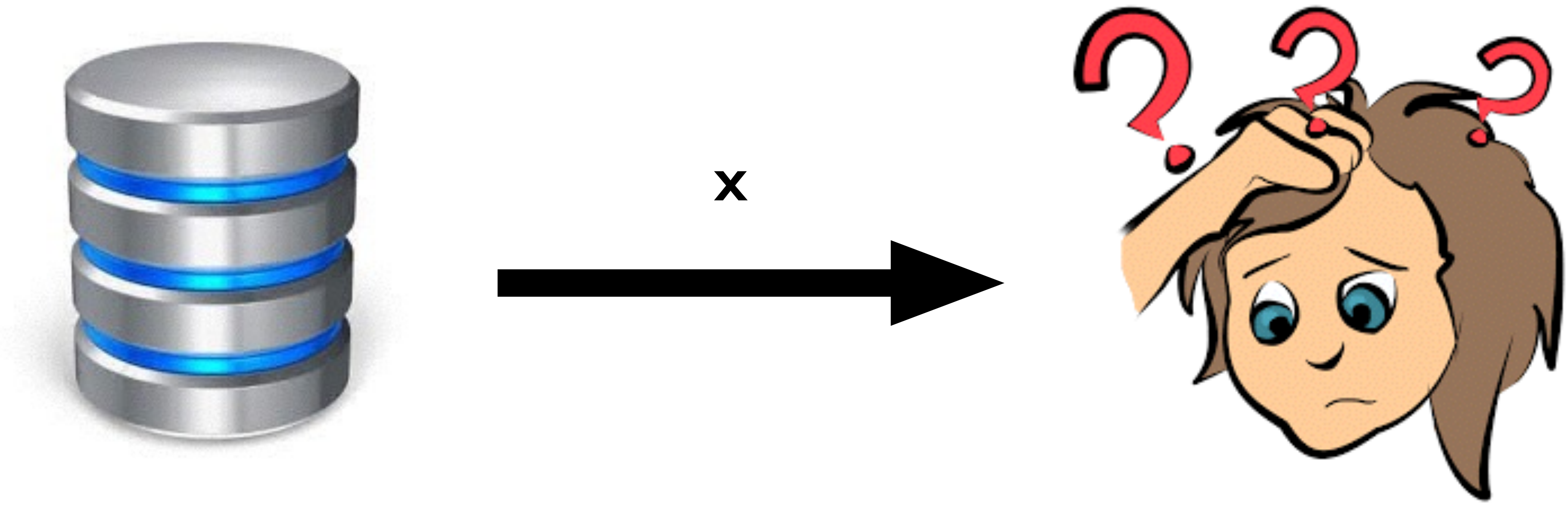
Compare:

#labels for disagreemt based active learning: $\approx \tilde{O}\left(\frac{d\theta(\nu^*)^2}{\epsilon^2}\right)$

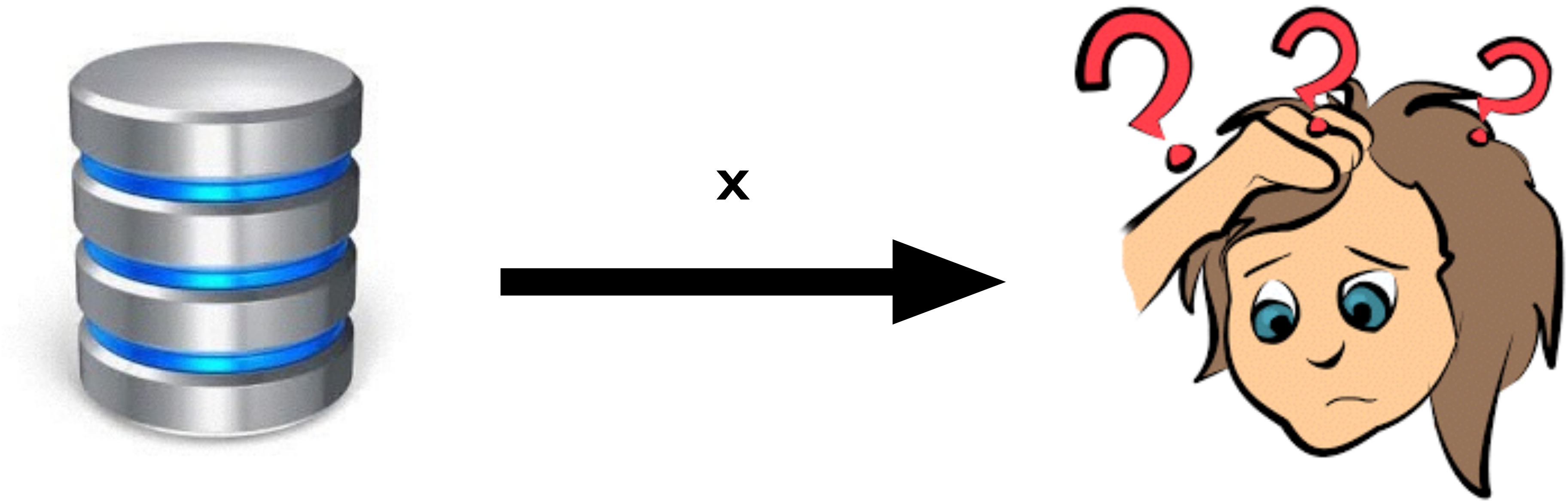
Talk Outline

1. Weak and Strong Labelers
    - the model
    - algorithm
    - analysis

2. Abstentions
    - the model



Labeler abstains on more difficult examples

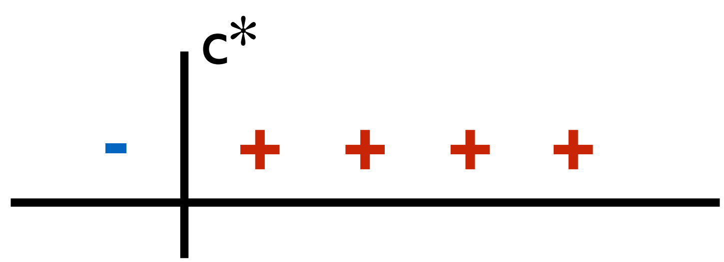


Labeler abstains on more difficult examples

Can we exploit abstentions to learn better?

Example: Learning Thresholds

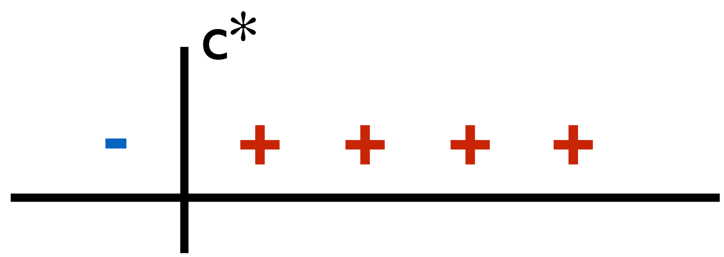
Concept class C = thresholds, instance space X = [0, 1]



Ground truth c*

Example: Learning Thresholds

Concept class C = thresholds, instance space X = [0, 1]



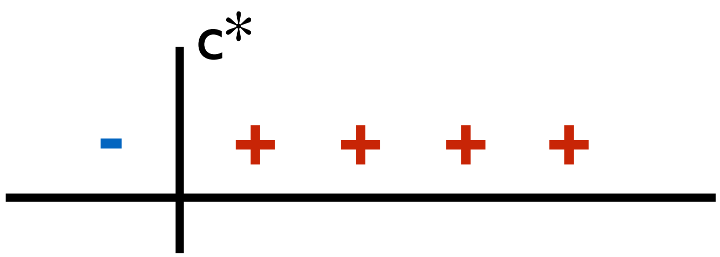
Ground truth c*

Responses

Learner can query any x in X

Example: Learning Thresholds

Concept class C = thresholds, instance space X = [0, 1]



Ground truth c*

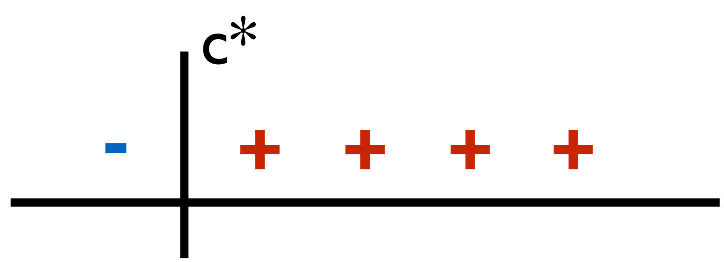
Responses

Learner can query any x in X

Responses +, -, ? drawn from unknown P(Y|x)

Example: Learning Thresholds

Concept class C = thresholds, instance space X = [0, 1]



Ground truth c*

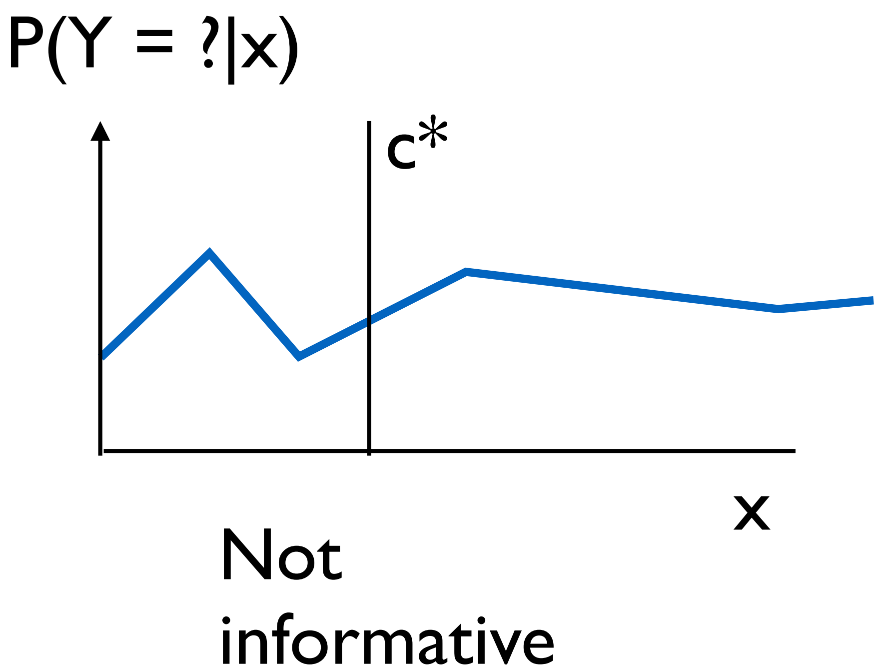
Responses

Learner can query any x in X

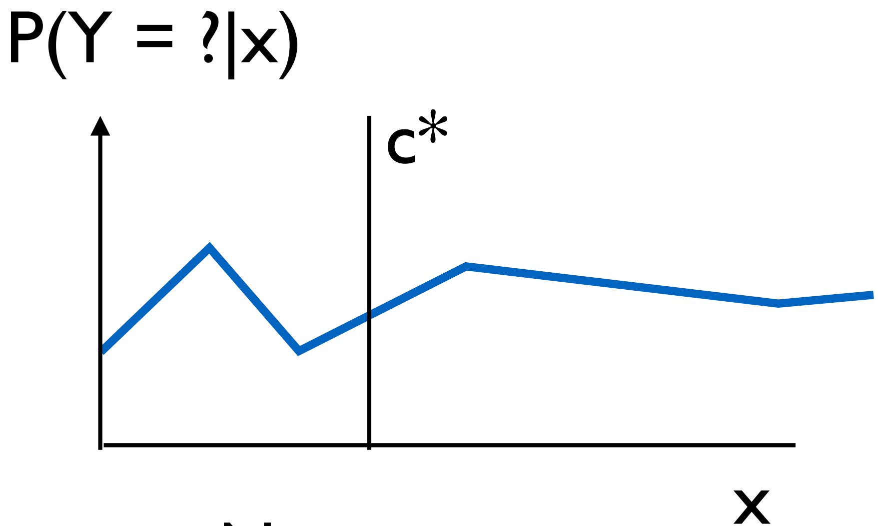
Responses +, -, ? drawn from unknown P(Y|x)

Goal: Find c s.t $|c - c^*| \leq \epsilon$ with min # queries

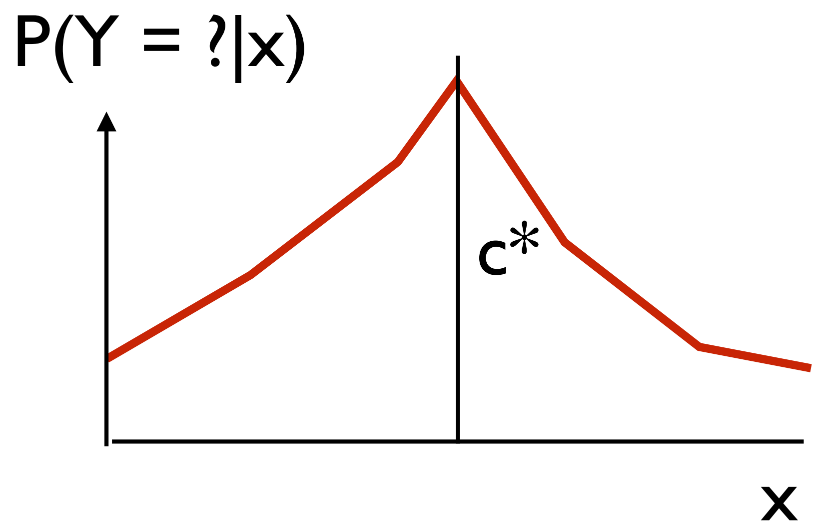
When can abstentions help?



When can abstentions help?

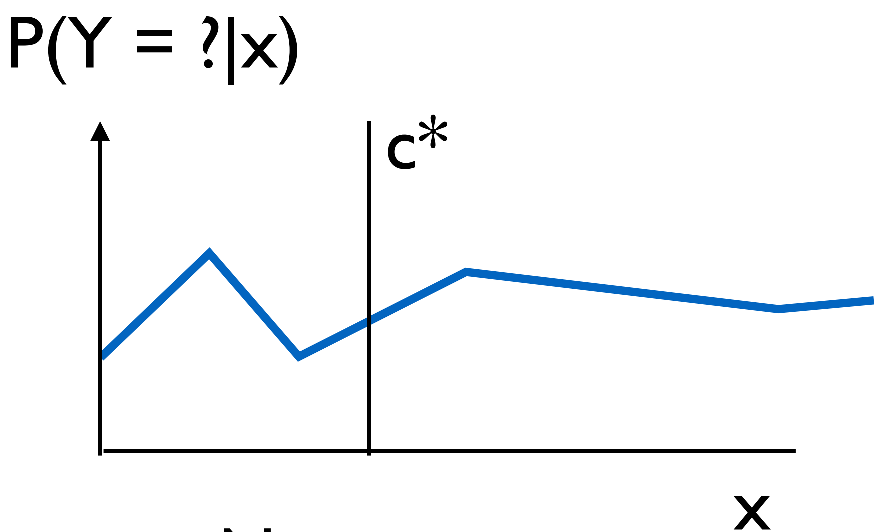


Not
informative

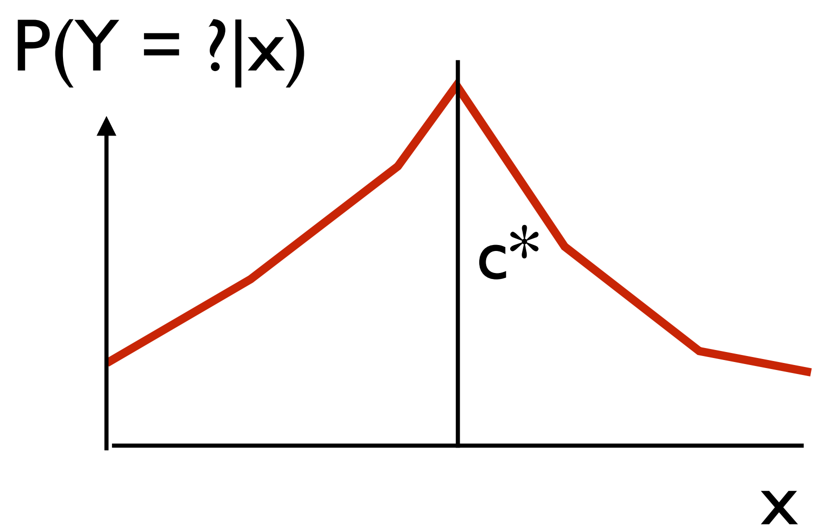


Informative

When can abstentions help?



Not informative



Informative

When abstention rates increase close to decision boundary

Talk Outline

1. Weak and Strong Labelers
    - the model
    - algorithm
    - analysis

2. Abstentions
    - the model
    - algorithm

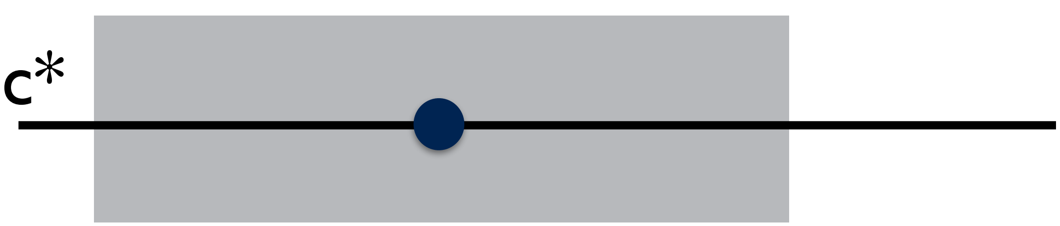
Basic Algorithm: Binary Search

Assume: Correct response

Divide plausible interval containing c* by half per query

Basic Algorithm: Binary Search

Interval
containing c*

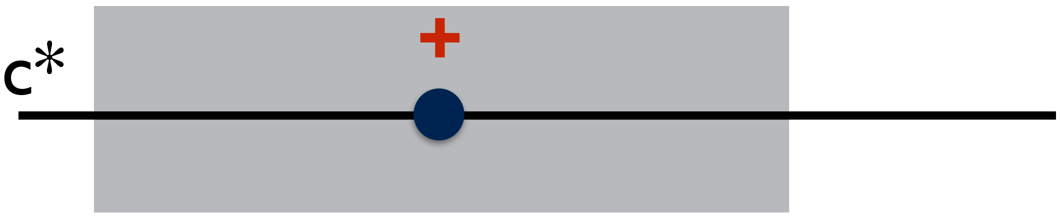


Assume: Correct response

Divide plausible interval containing c* by half per query

Basic Algorithm: Binary Search

Interval
containing c*

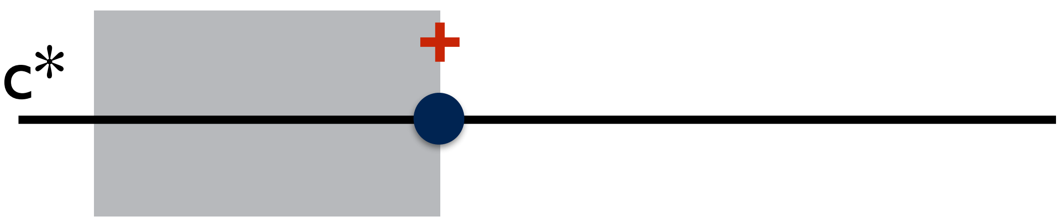


Assume: Correct response

Divide plausible interval containing c* by half per query

Basic Algorithm: Binary Search

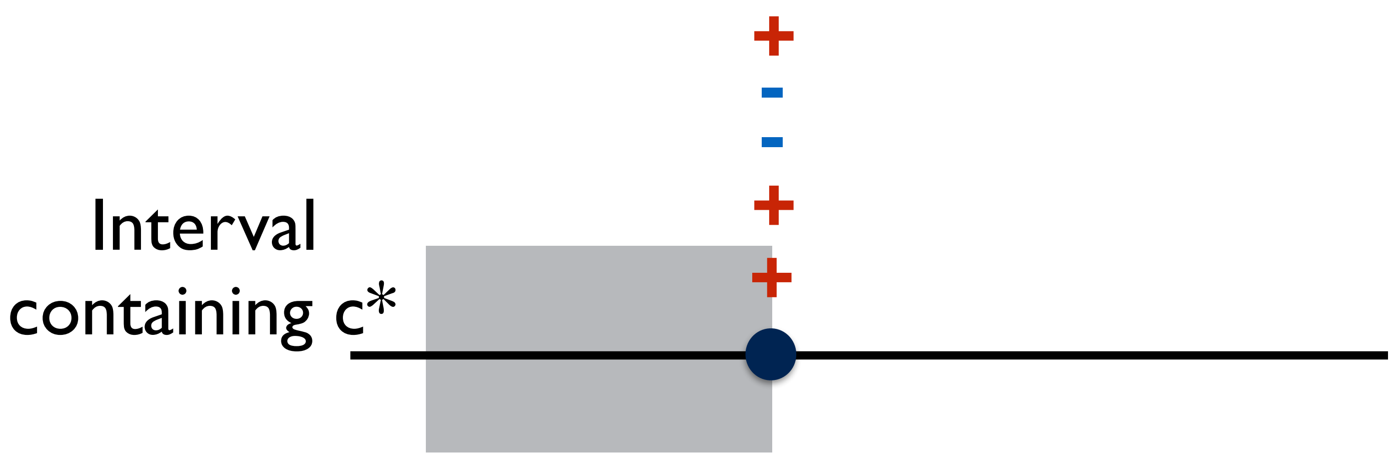
Interval containing c*



Assume: Correct response

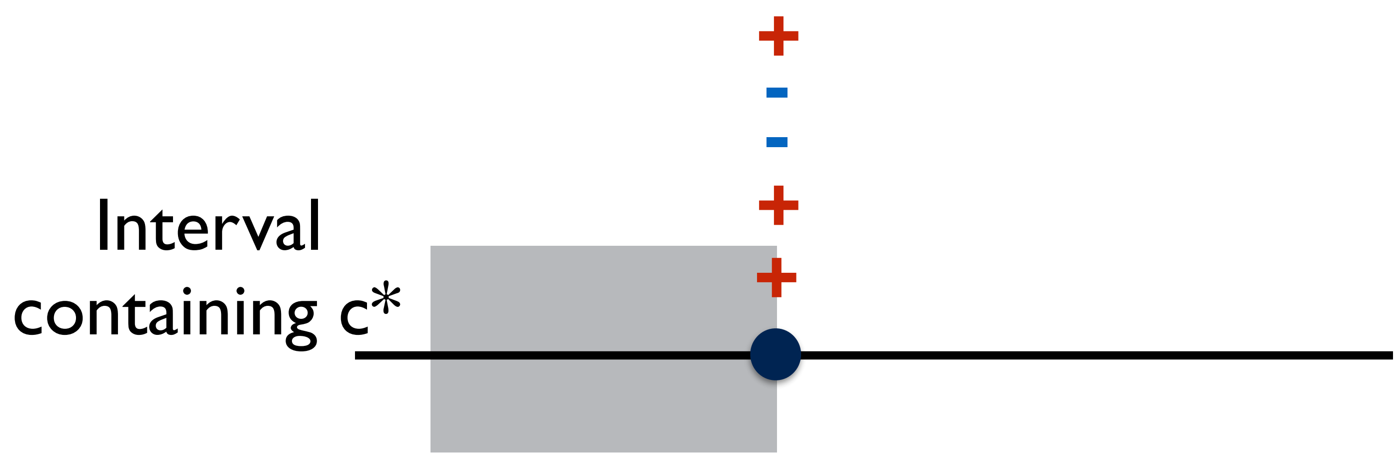
Divide plausible interval containing c* by half per query

Basic Algorithm: Binary Search



Noisy Response: Query multiple times, average to get ground truth label with high confidence

Basic Algorithm: Binary Search

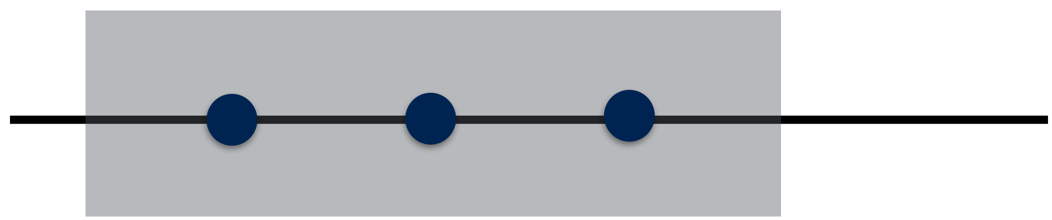


Noisy Response: Query multiple times, average to get ground truth label with high confidence

Increasing noise rate: Make an adaptive #queries till high confidence [BR16]

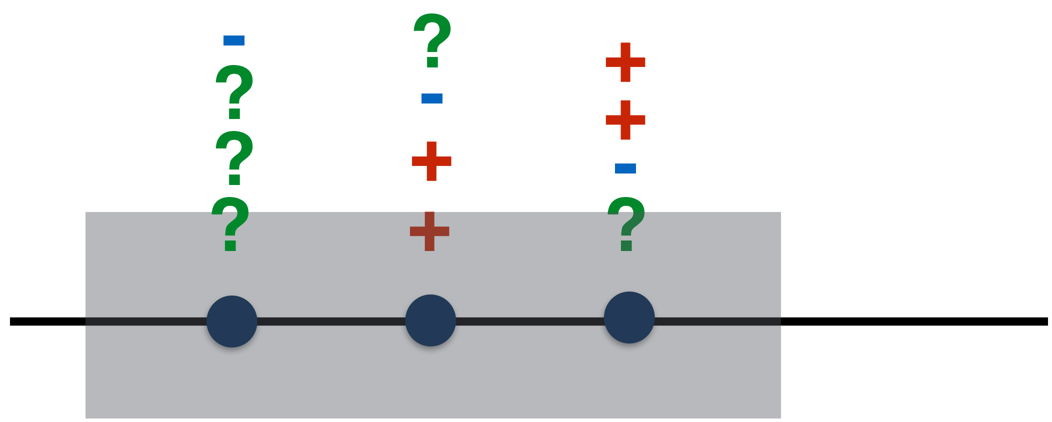
How to handle abstentions?

Modified Binary Search



Query: quartiles of interval

Modified Binary Search

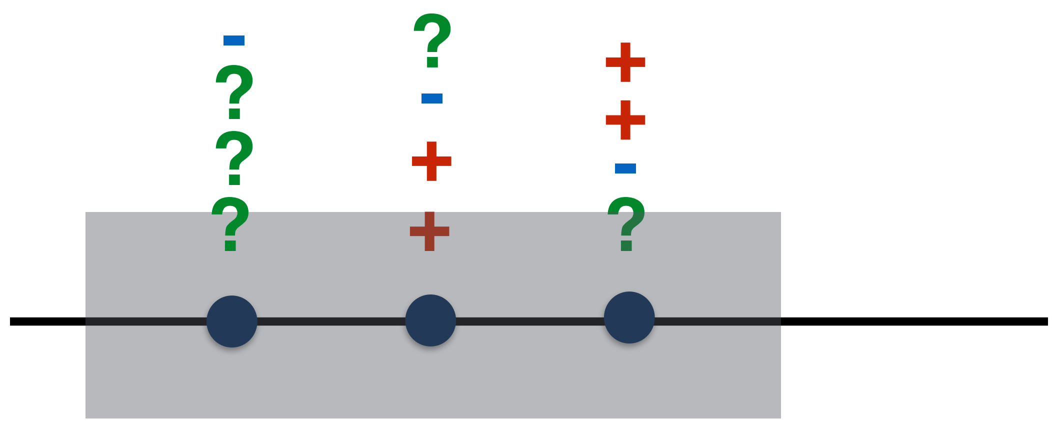


Query: quartiles of interval

After each query, determine if:

- We are confident in the label at any point

Modified Binary Search

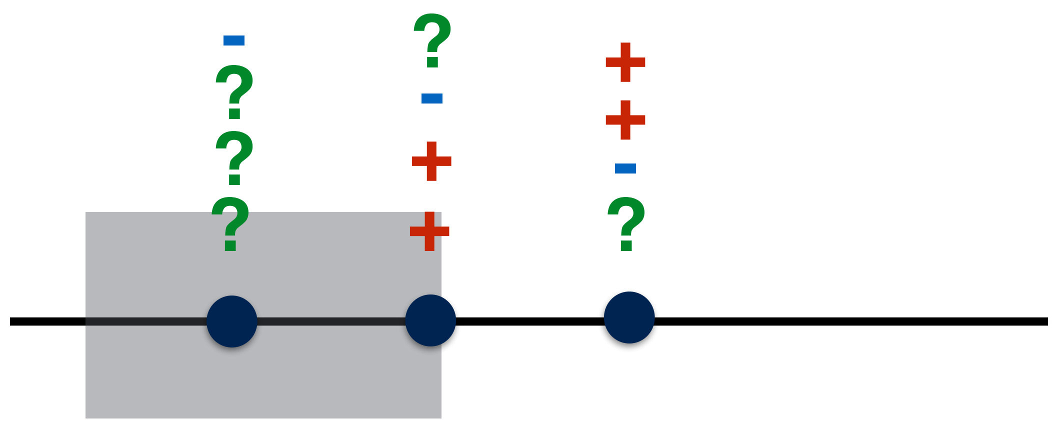


Query: quartiles of interval

After each query, determine if:

- We are confident in the label at any point
- or, if the abstention rate is increasing in some direction

Modified Binary Search



Query: quartiles of interval

After each query, determine if:

- We are confident in the label at any point
- or, if the abstention rate is increasing in some direction

Reduce interval correspondingly

Performance Guarantees

Completely adaptive - algorithm does not know response parameters

Performance Guarantees

Completely adaptive - algorithm does not know response parameters

Statistically consistent so long as abstention rate does not decrease closer to boundary

Performance Guarantees

Completely adaptive - algorithm does not know response parameters

Statistically consistent so long as abstention rate does not decrease closer to boundary

What about #queries?

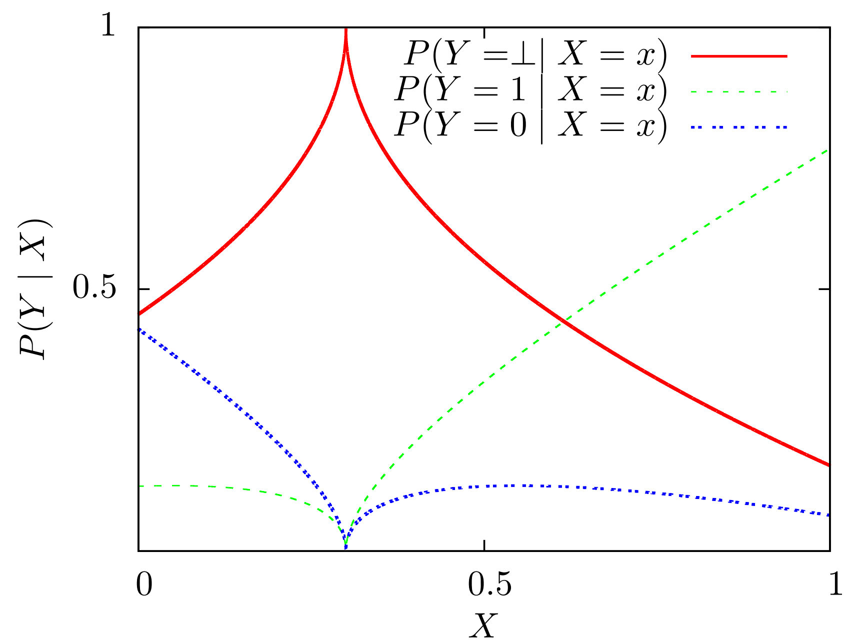
Example: An Informative Response Model

Response Model:

$$\Pr(Y = ?|x) = 1 - C_0|x - c^*|^\alpha$$

$$\Pr(Y \neq c^*(x)) \leq \frac{1}{2} - C_1|x - c^*|^\beta$$

$$\alpha, \beta \geq 1$$



#Queries to get $|c - c^*| \leq \epsilon$

$O(\epsilon^{-\alpha})$ (our method)

$O(\epsilon^{-\alpha-2\beta})$ (use only labels)

Summary

Abstentions may help if rate of abstentions increase close to decision boundary

Algorithms for thresholds and smooth boundary fragments [CN08]

Work in Progress: PAC model

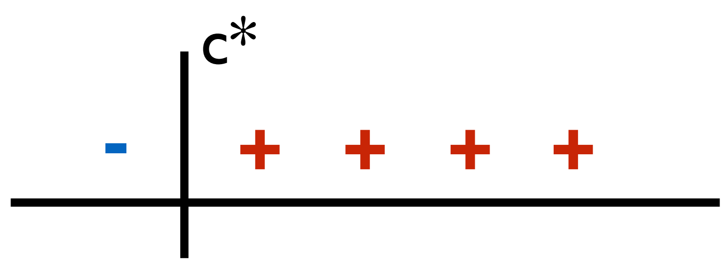
Conclusion

- More complex feedback helps active learning under certain conditions
- Need more sophisticated algorithms

Thank You!

Example: Learning Thresholds

Concept class C = thresholds, instance space X = [0, 1]



Ground truth c*

Responses

Learner can query any x in X

Responses +, -, ? drawn from unknown P(Y|x)

Goal: Find c close to c* with min #queries