Interactive learning of classifiers and other structures

Part I: Sanjoy Dasgupta
Part II: Rob Nowak

Simons Institute program in Foundations of ML

What is interactive learning?

The generic process of supervised learning:

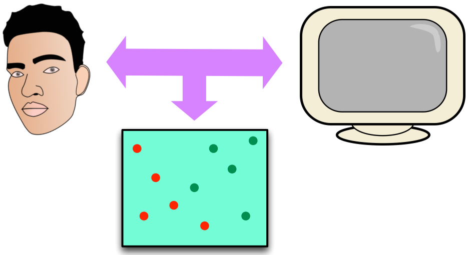
- A data set is obtained.
- A human labels this data set.
- The human goes away.
- A machine looks at the labeled data and chooses a classifier.

What is interactive learning?

The generic process of supervised learning:

- A data set is obtained.
- A human labels this data set.
- The human goes away.
- A machine looks at the labeled data and chooses a classifier.

Interactive learning: the learning machine engages adaptively with an information source (e.g. human) during learning



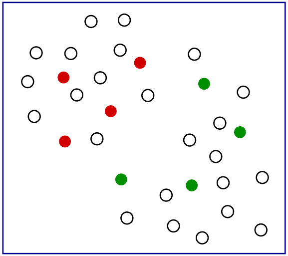
Example: learning a classifier via label queries

Unlabeled data is often plentiful and cheap: documents off the web, speech samples, images, video. *But labeling can be expensive.*

Example: learning a classifier via label queries

Unlabeled data is often plentiful and cheap: documents off the web, speech samples, images, video. *But labeling can be expensive.*

“Active learning”: Machine queries just a few labels, choosing wisely and adaptively.



- Good querying schemes?
- Tradeoff between # labels and error rate of final classifier?

Example: explanation-based learning

In addition to labels, the human might provide an explanation, for instance in the form of relevant features.

Example: explanation-based learning

In addition to labels, the human might provide an explanation, for instance in the form of relevant features.

HOW FINISHES ENGLAND

NZ JUST TOO GOOD

England's hopes of a Test World Medal title are all but over. Following New Zealand's convincing win at Lords, England now have to hope India beat the Black Caps in the first test in New Zealand, and then crush India in one match at Kolkata. Given how well India have been playing lately, it will be a tough task for England to have a comprehensive victory, particularly in India.

The reason for this is New Zealand's resolute, assured performance against England which saw them take the match inside four days. Martin Guptill, Tim Southee, Kane Williamson and Jamie How were the outstanding performers- Southee picking up the Npower Man of the Match Award. Only Pietersen made something of a start for England, with no bowler taking more than 4 wickets for England.

It was a match in which several players put up their hands when the experienced players did not perform as well as they would have liked to. The likes of Vettori, Oram, Bond and McCullum played their part in the match, but it was players like Guptill, How, Southee and Williamson who were the stars. Having always looked the part in test cricket, How and Guptill finally have some runs to show for their talent. Southee has always enjoyed bowling in and against England, and here he made the world sit up and take notice with a fine bowling performance in the 2nd Innings, showing resilience after a disappointing 1st innings.

As for Williamson, expect big things from this boys in the future- he played with outstanding flair and yet great maturity.

Unfortunately, Pietersen was the only performer for England, although Prior did keep and bat solidly. Again, it was a game of missed opportunities.

Example: explanation-based learning

In addition to labels, the human might provide an explanation, for instance in the form of relevant features.

HOW FINISHES ENGLAND

NZ JUST TOO GOOD

England's hopes of a Test World Medal title are all but over. Following New Zealand's convincing win at Lords, England now have to hope India beat the Black Caps in the first test in New Zealand, and then crush India in one match at Kolkata. Given how well India have been playing lately, it will be a tough task for England to have a comprehensive victory, particularly in India.

The reason for this is New Zealand's resolute, assured performance against England which saw them take the match inside four days. Martin Guptill, Tim Southee, Kane Williamson and Jamie How were the outstanding performers- Southee picking up the Npower Man of the Match Award. Only Pietersen made something of a start for England, with no bowler taking more than 4 wickets for England.

It was a match in which several players put up their hands when the experienced players did not perform as well as they would have liked to. The likes of Vettori, Oram, Bond and McCullum played their part in the match, but it was players like Guptill, How, Southee and Williamson who were the stars. Having always looked the part in test cricket, How and Guptill finally have some runs to show for their talent. Southee has always enjoyed bowling in and against England, and here he made the world sit up and take notice with a fine bowling performance in the 2nd Innings, showing resilience after a disappointing 1st innings.

As for Williamson, expect big things from this boys in the future- he played with outstanding flair and yet great maturity.

Unfortunately, Pietersen was the only performer for England, although Prior did keep and bat solidly. Again, it was a game of missed opportunities.

Example: explanation-based learning

In addition to labels, the human might provide an explanation, for instance in the form of relevant features.

HOW FINISHES ENGLAND

NZ JUST TOO GOOD

England's hopes of a Test World Medal title are all but over. Following New Zealand's convincing win at Lords, England now have to hope India beat the Black Caps in the first test in New Zealand, and then crush India in one match at Kolkata. Given how well India have been playing lately, it will be a tough task for England to have a comprehensive victory, particularly in India.

The reason for this is New Zealand's resolute, assured performance against England which saw them take the match inside four days. Martin Guptill, Tim Southee, Kane Williamson and Jamie How were the outstanding performers- Southee picking up the Npower Man of the Match Award. Only Pietersen made something of a start for England, with no bowler taking more than 4 wickets for England.

It was a match in which several players put up their hands when the experienced players did not perform as well as they would have liked to. The likes of Vettori, Oram, Bond and McCullum played their part in the match, but it was players like Guptill, How, Southee and Williamson who were the stars. Having always looked the part in test cricket, How and Guptill finally have some runs to show for their talent. Southee has always enjoyed bowling in and against England, and here he made the world sit up and take notice with a fine bowling performance in the 2nd Innings, showing resilience after a disappointing 1st innings.

As for Williamson, expect big things from this boys in the future- he played with outstanding flair and yet great maturity.

Unfortunately, Pietersen was the only performer for England, although Prior did keep and bat solidly. Again, it was a game of missed opportunities.

- Benefit of explanations over labels alone?

Example: explanation-based learning

In addition to labels, the human might provide an explanation, for instance in the form of relevant features.

HOW FINISHES ENGLAND

NZ JUST TOO GOOD

England's hopes of a Test World Medal title are all but over. Following New Zealand's convincing win at Lords, England now have to hope India beat the Black Caps in the first test in New Zealand, and then crush India in one match at Kolkata. Given how well India have been playing lately, it will be a tough task for England to have a comprehensive victory, particularly in India.

The reason for this is New Zealand's resolute, assured performance against England which saw them take the match inside four days. Martin Guptill, Tim Southee, Kane Williamson and Jamie How were the outstanding performers- Southee picking up the Npower Man of the Match Award. Only Pietersen made something of a start for England, with no bowler taking more than 4 wickets for England.

It was a match in which several players put up their hands when the experienced players did not perform as well as they would have liked to. The likes of Vettori, Oram, Bond and McCullum played their part in the match, but it was players like Guptill, How, Southee and Williamson who were the stars. Having always looked the part in test cricket, How and Guptill finally have some runs to show for their talent. Southee has always enjoyed bowling in and against England, and here he made the world sit up and take notice with a fine bowling performance in the 2nd Innings, showing resilience after a disappointing 1st innings.

As for Williamson, expect big things from this boys in the future- he played with outstanding flair and yet great maturity.

Unfortunately, Pietersen was the only performer for England, although Prior did keep and bat solidly. Again, it was a game of missed opportunities.

- Benefit of explanations over labels alone?
- How to deal with ambiguity of feedback?

Example: interaction for unsupervised learning

Example: interaction for unsupervised learning

E.g. Machine has a clustering C of data X and wants feedback.

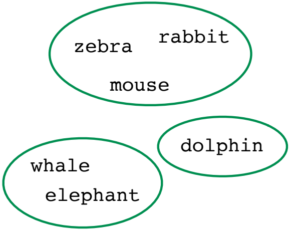
- Show human the restriction of C to $O(1)$ points from X.

Example: interaction for unsupervised learning

E.g. Machine has a clustering C of data X and wants feedback.

- Show human the restriction of C to $O(1)$ points from X.

Flat clustering

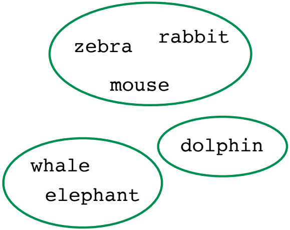


Example: interaction for unsupervised learning

E.g. Machine has a clustering C of data X and wants feedback.

- Show human the restriction of C to $O(1)$ points from X.

Flat clustering



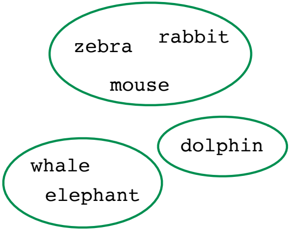
E.g. must-link `dolphin-whale`

Example: interaction for unsupervised learning

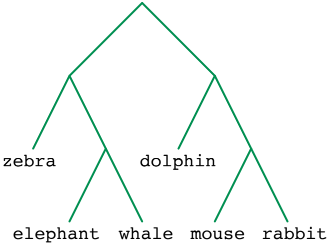
E.g. Machine has a clustering C of data X and wants feedback.

- Show human the restriction of C to $O(1)$ points from X.

Flat clustering



Hierarchical clustering



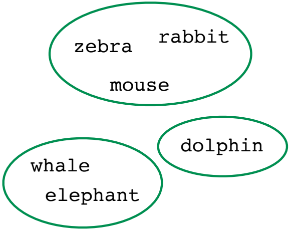
E.g. must-link `dolphin-whale`

Example: interaction for unsupervised learning

E.g. Machine has a clustering C of data X and wants feedback.

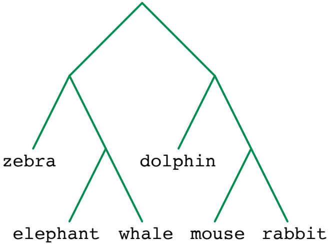
- Show human the restriction of C to $O(1)$ points from X.

Flat clustering



E.g. must-link `dolphin-whale`

Hierarchical clustering



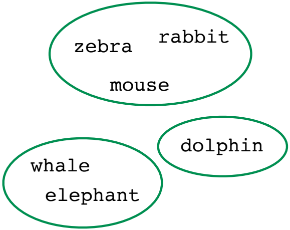
E.g. triplet constraint
$(\{\texttt{dolphin}, \texttt{whale}\}, \texttt{zebra})$

Example: interaction for unsupervised learning

E.g. Machine has a clustering C of data X and wants feedback.

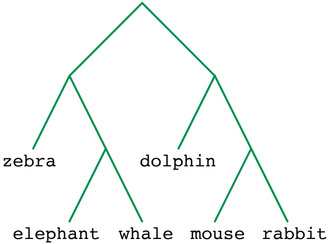
- Show human the restriction of C to $O(1)$ points from X.

Flat clustering



E.g. must-link `dolphin-whale`

Hierarchical clustering



E.g. triplet constraint
$(\{\texttt{dolphin}, \texttt{whale}\}, \texttt{zebra})$

How to choose substructure? How much feedback is needed?

Some questions of interest

1. Efficient interaction algorithms.
   How much interaction is needed to learn?

Some questions of interest

1. Efficient interaction algorithms.
   How much interaction is needed to learn?
2. Interaction versus computational complexity.
   Situations where interaction circumvents computational hardness.

Some questions of interest

1. Efficient interaction algorithms.
   How much interaction is needed to learn?
2. Interaction versus computational complexity.
   Situations where interaction circumvents computational hardness.
3. Modes of interaction.
   - What kinds of interaction are easy and pleasant for the human, and produce reliable feedback?
   - Does it help to have a “don’t know” option?

Some questions of interest

1. Efficient interaction algorithms.
   How much interaction is needed to learn?
2. Interaction versus computational complexity.
   Situations where interaction circumvents computational hardness.
3. Modes of interaction.
   - What kinds of interaction are easy and pleasant for the human, and produce reliable feedback?
   - Does it help to have a "don't know" option?
4. The communication gap between human and machine.

Outline

1. What is interactive learning?
2. Query learning of classifiers
3. Query learning of other structures
4. Interaction in practice

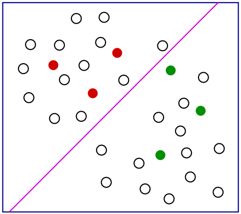
Typical heuristics for "active learning"

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,...)



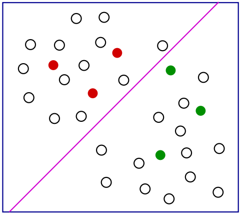
Typical heuristics for "active learning"

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,...)



How to analyze such schemes?

The statistical learning theory framework

Unknown, underlying distribution $\mathbb{P}$ on the (data, label) space.
Hypothesis class $\mathcal{H}$ of candidate classifiers.
Target: the $h^* \in \mathcal{H}$ that has fewest errors on $\mathbb{P}$.

Get n samples from $\mathbb{P}$, choose $h_n \in \mathcal{H}$ that does well on these.

We'd like: $h_n \to h^*$, as rapidly as possible.

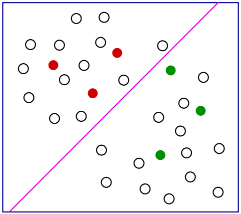
Typical heuristics for "active learning"

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,...)



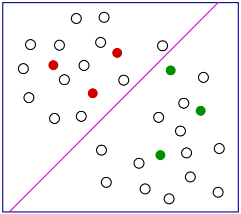
Typical heuristics for "active learning"

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,...)



Biased sampling: the labeled points are not representative of the underlying distribution.

Sampling bias

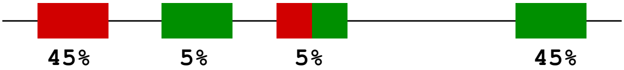
Start with a pool of unlabeled data

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,...)

Example: data in $\mathbb{R}$, $\mathcal{H} = \{\text{thresholds}\}$.



Sampling bias

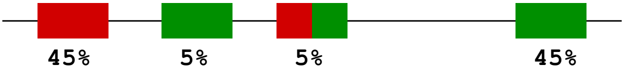
Start with a pool of unlabeled data

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,...)

Example: data in $\mathbb{R}$, $\mathcal{H} = \{\text{thresholds}\}$.



Even with infinitely many labels, converges to a classifier with 5% error instead of the best achievable, 2.5%. *Not consistent.*

Sampling bias

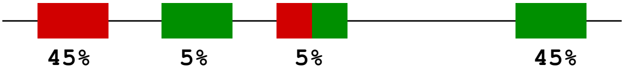
Start with a pool of unlabeled data

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,...)

Example: data in $\mathbb{R}$, $\mathcal{H} = \{\text{thresholds}\}$.



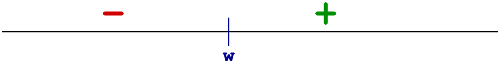
Even with infinitely many labels, converges to a classifier with 5% error instead of the best achievable, 2.5%. *Not consistent.*

Question: Is there a generic fix to uncertainty-based heuristics that makes them consistent?

How much can active learning help?

Threshold functions on the real line ($\mathcal{X} = \mathbb{R}, \mathcal{Y} = \{+1, -1\}$):

$\mathcal{H} = \{h_w : w \in \mathbb{R}\}$
$h_w(x) = 1(x \geq w)$

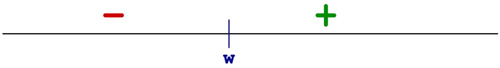


Supervised: for misclassification error $\leq \epsilon$, need $\approx 1/\epsilon$ labeled points.

How much can active learning help?

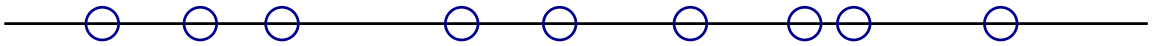
Threshold functions on the real line ($\mathcal{X} = \mathbb{R}, \mathcal{Y} = \{+1, -1\}$):

$\mathcal{H} = \{h_w : w \in \mathbb{R}\}$
$h_w(x) = 1(x \geq w)$



Supervised: for misclassification error $\leq \epsilon$, need $\approx 1/\epsilon$ labeled points.

Active learning: instead, start with $1/\epsilon$ *unlabeled* points.

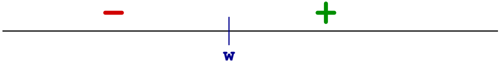


Binary search: need just $\log 1/\epsilon$ labels, from which the rest can be inferred. *Exponential improvement in label complexity*.

How much can active learning help?

Threshold functions on the real line ($\mathcal{X} = \mathbb{R}, \mathcal{Y} = \{+1, -1\}$):

$\mathcal{H} = \{h_w : w \in \mathbb{R}\}$
$h_w(x) = 1(x \geq w)$



Supervised: for misclassification error $\leq \epsilon$, need $\approx 1/\epsilon$ labeled points.

Active learning: instead, start with $1/\epsilon$ *unlabeled* points.

Binary search: need just $\log 1/\epsilon$ labels, from which the rest can be inferred. *Exponential improvement in label complexity*.

What about other hypothesis classes?

Generalized binary search?

For supervised learning of a hypothesis class $\mathcal{H}$ of VC dimension d, we need about d/ϵ labeled points.

Generalized binary search?

For supervised learning of a hypothesis class $\mathcal{H}$ of VC dimension d, we need about d/ϵ labeled points.

- Start with d/ϵ unlabeled points.
- At most $(d/\epsilon)^d$ different ways to classify these using $\mathcal{H}$.

Generalized binary search?

For supervised learning of a hypothesis class $\mathcal{H}$ of VC dimension d, we need about d/ϵ labeled points.

- Start with d/ϵ unlabeled points.
- At most $(d/\epsilon)^d$ different ways to classify these using $\mathcal{H}$.
- Ask queries that cut this space in half each time.
- Then just $d \log(d/\epsilon)$ queries are needed.

Generalized binary search?

For supervised learning of a hypothesis class $\mathcal{H}$ of VC dimension d, we need about d/ϵ labeled points.

- Start with d/ϵ unlabeled points.
- At most $(d/\epsilon)^d$ different ways to classify these using $\mathcal{H}$.
- Ask queries that cut this space in half each time.
- Then just $d \log(d/\epsilon)$ queries are needed.

Problems:

- Halving queries might not exist.
- Computational complexity of maintaining the version space.
- What if there is no classifier with zero error?

Generalized binary search?

For supervised learning of a hypothesis class $\mathcal{H}$ of VC dimension d, we need about d/ϵ labeled points.

- Start with d/ϵ unlabeled points.
- At most $(d/\epsilon)^d$ different ways to classify these using $\mathcal{H}$.
- Ask queries that cut this space in half each time.
- Then just $d \log(d/\epsilon)$ queries are needed.

Problems:

- Halving queries might not exist.
- Computational complexity of maintaining the version space.
- What if there is no classifier with zero error?

Several methods: variants of greedy (Bilmes-Guillory, D, Golovin-Krause, Nowak), query-by-committee (Freund-Shamir-Sompolinsky-Tishby), ...

Three types of active learning results

1. Mellow active learning.
2. Margin-based active learning.
3. Active annotation.

A mellow active learner (Cohn-Atlas-Ladner)

For *separable* data that is streaming in.

$\mathcal{H}_1$ = hypothesis class

Repeat for $t = 1, 2, \ldots$

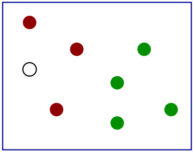
Receive unlabeled point $x_t \in \mathcal{X}$

If there is any disagreement within $\mathcal{H}_t$ about x_t's label:

query label y_t and set $\mathcal{H}_{t+1} = \{h \in \mathcal{H}_t : h(x_t) = y_t\}$

else

$\mathcal{H}_{t+1} = \mathcal{H}_t$



Is a label needed?

A mellow active learner (Cohn-Atlas-Ladner)

For *separable* data that is streaming in.

$\mathcal{H}_1$ = hypothesis class

Repeat for $t = 1, 2, \ldots$

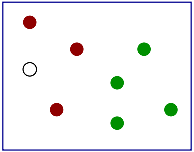
Receive unlabeled point $x_t \in \mathcal{X}$

If there is any disagreement within $\mathcal{H}_t$ about x_t's label:

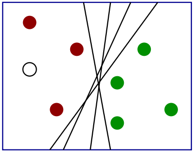
query label y_t and set $\mathcal{H}_{t+1} = \{h \in \mathcal{H}_t : h(x_t) = y_t\}$

else

$\mathcal{H}_{t+1} = \mathcal{H}_t$



Is a label needed?



$\mathcal{H}_t$ = current candidate hypotheses

A mellow active learner (Cohn-Atlas-Ladner)

For *separable* data that is streaming in.

$\mathcal{H}_1$ = hypothesis class

Repeat for $t = 1, 2, \ldots$

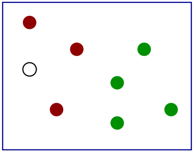
Receive unlabeled point $x_t \in \mathcal{X}$

If there is any disagreement within $\mathcal{H}_t$ about x_t's label:

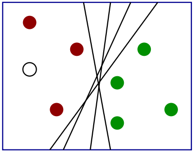
query label y_t and set $\mathcal{H}_{t+1} = \{h \in \mathcal{H}_t : h(x_t) = y_t\}$

else

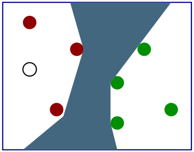
$\mathcal{H}_{t+1} = \mathcal{H}_t$



Is a label needed?



$\mathcal{H}_t$ = current candidate hypotheses



Region of disagreement

A mellow active learner (Cohn-Atlas-Ladner)

For *separable* data that is streaming in.

$\mathcal{H}_1$ = hypothesis class

Repeat for $t = 1, 2, \ldots$

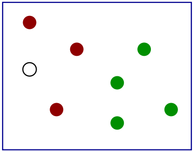
Receive unlabeled point $x_t \in \mathcal{X}$

If there is any disagreement within $\mathcal{H}_t$ about x_t's label:

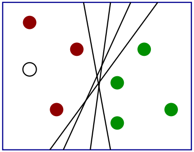
query label y_t and set $\mathcal{H}_{t+1} = \{h \in \mathcal{H}_t : h(x_t) = y_t\}$

else

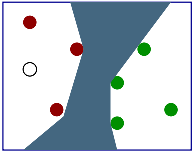
$\mathcal{H}_{t+1} = \mathcal{H}_t$



Is a label needed?



$\mathcal{H}_t$ = current candidate hypotheses



Region of disagreement

No need to explicitly maintain $\mathcal{H}_t$.

Label complexity bounds (Hanneke)

Label complexity can be upper-bounded in terms of:

- the VC dimension d of $\mathcal{H}$
- the **disagreement coefficient** θ, which depends on $\mathcal{H}$ and also on the distribution $\mathbb{P}$ on $\mathcal{X}$

To achieve misclassification rate ϵ w.p. 0.9, suffices to have

$$\#\ \text{labels} \ \approx \ \theta d \log \frac{d}{\epsilon}.$$

Usual supervised requirement: d/ϵ.

Label complexity bounds (Hanneke)

Label complexity can be upper-bounded in terms of:

- the VC dimension d of $\mathcal{H}$
- the **disagreement coefficient** θ, which depends on $\mathcal{H}$ and also on the distribution $\mathbb{P}$ on $\mathcal{X}$

To achieve misclassification rate ϵ w.p. 0.9, suffices to have

$$\# \text{ labels } \approx \theta d \log \frac{d}{\epsilon}.$$

Usual supervised requirement: d/ϵ.

A variety of generalizations to non-separable situations (by various subsets of Balcan, Beygelzimer, Chaudhuri, D, Hanneke, Hsu, Langford, Monteleoni, Zhang, ...).

Label complexity: intuition

$\mathbb{P}$ = underlying distribution on input space $\mathcal{X}$.

- After t points are seen, version space $\mathcal{H}_t$ consists of classifiers with error at most about $\Delta_t = d/t$.
- Let $\mathsf{DIS}(\mathcal{H}_t) \subseteq \mathcal{X}$ be the part of the input space on which there is disagreement within $\mathcal{H}_t$.
  Any point outside $\mathsf{DIS}(\mathcal{H}_t)$ is not queried.
- The *disagreement coefficient* θ tells us $\mathbb{P}(\mathsf{DIS}(\mathcal{H}_t)) \leq \theta\Delta_t$.

Label complexity: intuition

$\mathbb{P}$ = underlying distribution on input space $\mathcal{X}$.

- After t points are seen, version space $\mathcal{H}_t$ consists of classifiers with error at most about $\Delta_t = d/t$.
- Let $\text{DIS}(\mathcal{H}_t) \subseteq \mathcal{X}$ be the part of the input space on which there is disagreement within $\mathcal{H}_t$.
  Any point outside $\text{DIS}(\mathcal{H}_t)$ is not queried.
- The *disagreement coefficient* θ tells us $\mathbb{P}(\text{DIS}(\mathcal{H}_t)) \leq \theta \Delta_t$.
- The expected number of queries, upto time T, is thus:

$$\sum_{t=1}^{T} \mathbb{P}(\text{DIS}(\mathcal{H}_t)) \ \leq\ \theta \sum_{t=1}^{T} \Delta_t \ =\ \theta \sum_{t=1}^{T} \frac{d}{t} \ \approx\ \theta d \log T.$$

- To get error $\leq \epsilon$, take $T \approx d/\epsilon$.

Label complexity: intuition

$\mathbb{P}$ = underlying distribution on input space $\mathcal{X}$.

- After t points are seen, version space $\mathcal{H}_t$ consists of classifiers with error at most about $\Delta_t = d/t$.
- Let $\mathrm{DIS}(\mathcal{H}_t) \subseteq \mathcal{X}$ be the part of the input space on which there is disagreement within $\mathcal{H}_t$.
  Any point outside $\mathrm{DIS}(\mathcal{H}_t)$ is not queried.
- The *disagreement coefficient* θ tells us $\mathbb{P}(\mathrm{DIS}(\mathcal{H}_t)) \leq \theta \Delta_t$.
- The expected number of queries, upto time T, is thus:

$$\sum_{t=1}^{T} \mathbb{P}(\mathrm{DIS}(\mathcal{H}_t)) \;\leq\; \theta \sum_{t=1}^{T} \Delta_t \;=\; \theta \sum_{t=1}^{T} \frac{d}{t} \;\approx\; \theta d \log T.$$

- To get error $\leq \epsilon$, take $T \approx d/\epsilon$.

The disagreement coefficient bounds the probability mass of the region of disagreement in $\mathcal{X}$... how is it defined?

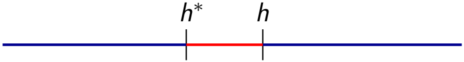
Geometry of hypothesis class

$\mathbb{P}$ = probability distribution on input space $\mathcal{X}$.
Induced pseudo-metric on hypotheses: $d(h, h') = \mathbb{P}[h(X) \neq h'(X)]$.
Corresponding notion of *ball* $B(h, r) = \{h' \in \mathcal{H} : d(h, h') < r\}$.

Geometry of hypothesis class

$\mathbb{P}$ = probability distribution on input space $\mathcal{X}$.
Induced pseudo-metric on hypotheses: $d(h, h') = \mathbb{P}[h(X) \neq h'(X)]$.
Corresponding notion of *ball* $B(h, r) = \{h' \in \mathcal{H} : d(h, h') < r\}$.

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{\text{thresholds}\}$.

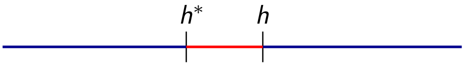


$$d(h^*, h) \;=\; \mathbb{P}[h^*(X) \neq h(X)] \;=\; \text{probability mass of red region}$$

Geometry of hypothesis class

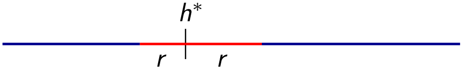
$\mathbb{P}$ = probability distribution on input space $\mathcal{X}$.
Induced pseudo-metric on hypotheses: $d(h, h') = \mathbb{P}[h(X) \neq h'(X)]$.
Corresponding notion of *ball* $B(h, r) = \{h' \in \mathcal{H} : d(h, h') < r\}$.

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{\text{thresholds}\}$.



$$d(h^*, h) \;=\; \mathbb{P}[h^*(X) \neq h(X)] \;=\; \text{probability mass of red region}$$

$B(h^*, r)$ consists of thresholds within probability mass r of h^*:



Disagreement coefficient

Disagreement region of any set of candidate hypotheses $V \subseteq \mathcal{H}$:

$$\mathrm{DIS}(V) = \{x \in \mathcal{X} : \exists h, h' \in V \text{ such that } h(x) \neq h'(x)\}.$$

Need only consider $V = B(h^*, r)$, where $h^* =$ target hypothesis.

Disagreement coefficient

Disagreement region of any set of candidate hypotheses $V \subseteq \mathcal{H}$:

$$\mathrm{DIS}(V) = \{x \in \mathcal{X} : \exists h, h' \in V \text{ such that } h(x) \neq h'(x)\}.$$

Need only consider $V = B(h^*, r)$, where h^* = target hypothesis.

Disagreement coefficient:

$$\theta = \sup_r \frac{\mathbb{P}[\mathrm{DIS}(B(h^*, r))]}{r}.$$

Disagreement coefficient

Disagreement region of any set of candidate hypotheses $V \subseteq \mathcal{H}$:

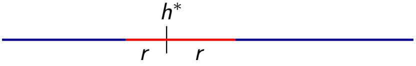
$$\mathrm{DIS}(V) = \{x \in \mathcal{X} : \exists h, h' \in V \text{ such that } h(x) \neq h'(x)\}.$$

Need only consider $V = B(h^*, r)$, where h^* = target hypothesis.

Disagreement coefficient:

$$\theta = \sup_r \frac{\mathbb{P}[\mathrm{DIS}(B(h^*, r))]}{r}.$$

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{\text{thresholds}\}$.
$B(h^*, r)$ consists of thresholds within r probability mass of h^*:



Therefore $\theta = 2$, implying label complexity $O(\log 1/\epsilon)$.

Disagreement coefficient: linear separators

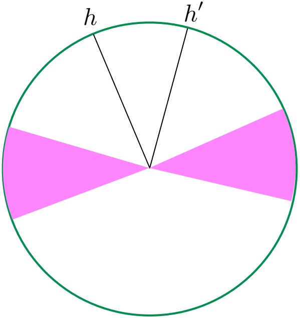
$\mathcal{H}$: through-the-origin linear separators in $\mathbb{R}^d$
$\mathcal{X}$: unit sphere, $\mathbb{P}$: uniform distribution

Then $\theta \leq \sqrt{d}$, implying label complexity $O(d^{3/2} \log d/\epsilon)$.

Disagreement coefficient: linear separators

$\mathcal{H}$: through-the-origin linear separators in $\mathbb{R}^d$
$\mathcal{X}$: unit sphere, $\mathbb{P}$: uniform distribution

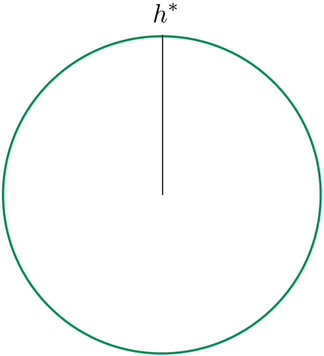
Then $\theta \leq \sqrt{d}$, implying label complexity $O(d^{3/2} \log d/\epsilon)$.



$$d(h, h') = \mathbb{P}(h(X) \neq h'(X)) = \frac{\text{angle between } h, h'}{\pi}.$$

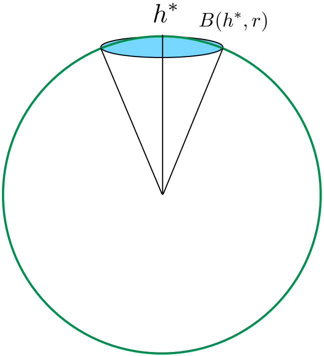
Disagreement coefficient: linear separators

$\mathcal{H}$ = through-the-origin linear separators in $\mathbb{R}^d$ = unit sphere
$\mathcal{X}$: unit sphere, $\mathbb{P}$: uniform distribution



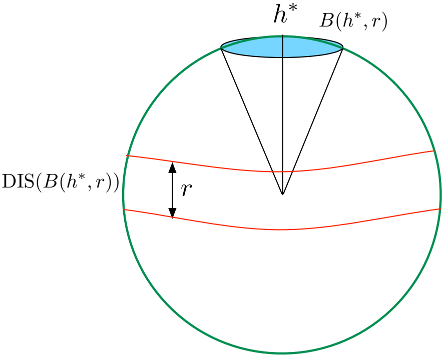
Disagreement coefficient: linear separators

$\mathcal{H}$ = through-the-origin linear separators in $\mathbb{R}^d$ = unit sphere
$\mathcal{X}$: unit sphere, $\mathbb{P}$: uniform distribution



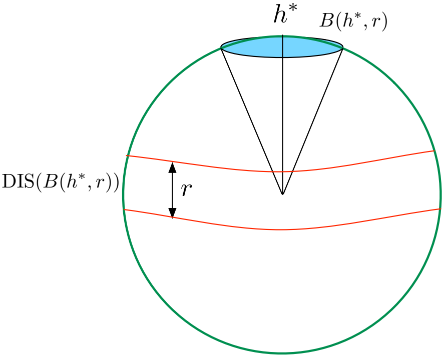
Disagreement coefficient: linear separators

$\mathcal{H}$ = through-the-origin linear separators in $\mathbb{R}^d$ = unit sphere
$\mathcal{X}$: unit sphere, $\mathbb{P}$: uniform distribution



Disagreement coefficient: linear separators

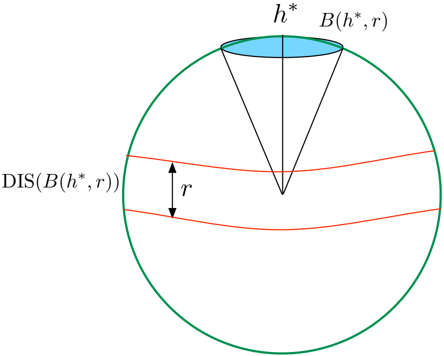
$\mathcal{H}$ = through-the-origin linear separators in $\mathbb{R}^d$ = unit sphere
$\mathcal{X}$: unit sphere, $\mathbb{P}$: uniform distribution



- Uniform distribution on unit sphere $\approx N(0, (1/d)I_d)$

Disagreement coefficient: linear separators

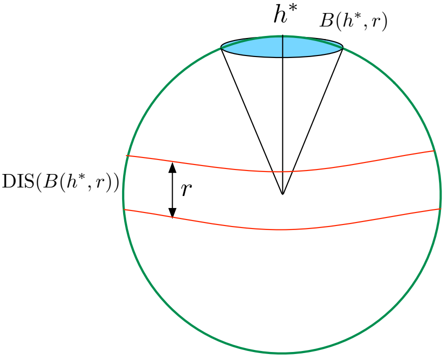
$\mathcal{H}$ = through-the-origin linear separators in $\mathbb{R}^d$ = unit sphere
$\mathcal{X}$: unit sphere, $\mathbb{P}$: uniform distribution



- Uniform distribution on unit sphere $\approx N(0, (1/d)I_d)$
- Marginal in h^*-direction $\approx N(0, 1/d)$

Disagreement coefficient: linear separators

$\mathcal{H}$ = through-the-origin linear separators in $\mathbb{R}^d$ = unit sphere
$\mathcal{X}$: unit sphere, $\mathbb{P}$: uniform distribution



- Uniform distribution on unit sphere $\approx N(0, (1/d)I_d)$
- Marginal in h^*-direction $\approx N(0, 1/d)$
- Therefore $\mathbb{P}(\mathrm{DIS}(h^*, r)) \approx r\sqrt{d} \;\Rightarrow\; \theta \approx \sqrt{d}$.

Three types of active learning results

1. Mellow active learning.
2. Margin-based active learning.
3. Active annotation.

Margin-based active learning (Balcan-Long)

An active learning blueprint for linear separators (D-Kalai-Monteleoni, Balcan-Broder-Zhang, CesaBianchi-Gentile-Orabona, Balcan-Long):

- Let's say all x have $\|x\| = 1$.
- For $t = 1, 2, 3, \ldots$:
  - w_t = classifier based on data so far
  - Randomly choose points amongst those with $|x \cdot w_t| \leq m_t$
  - Query their labels

Here (m_t) is a schedule of margins that decreases to zero.

Margin-based active learning (Balcan-Long)

An active learning blueprint for linear separators (D-Kalai-Monteleoni, Balcan-Broder-Zhang, CesaBianchi-Gentile-Orabona, Balcan-Long):

- Let's say all x have $\|x\| = 1$.
- For $t = 1, 2, 3, \ldots$:
  - w_t = classifier based on data so far
  - Randomly choose points amongst those with $|x \cdot w_t| \leq m_t$
  - Query their labels

Here (m_t) is a schedule of margins that decreases to zero.

Results:

- Yields a classifier of error $\leq \epsilon$ using $O(d \log(1/\epsilon))$ labels if the marginal distribution of x is logconcave and isotropic.
- Can handle a variant of "Tsybakov noise".

Margin-based active learning (Balcan-Long)

An active learning blueprint for linear separators (D-Kalai-Monteleoni, Balcan-Broder-Zhang, CesaBianchi-Gentile-Orabona, Balcan-Long):

- Let's say all x have $\|x\| = 1$.
- For $t = 1, 2, 3, \ldots$:
  - w_t = classifier based on data so far
  - Randomly choose points amongst those with $|x \cdot w_t| \leq m_t$
  - Query their labels

Here (m_t) is a schedule of margins that decreases to zero.

Results:

- Yields a classifier of error $\leq \epsilon$ using $O(d \log(1/\epsilon))$ labels if the marginal distribution of x is logconcave and isotropic.
- Can handle a variant of "Tsybakov noise".

Question: Make this practical while retaining statistical guarantees.

Three types of active learning results

1. Mellow active learning.
2. Margin-based active learning.
3. Active annotation.

Active annotation

Input:

- Finite set of data points $\{x_1, \ldots, x_n\}$, each of which has an associated label y_i that is initially missing.
- Parameters $0 < \delta, \epsilon < 1$.
- Access to an oracle that can supply any label y_i.

Output:
A set of labels $\widehat{y}_1, \ldots, \widehat{y}_n$ such that with probability at least $1 - \delta$, at most an ϵ fraction of these labels are incorrect, that is,

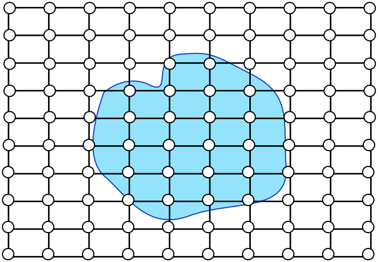
$$\sum_i 1(y_i \neq \widehat{y}_i) \ \leq \ \epsilon n.$$

Goal: Minimize calls to the oracle.

Active learning on graphs

Input: a **neighborhood graph** G whose nodes are the data points x.

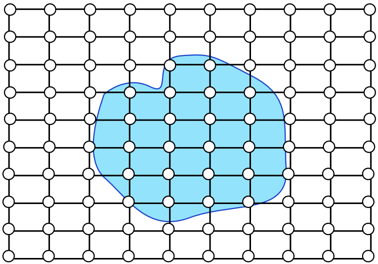
- Each node has an unknown label.
- Goal: find the *cut-edges* in this graph that separate two labels.



Active learning on graphs

Input: a **neighborhood graph** G whose nodes are the data points x.

- Each node has an unknown label.
- Goal: find the *cut-edges* in this graph that separate two labels.



What should label complexity depend upon?

- # cut edges
- log(diameter of graph)
- 1/(proportion of each class)

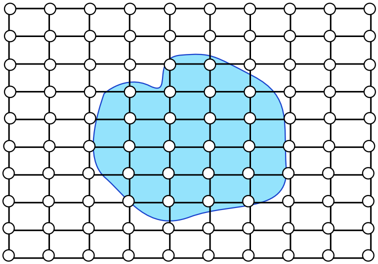
The S^2 algorithm (Dasarthy-Nowak-Zhu)

(For binary labels)

Keep going until budget runs out:

- If $\exists$ labeled nodes of opposite polarity that are connected in G:
  - Find the shortest path connecting nodes of opposite label.
  - Query its midpoint.

  Else:
  - Pick a random point and query it.
- Remove any newly-revealed cut edges from the graph G.



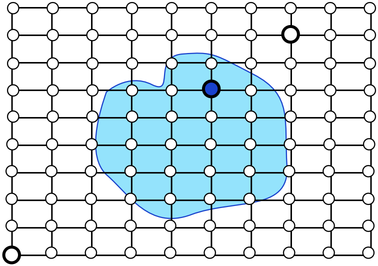
The S^2 algorithm (Dasarthy-Nowak-Zhu)

(For binary labels)

Keep going until budget runs out:

- If $\exists$ labeled nodes of opposite polarity that are connected in G:
  - Find the shortest path connecting nodes of opposite label.
  - Query its midpoint.

  Else:
  - Pick a random point and query it.
- Remove any newly-revealed cut edges from the graph G.



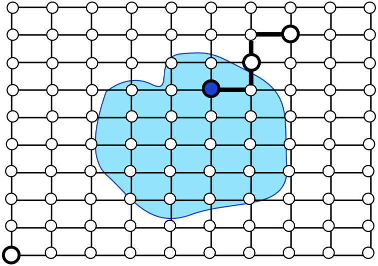
The S^2 algorithm (Dasarthy-Nowak-Zhu)

(For binary labels)

Keep going until budget runs out:

- If $\exists$ labeled nodes of opposite polarity that are connected in G:
  - Find the shortest path connecting nodes of opposite label.
  - Query its midpoint.

  Else:
  - Pick a random point and query it.
- Remove any newly-revealed cut edges from the graph G.



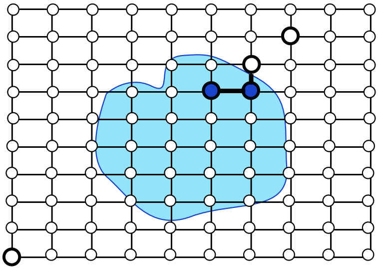
The S^2 algorithm (Dasarthy-Nowak-Zhu)

(For binary labels)

Keep going until budget runs out:

- If $\exists$ labeled nodes of opposite polarity that are connected in G:
  - Find the shortest path connecting nodes of opposite label.
  - Query its midpoint.

  Else:
  - Pick a random point and query it.
- Remove any newly-revealed cut edges from the graph G.



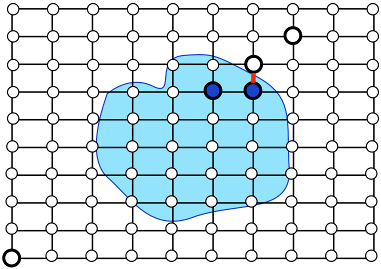
The S^2 algorithm (Dasarthy-Nowak-Zhu)

(For binary labels)

Keep going until budget runs out:

- If $\exists$ labeled nodes of opposite polarity that are connected in G:
  - Find the shortest path connecting nodes of opposite label.
  - Query its midpoint.

  Else:
  - Pick a random point and query it.
- Remove any newly-revealed cut edges from the graph G.



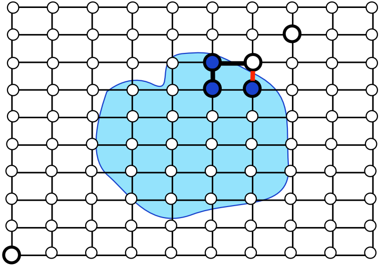
The S^2 algorithm (Dasarthy-Nowak-Zhu)

(For binary labels)

Keep going until budget runs out:

- If $\exists$ labeled nodes of opposite polarity that are connected in G:
  - Find the shortest path connecting nodes of opposite label.
  - Query its midpoint.

  Else:
  - Pick a random point and query it.
- Remove any newly-revealed cut edges from the graph G.



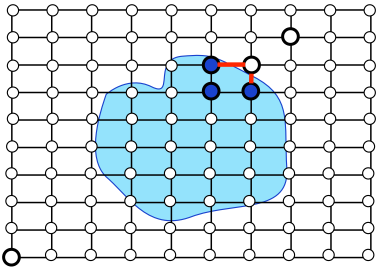
The S^2 algorithm (Dasarthy-Nowak-Zhu)

(For binary labels)

Keep going until budget runs out:

- If $\exists$ labeled nodes of opposite polarity that are connected in G:
  - Find the shortest path connecting nodes of opposite label.
  - Query its midpoint.

  Else:
  - Pick a random point and query it.
- Remove any newly-revealed cut edges from the graph G.



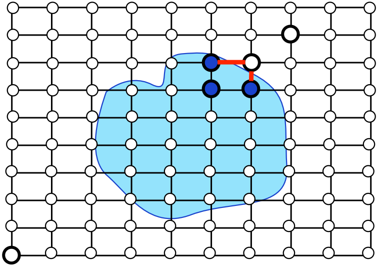
The S^2 algorithm (Dasarthy-Nowak-Zhu)

(For binary labels)

Keep going until budget runs out:

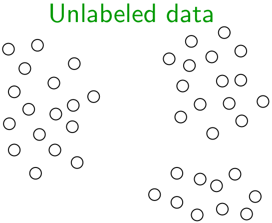
- If $\exists$ labeled nodes of opposite polarity that are connected in G:
  - Find the shortest path connecting nodes of opposite label.
  - Query its midpoint.

  Else:
  - Pick a random point and query it.
- Remove any newly-revealed cut edges from the graph G.



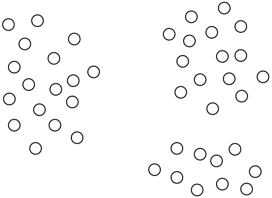
Graph-specific label complexity + nonparametric generalization bounds.

A cluster-based approach (D-Hsu)

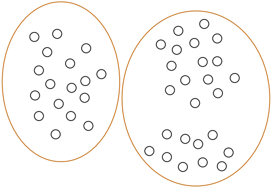


A cluster-based approach (D-Hsu)

Unlabeled data

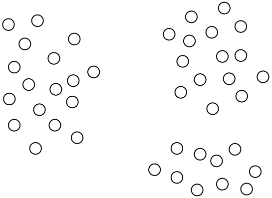


Find a clustering

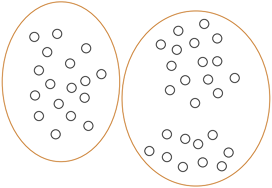


A cluster-based approach (D-Hsu)

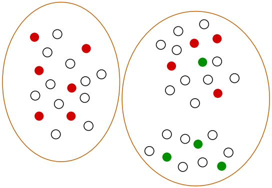
Unlabeled data



Find a clustering



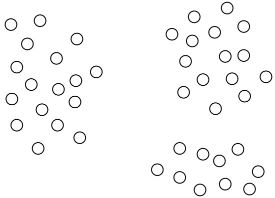
Ask for some labels



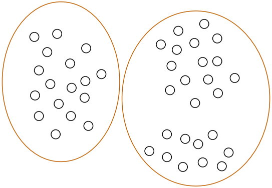
(random sampling within clusters)

A cluster-based approach (D-Hsu)

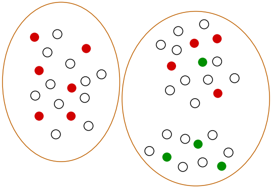
Unlabeled data



Find a clustering



Ask for some labels

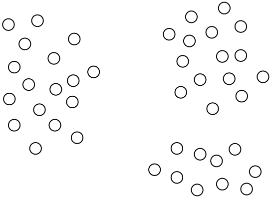


(random sampling within clusters)

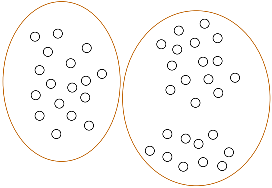
Now what?

A cluster-based approach (D-Hsu)

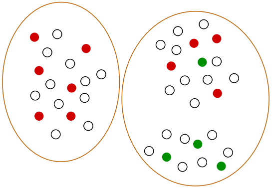
Unlabeled data



Find a clustering



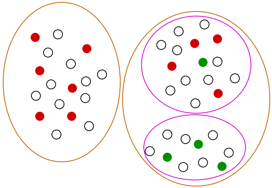
Ask for some labels



(random sampling within clusters)

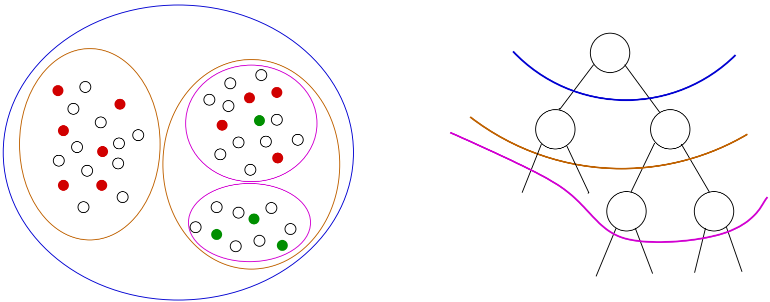
Now what?

Refine the clustering



Queried points are also randomly distributed within the new clusters.

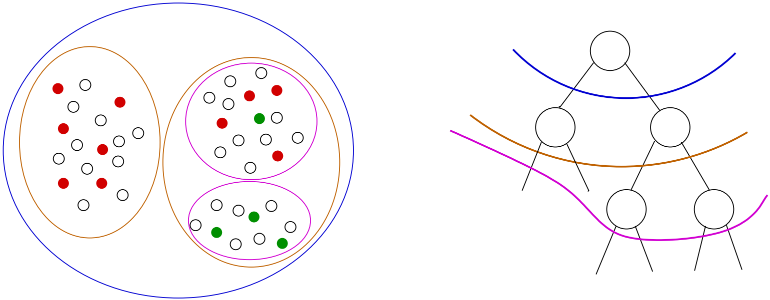
Hierarchical sampling



Rules:

- Always work with some pruning of the hierarchy: a clustering induced by the tree.
- Pick a cluster, query a *random* point in it.
- For each tree node (cluster) maintain majority label and confidence intervals on label frequencies.

Hierarchical sampling



Rules:

- Always work with some pruning of the hierarchy: a clustering induced by the tree.
- Pick a cluster, query a *random* point in it.
- For each tree node (cluster) maintain majority label and confidence intervals on label frequencies.

Ben David-Kpotufe-Urner '14: Label complexity under smoothness.

Three types of active learning results

1. Mellow active learning.
2. Margin-based active learning.
3. Active annotation.