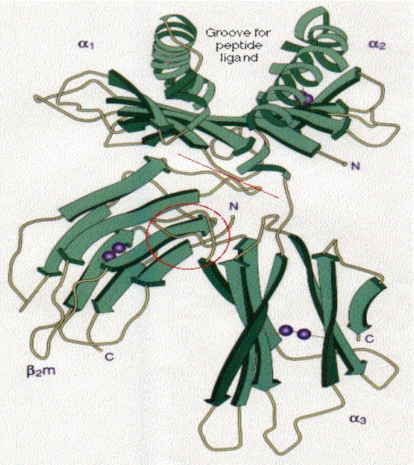
Active Learning

Lectured by Shangsong Liang

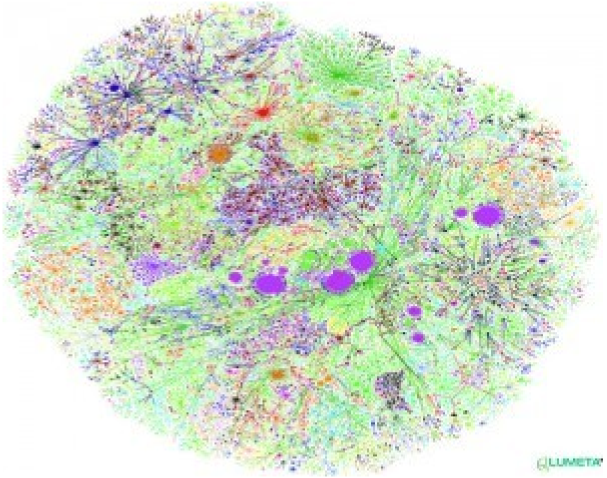
Classic Fully Supervised Learning Paradigm Insufficient Nowadays

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.



Protein sequences



Billions of webpages

Images

Modern ML: New Learning Approaches

Modern applications: massive amounts of raw data.

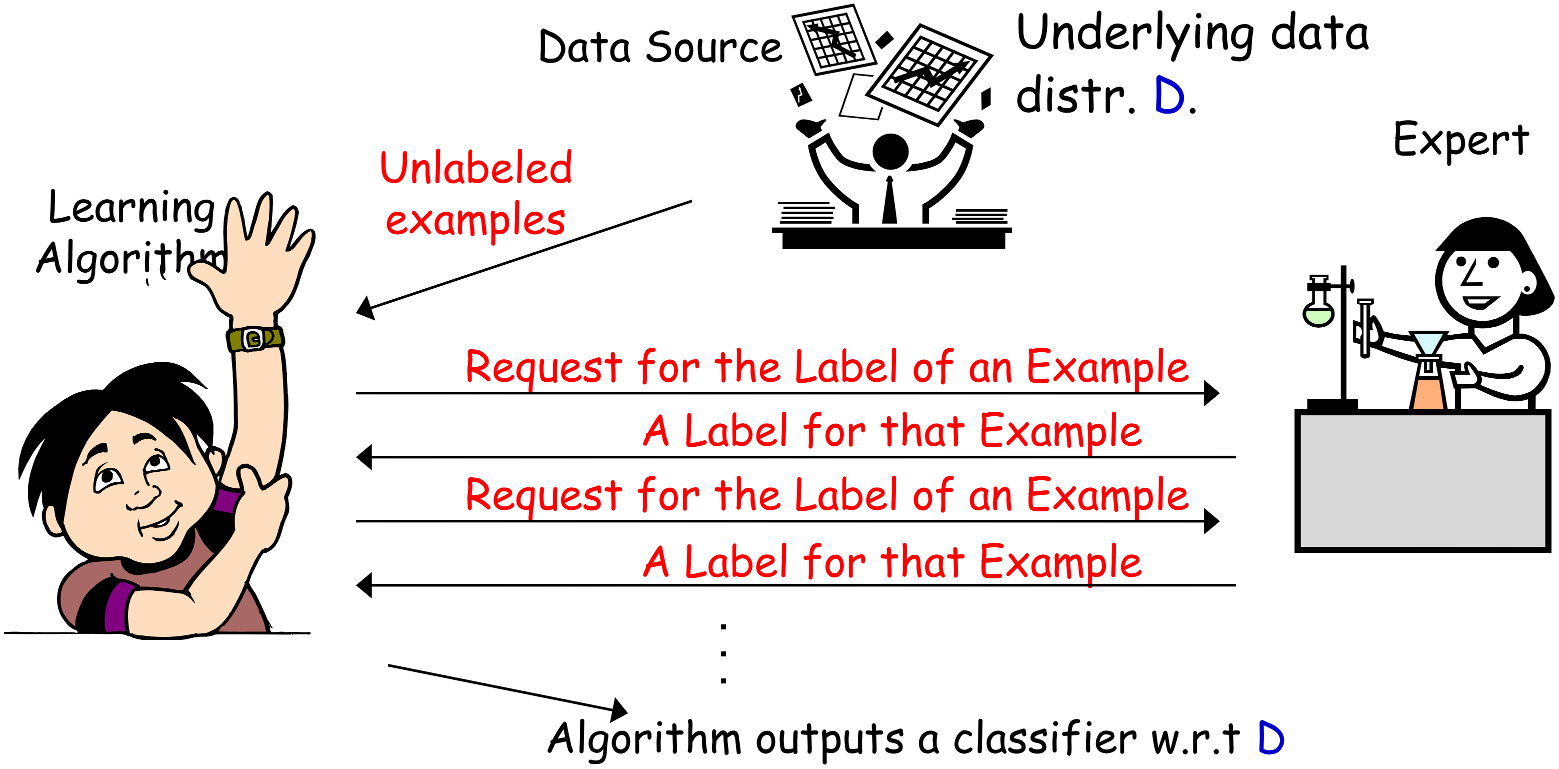
Active learning: techniques that best utilize data, minimizing need for expert/human intervention.

Active Learning

Additional resources:

- Two faces of active learning. Sanjoy Dasgupta. 2011._
- Active Learning. Bur Settles. 2012.
- Active Learning. Balcan-Urner. Encyclopedia of Algorithms. 2015

Batch Active Learning



Data Source

Underlying data distr. D.

Expert

Learning Algorithm

Unlabeled examples

Request for the Label of an Example

A Label for that Example

Request for the Label of an Example

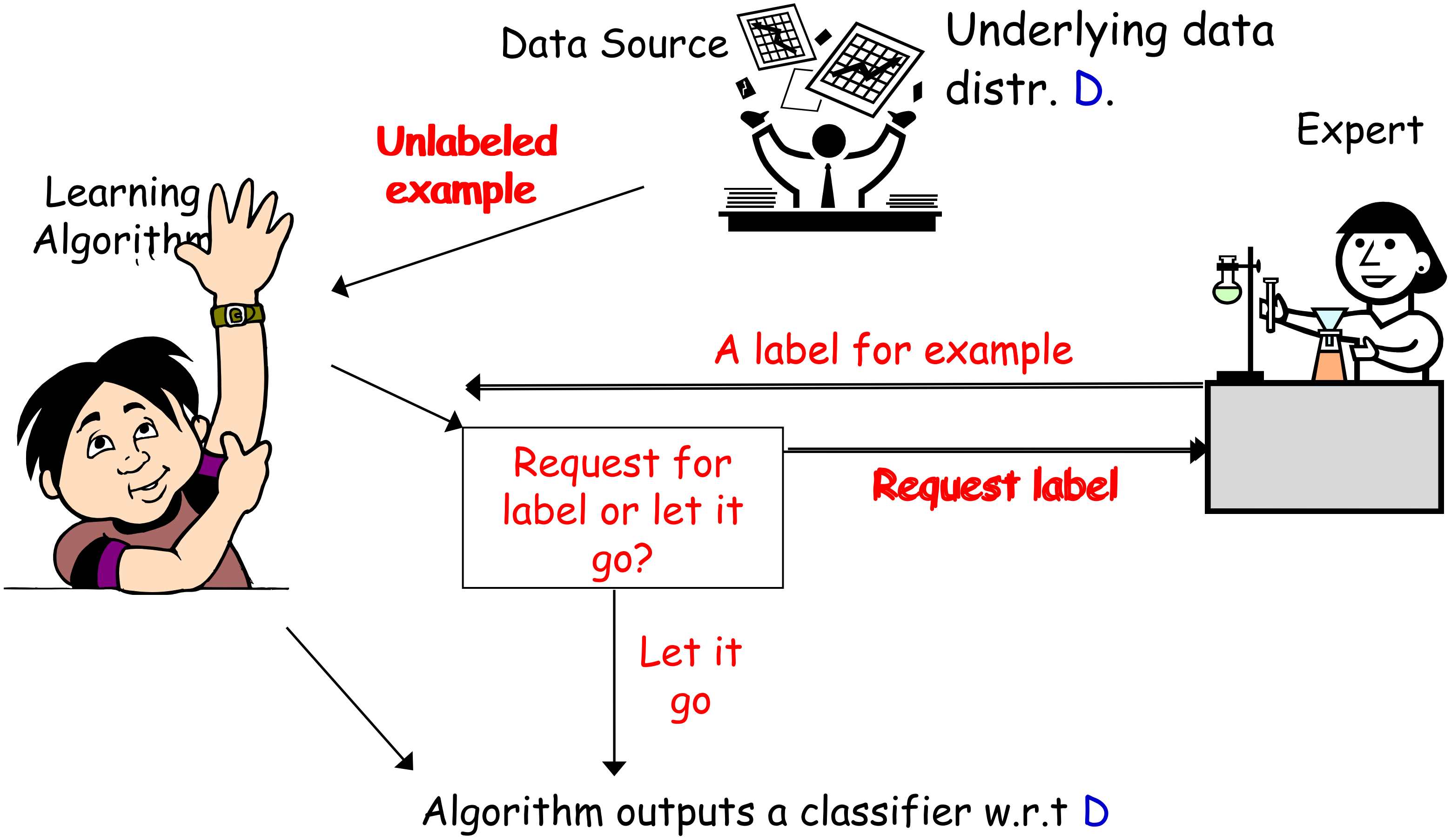
A Label for that Example

⋮

Algorithm outputs a classifier w.r.t D

- Learner can choose specific examples to be labeled.
- Goal: use fewer labeled examples [pick informative examples to be labeled].

Selective Sampling Active Learning


Data Source
Underlying data distr. D.
Expert
Unlabeled example
Learning Algorithm
A label for example
Request for label or let it go?
Request label
Let it go
Algorithm outputs a classifier w.r.t D

- Selective sampling AL (Online AL): stream of unlabeled examples, when each arrives make a decision to ask for label or not.
- Goal: use fewer labeled examples [pick informative examples to be labeled].

What Makes a Good Active Learning Algorithm?

- Guaranteed to output a relatively good classifier for most learning problems.

- Doesn't make too many label requests.

  Hopefully a lot less than passive(被动的，消极的) learning and SSL.

- Need to choose the label requests carefully, to get informative labels.

Can adaptive querying really do better than passive/random sampling?

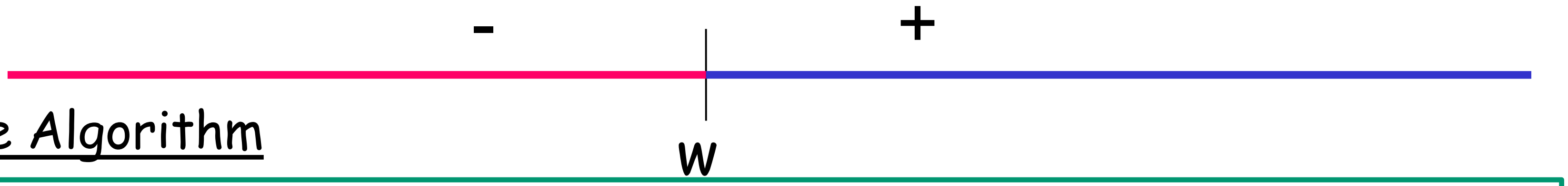
- YES! (sometimes)

- We often need far fewer labels for active learning than for passive.

- This is predicted by theory and has been observed in practice.

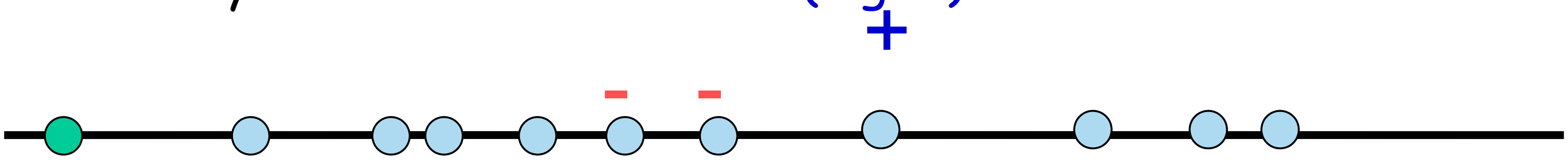
Can adaptive querying help? [CAL92, Dasgupta04]

- Threshold fns on the real line: $h_w(x) = 1(x$ ¸ $w)$, $C = \{h_w: w$ 2 $R\}$



Active Algorithm

- Get N unlabeled examples
- How can we recover the correct labels with queries?
- Do binary search! Just need O(log N) labels!



- Output a classifier consistent with the N inferred labels.

- we are guaranteed to get a classifier of error .

Passive supervised: labels to find an ε-accurate threshold.

Active: only labels. Exponential improvement.

Common Technique in Practice

Uncertainty sampling in SVMs common and quite useful in practice. E.g., [Tong & Koller, ICML 2000; Jain, Vijayanarasimhan & Grauman, NIPS 2010; Schohon Cohn, ICML 2000]

Active SVM Algorithm

- At any time during the alg., we have a "current guess" of the separator: the max-margin separator of all labeled points so far.
- Request the label of the example closest to the current separator.

Common Technique in Practice

Active SVM seems to be quite useful in practice.

[Tong & Koller, ICML 2000; Jain, Vijayanarasimhan & Grauman, NIPS 2010]

Algorithm (batch version)

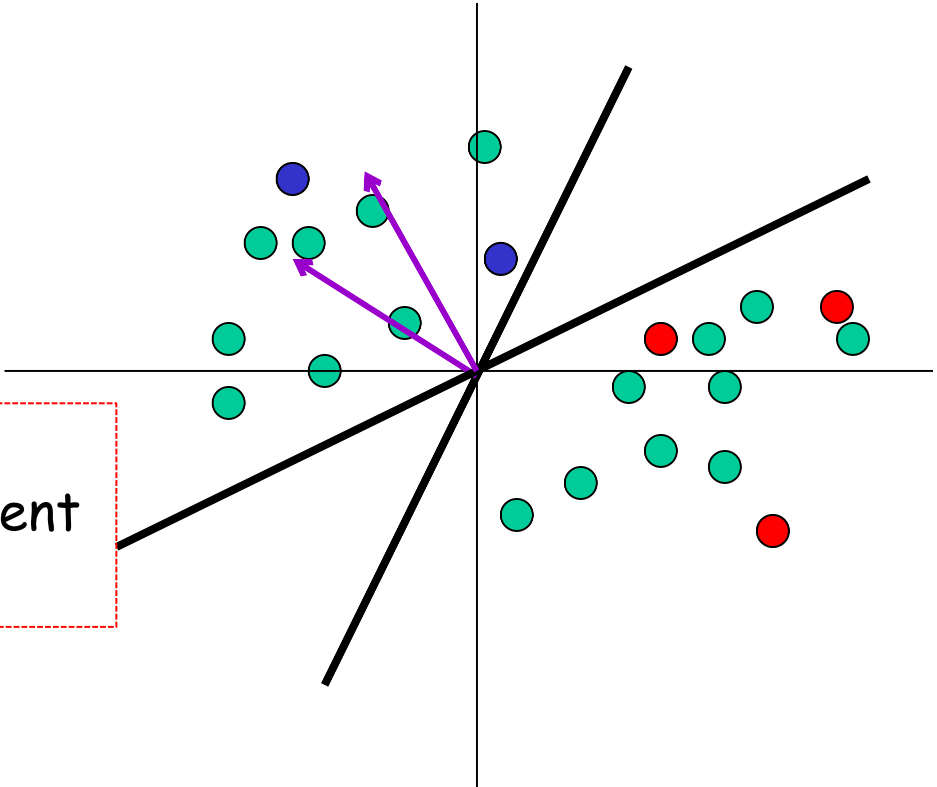
Input ={, …,} drawn i.i.d from the underlying source D

Start: query for the labels of a few random s.

For , …. ,

- Find the max-margin separator of all labeled points so far.
- Request the label of the example closest to the current separator: minimizing .

(highest uncertainty)

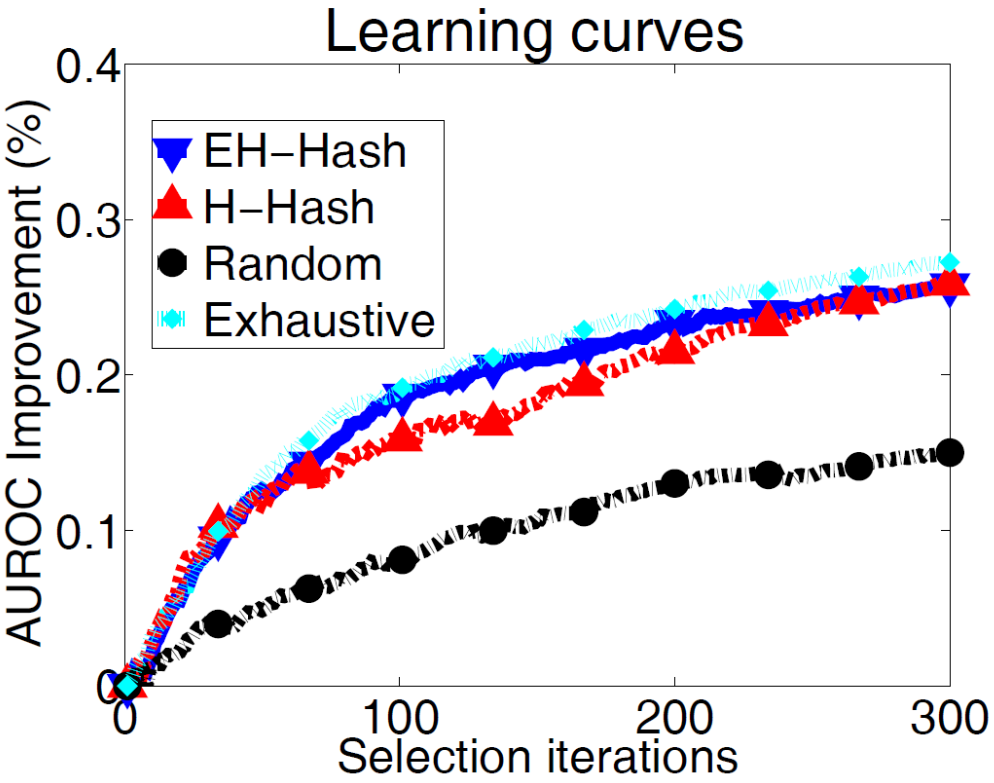


Common Technique in Practice

Active SVM seems to be quite useful in practice.

E.g., Jain, Vijayanarasimhan & Grauman, NIPS 2010

Newsgroups dataset (20.000 documents from 20 categories)

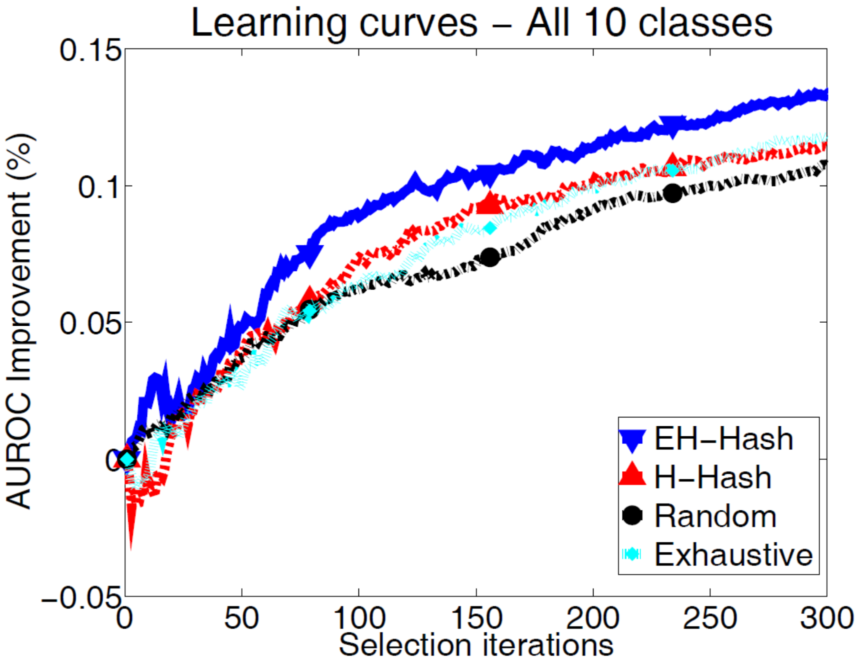


Common Technique in Practice

Active SVM seems to be quite useful in practice.

E.g., Jain, Vijayanarasimhan & Grauman, NIPS 2010

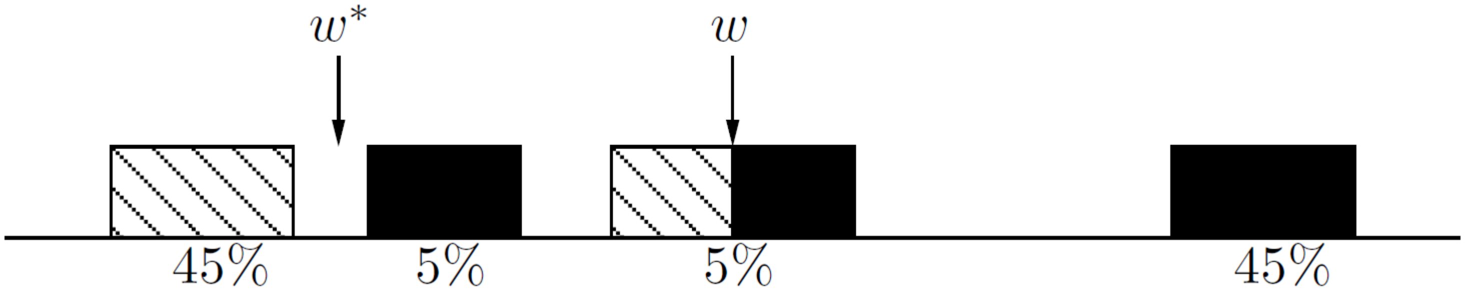
CIFAR-10 image dataset (60.000 images from 10 categories)



Active SVM/Uncertainty Sampling

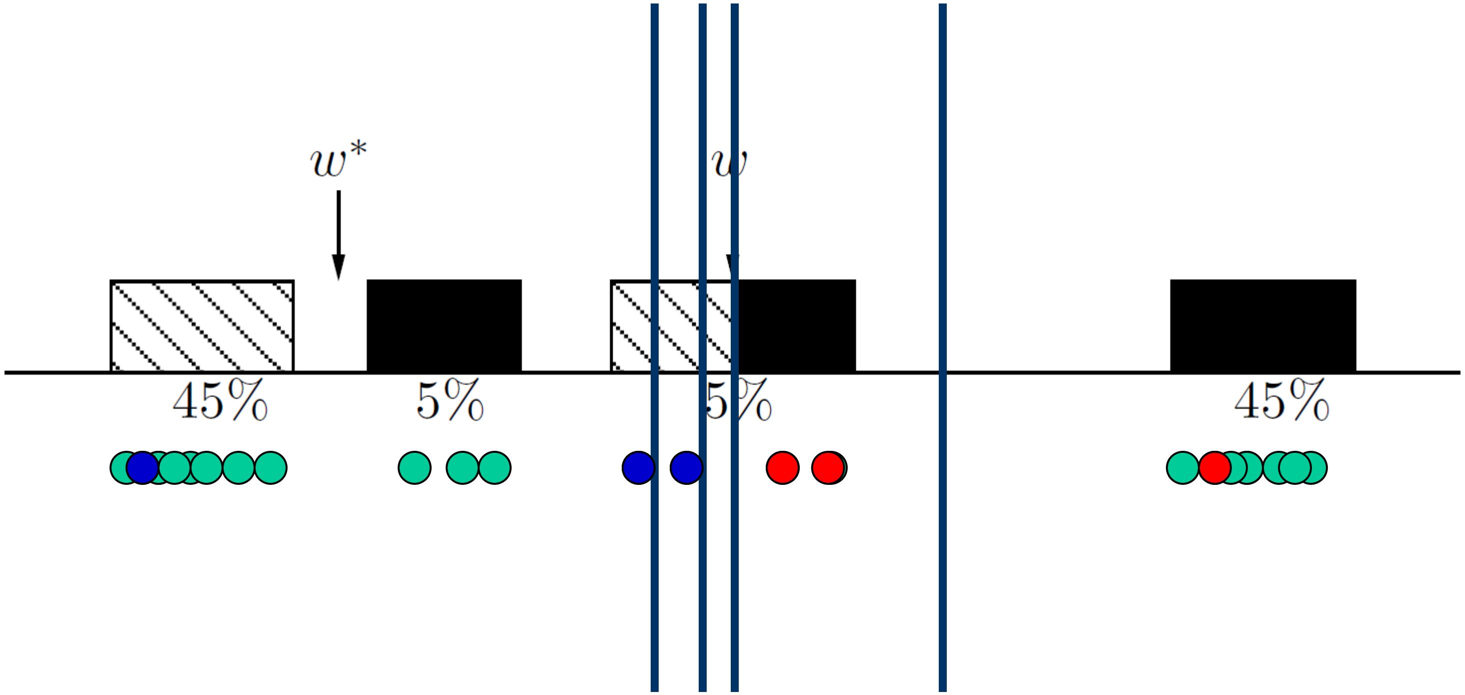
- Works sometimes….
- **However, we need to be very very very careful!!!**
  - Myopic(缺乏远见的), greedy technique can suffer from sampling bias.
  - A bias created because of the querying strategy; as time goes on the sample is less and less representative of the true data source.

[Dasgupta10]



Active SVM/Uncertainty Sampling

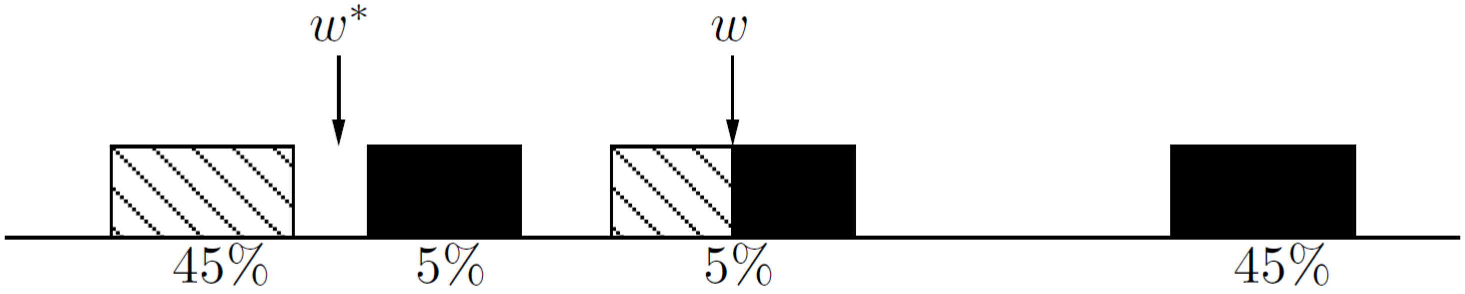
- Works sometimes….
- **However, we need to be very very careful!!!**



Active SVM/Uncertainty Sampling

- Works sometimes….
- **However, we need to be very very careful!!!**
  - Myopic, greedy technique can suffer from sampling bias.
  - Bias created because of the querying strategy; as time goes on the sample is less and less representative of the true source.
  - Observed in practice too!!!!

- **Main tension**: want to choose informative points, but also want to guarantee that the classifier we output does well on true random examples from the underlying distribution.



Safe Active Learning Schemes

Disagreement Based Active Learning

Hypothesis Space Search

[CAL92] [BBL06]

[Hanneke'07, DHM'07, Wang'09 , Fridman'09, Kolt10, BHW'08, BHLZ'10, H'10, Ailon'12, …]

Version Spaces

- X – feature/instance space; distr. D over X; target fnc
- Fix hypothesis space H.

Definition (Mitchell'82) Assume realizable case: .

Given a set of labeled examples (), …,(), $y_i = c^*(x_i)$

Version space of H: part of H consistent with labels so far.

I.e., iff .

Version Spaces

- X – feature/instance space; distr. D over X; target fnc
- Fix hypothesis space H.

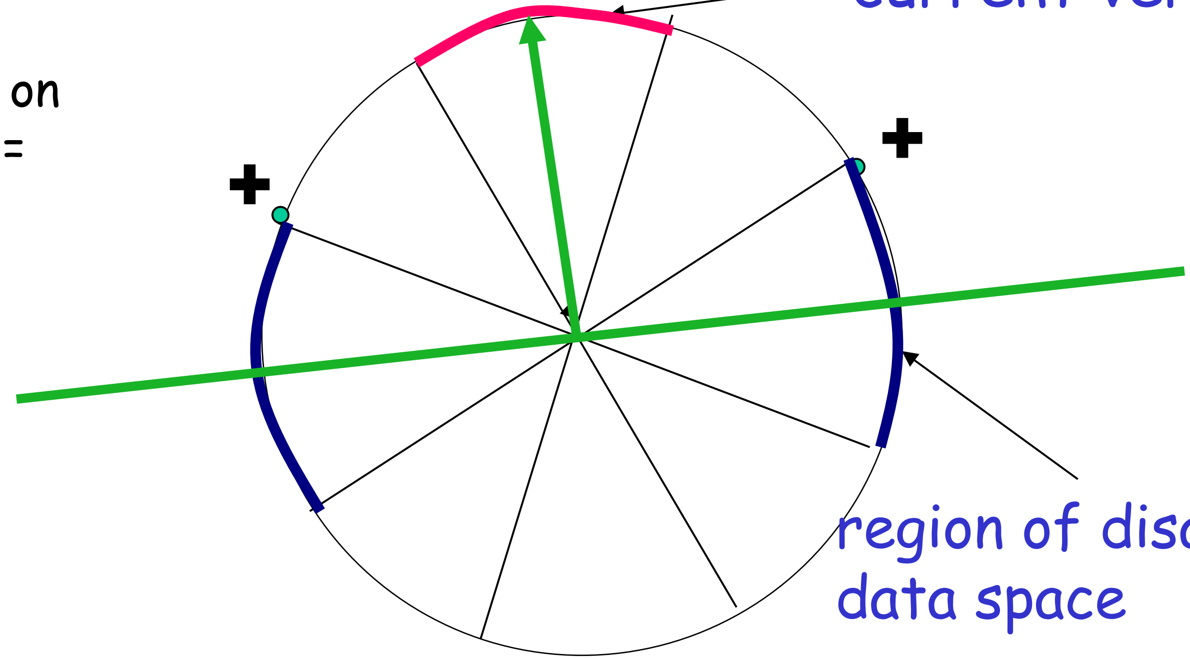
Definition (Mitchell'82) Assume realizable case: .

Given a set of labeled examples (), …,(),

$$y_i = c^*(x_i)$$

Version space of H: part of H consistent with labels so far.

E.g.,: data lies on circle in R^2, H = homogeneous linear seps.



current version space

region of disagreement in data space

Version Spaces. Region of Disagreement

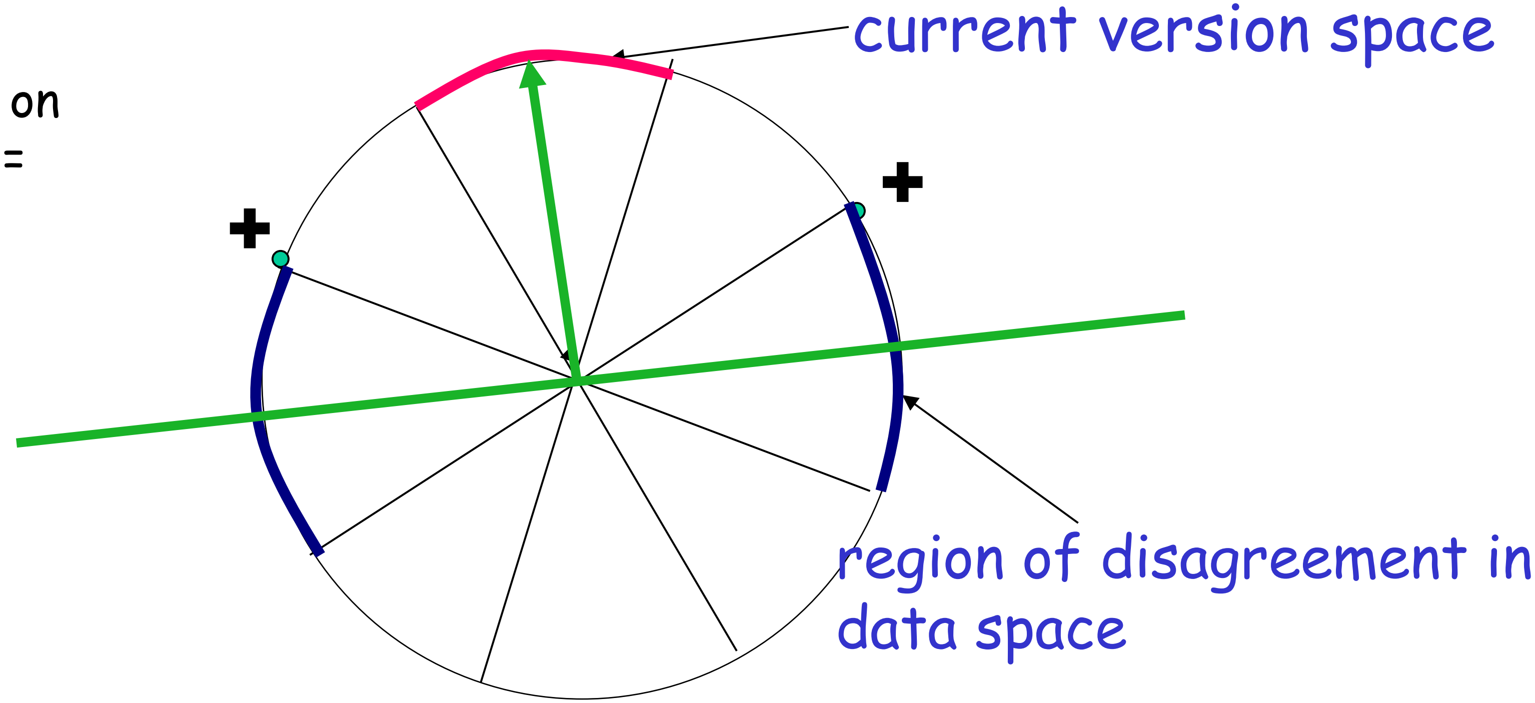
Definition (CAL'92)

Version space: part of H consistent with labels so far.

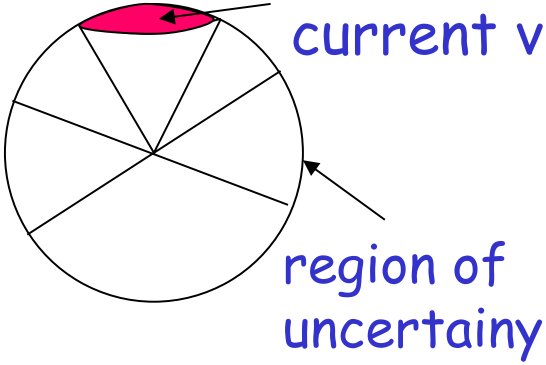
Region of disagreement = part of data space about which there is still some uncertainty (i.e. disagreement within version space)

iff

E.g.,: data lies on circle in R^2, H = homogeneous linear seps.



Disagreement Based Active Learning [CAL92]



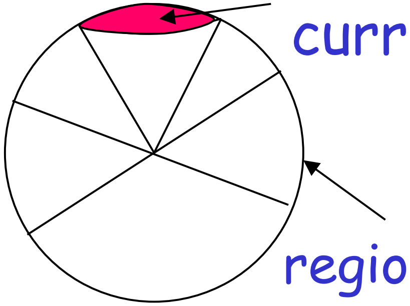
Algorithm:

Pick a few points at random from the current region of uncertainty and query their labels.

Stop when region of uncertainty is small.

Note: it is active since we do not waste labels by querying in regions of space we are certain about the labels.

Disagreement Based Active Learning [CAL92]



current version space

region of uncertainy

Algorithm:

Query for the labels of a few random s.

Let be the current version space.

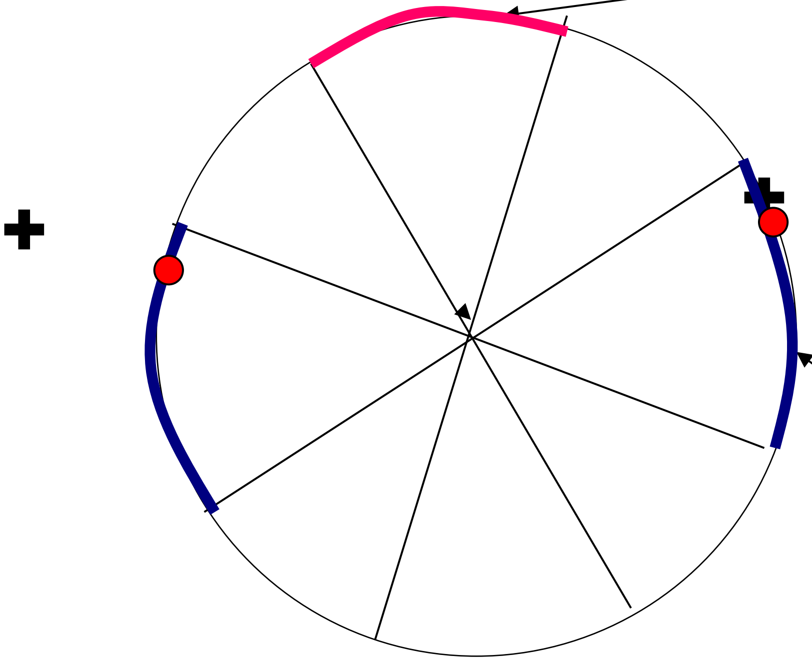
For , ….,

Pick a few points at random from the current region of disagreement and query their labels.

Let be the new version space.

Region of uncertainty [CAL92]

- Current version space: part of *C* consistent with labels so far.
- "Region of uncertainty" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)

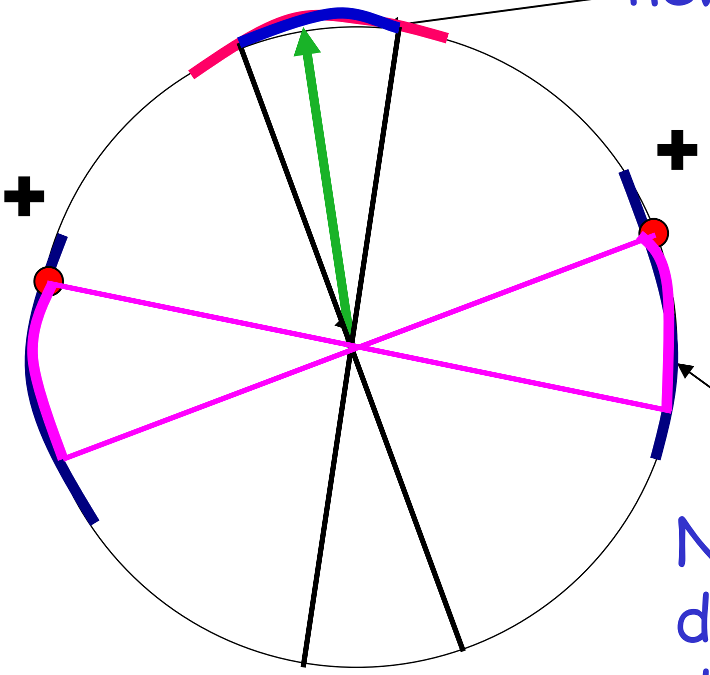


current version space

region of uncertainty in data space

Region of uncertainty [CAL92]

- Current version space: part of *C* consistent with labels so far.
- "Region of uncertainty" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)

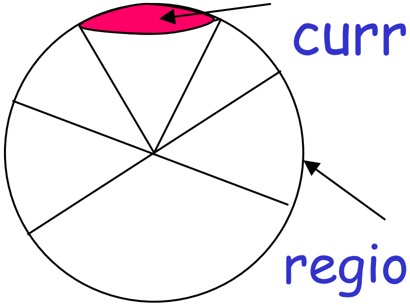


new version space

New region of disagreement in data space

How about the agnostic(不可知的) case where the target might not belong the H?

A^2 Agnostic Active Learner [BBL'06]



current version space

region of disagreement

Algorithm:

Let $H_1 = H$.

For , ….,

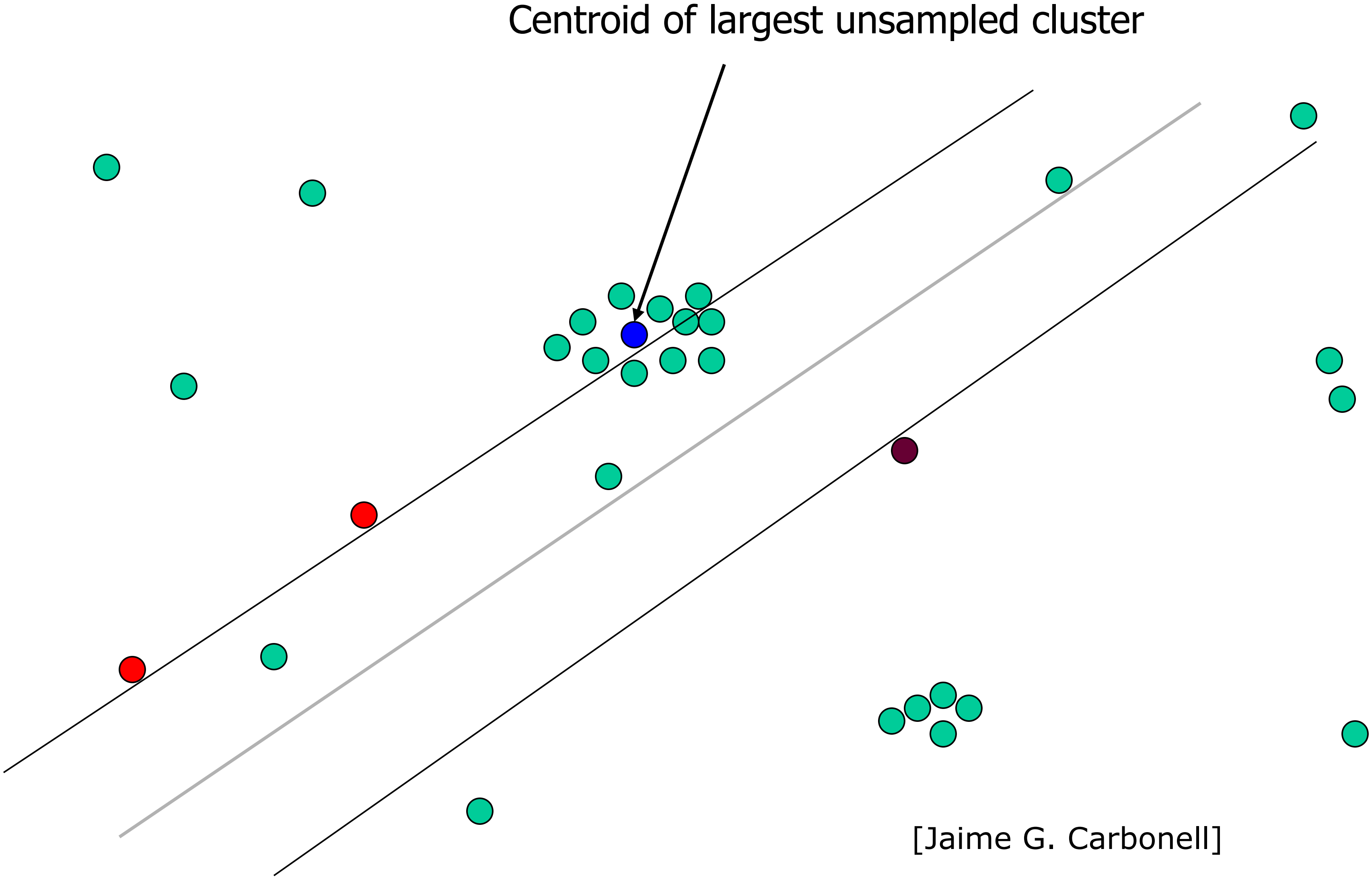
- Pick a few points at random from the current region of disagreement and query their labels.
- Throw out hypothesis if you are statistically confident they are suboptimal.

Careful use of generalization bounds; Avoid the sampling bias!!!!

Other Interesting ALTechniques used in Practice

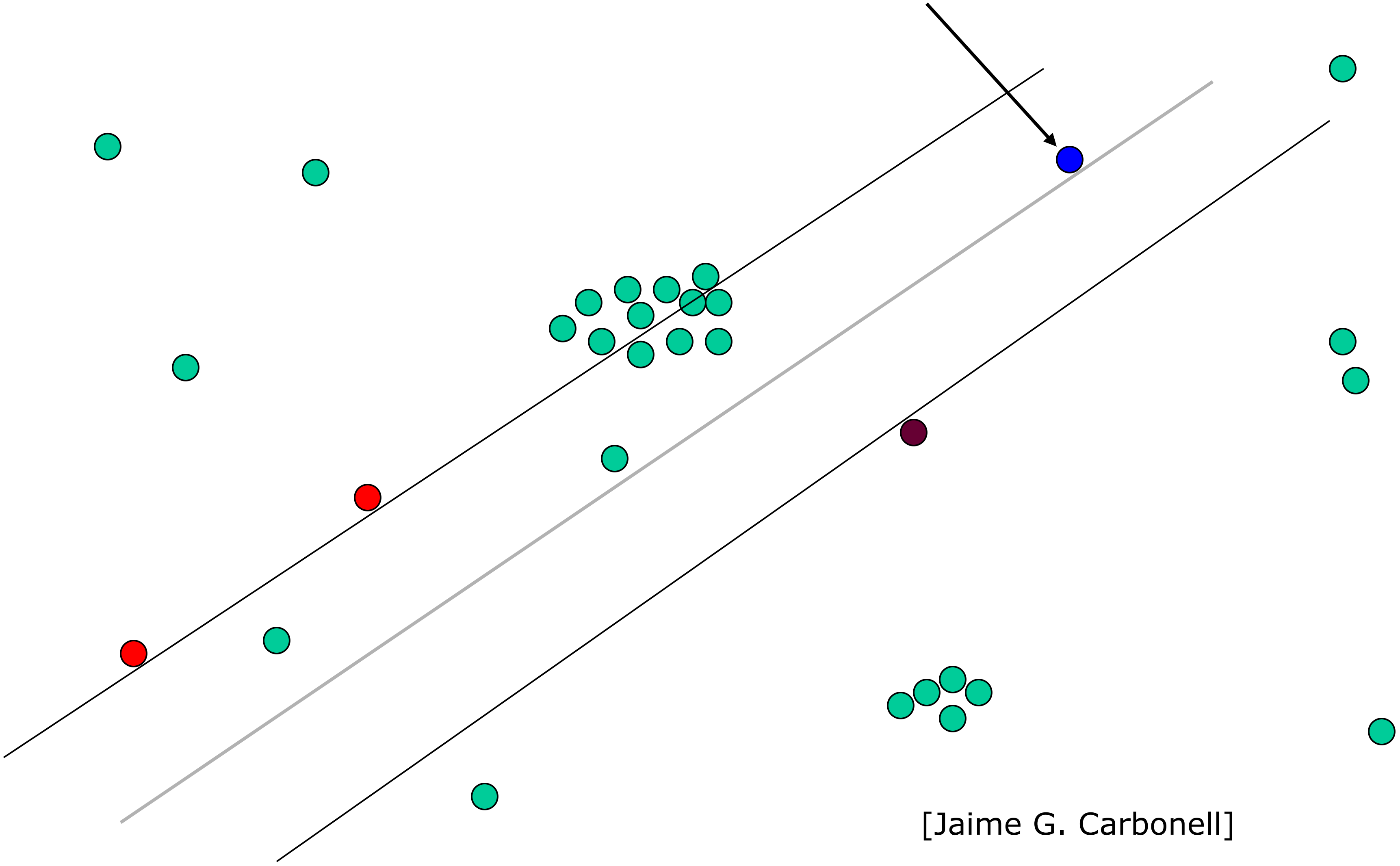
Interesting open question to analyze under what conditions they are successful.

Density-Based Sampling


Centroid of largest unsampled cluster

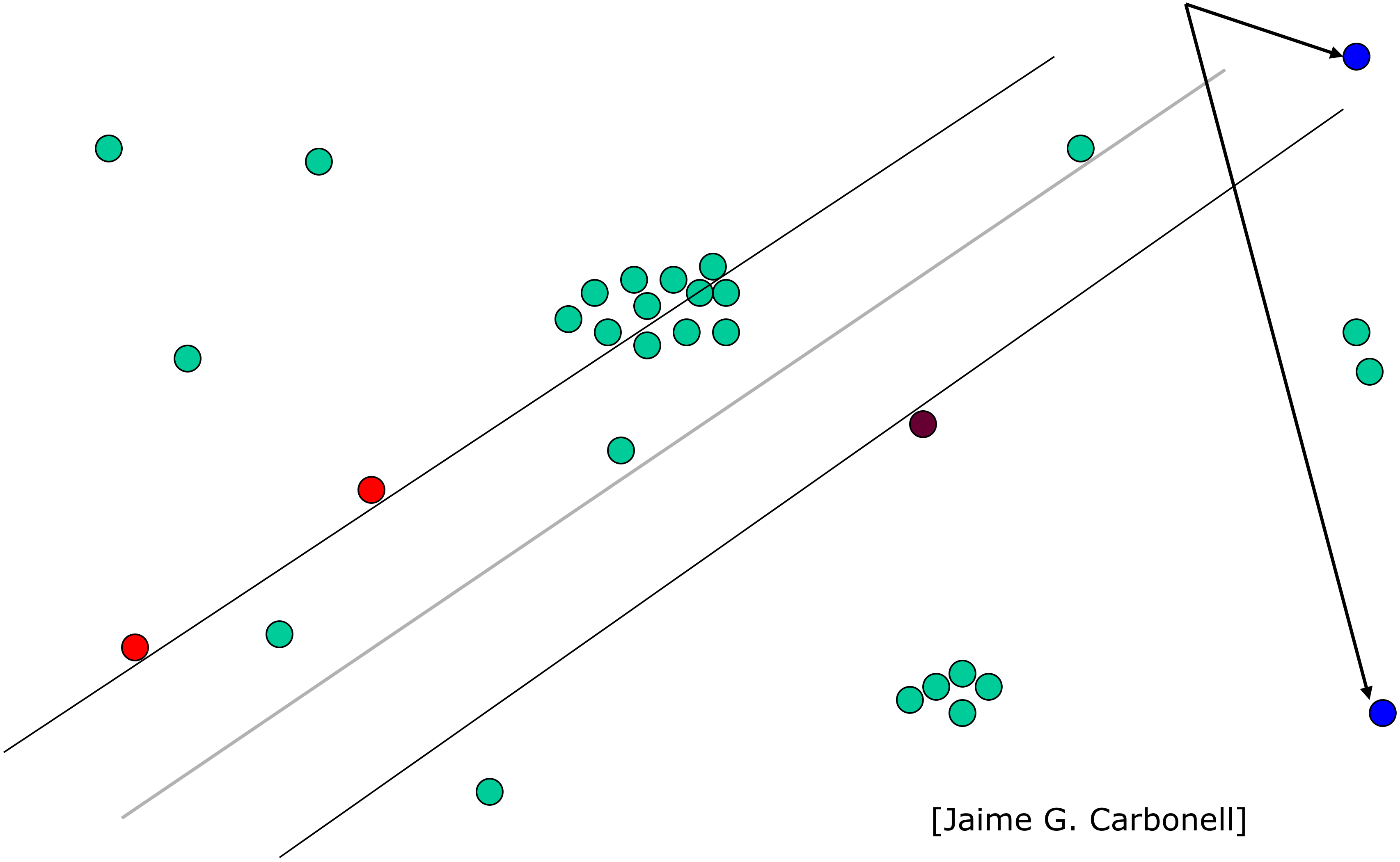
[Jaime G. Carbonell]

Uncertainty Sampling


Closest to decision boundary (Active SVM)

[Jaime G. Carbonell]

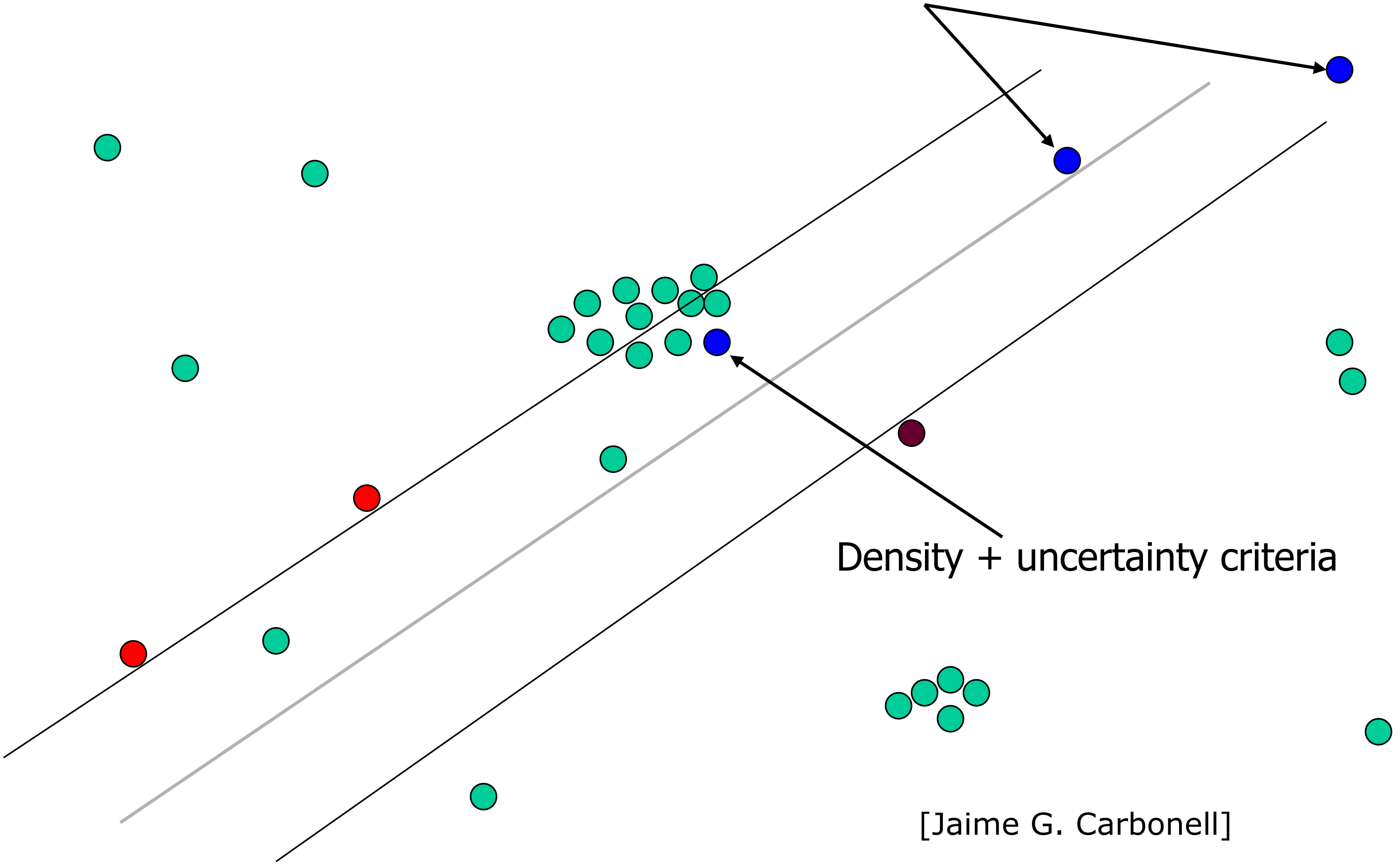
Maximal Diversity Sampling



Maximally distant from labeled x’s

[Jaime G. Carbonell]

Ensemble-Based Possibilities



Uncertainty + Diversity criteria

Density + uncertainty criteria

[Jaime G. Carbonell]

What You Should Know

- Active learning could be really helpful, could provide exponential improvements in label complexity (both theoretically and practically)!

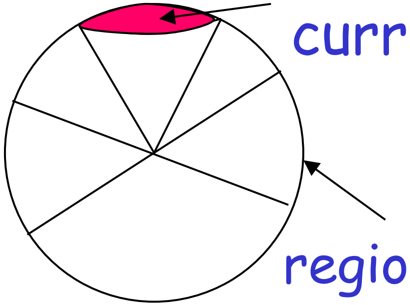
- Common heuristics (e.g., those based on uncertainty sampling). Need to be very careful due to sampling bias.

- Safe Disagreement Based Active Learning Schemes.
  - Understand how they operate precisely in the realizable case (noise free scenarios).

Advanced additional (not required material)

Disagreement based algorithms: How about the agnostic case where the target might not belong the H?

A^2 Agnostic Active Learner [BBL'06]



current version space

region of disagreement

Algorithm:

Careful use of generalization bounds; Avoid the sampling bias!!!!

Let $H_1 = H$.

For , ….,

- Pick a few points at random from the current region of disagreement and query their labels.
- Throw out hypothesis if you are statistically confident they are suboptimal.

Formal General Guarantees for Agnostic AL

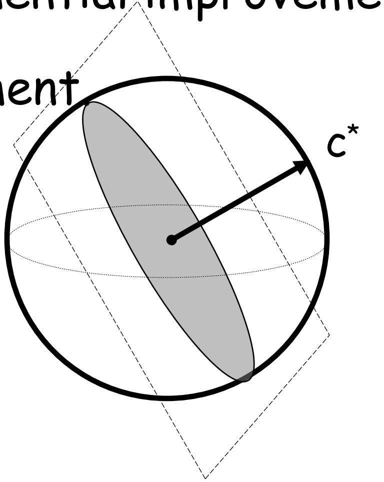
A^2 the first algorithm which is robust to noise.

[Balcan, Beygelzimer, Langford, ICML'06] [Balcan, Beygelzimer, Langford, JCSS'08]

"Region of disagreement" style: Pick a few points at random from the current region of disagreement, query their labels, throw out hypothesis if you are statistically confident they are suboptimal.

Guarantees for A^2 [BBL'06,'08]:

- It is safe (never worse than passive learning) & exponential improvements.
  - C – thresholds, low noise, exponential improvement.
  - C - homogeneous linear separators in R^d, D - uniform, low noise, only $d^2 \log(1/\varepsilon)$ labels.



A lot of subsequent work.

[Hanneke'07, DHM'07, Wang'09 , Fridman'09, Kolt10, BHW'08, BHLZ'10, H'10, Ailon'12, …]

General guarantees for A² Agnostic Active Learner

"Disagreement based": Pick a few points at random from the current region of uncertainty, query their labels, throw out hypothesis if you are statistically confident they are suboptimal. [BBL'06]

How quickly the region of disagreement collapses as we get closer and closer to optimal classifier

<u>Guarantees for A²</u> [Hanneke'07]:

Disagreement coefficient $\theta_{c^*} = \sup_{r \geq \eta+\epsilon} \frac{\Pr(DIS(B(c^*,r)))}{r}$

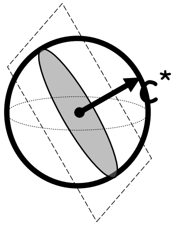
Theorem

$$m = \left(1 + \frac{\eta^2}{\epsilon^2}\right) VCdim(C)\theta_{c^*}^2 \log(\frac{1}{\varepsilon})$$

labels are sufficient s.t. with prob. $\geq 1-\delta$ output h with $err(h) \leq \eta+\epsilon$.

Realizable case: $m = VCdim(C)\theta_{c^*} \log(\frac{1}{\varepsilon})$

Linear Separators, uniform distr.: $\theta_{c^*} = \sqrt{d}$



Disagreement Based Active Learning

"Disagreement based " algos: query points from current region of disagreement, throw out hypotheses when statistically confident they are suboptimal.

- Generic (any class), adversarial label noise.
- Computationally efficient for classes of small VC-dimension

Still, could be suboptimal in label complex & computationally inefficient in general.

Lots of subsequent work trying to make is more efficient computationally and more aggressive too: [Hanneke07, DasguptaHsuMontleoni'07, Wang'09 , Fridman'09, Koltchinskii10, BHW'08, BeygelzimerHsuLangfordZhang'10, Hsu'10, Ailon'12, …]

applications

- Text classification
- Web page classification
- Junk mail recognition

active learning with different methods

- 1, Neural Networks
- 2, Bayesian rule
- 3, SVM
- No matter which method will be used, the core problem will be the same.

active learning with different methods

- The core problem is how to select training points actively?
- In other words, which training points will be informative to the model?

Apply active learning to Neural Networks

- Combined with query by committee
- Algorithm:

1, Samples two Neural Networks from distribution

2, when the unlabeled data arrives, use the committee to predict the label

3, if they disagree with each other, select it.

Apply active learning to Neural Networks

- Usually:
- Committee may contain more than two members.
- Classification problem will count #(+) and #(-) to see whether they are close.
- Regression problem use the variance of the outputs as the criteria of disagreement.
- Stop criteria is maximum model variance dropped below a set threshold.

Apply active learning to Baysian theory

- Characteristic:
- build a probabilistic classifier which not only make classification decisions, but estimate their uncertainty
- Try to estimate P(Ci | w), posterior probability that an example with pattern w belongs to class Ci.
- P(Ci | w) will directly guide to select training data.

Apply active learning to SVM

- Problem is also what is the criteria for uncertainty sampling?
- we can improve the model by attempting to maximally narrow the existing margin.
- If the points which lie on or close to the dividing hyperplane are added into training points, it will on average narrow the margin most.

Apply active learning to Baysian theory

- About the stopping criteria:
- All unlabeled data in the margin have been exhausted, we will stop.
- Why?
- Only unlabeled data within the margin will have great effect on our learner.
- Labeling an example in the margin may shift the margin such that examples that were previously outside are now inside.

Employing EM and Pool-based Active Learning for Text Classification

- Motivation:
- Obtaining labeled training examples for text classification is often expensive, while gathering large quantities of unlabeled examples is very cheap.
- Here, we will present techniques for using a large pool of unlabeled documents to improve text classification when labeled training data is sparse.

How data are produced

- We approach the task of text classification from a bayesian learning perspective, we assume that the documents are generated by a particular parametric model, mixture of naïve nayes, and one-to-one correspondence between class labels and the mixture components.

How data are produced

The likelihood of a document is a sum of total probability over all generative components

$$\mathrm{P}(d_i|\theta) = \sum_{j=1}^{|C|} \mathrm{P}(c_j|\theta)\mathrm{P}(d_i|c_j;\theta).$$

$c_j \in C = \{c_1, \ldots, c_{|C|}\}$,Indicate the jth component and jth class

Each component cj is parameterized by a disjoint subset of θ

How data are produced

- Document di is considered to be an ordered list of word events.
- Wdik represents the word in position k of document di. The subscript of w indicates an index into the vocabulary V=<w1,w2,…,w|v|>.
- Combined with standard naïve bayes assumption: words are independent from other words in the same document.

$$P(d_i|c_j;\theta) = \prod_{k=1}^{|d_i|} P(w_{d_{ik}}|c_j;\theta)$$

goal

- Given these underlying assumption of how data are produced, the task of learning a text $\hat{\theta}$ classifier consists of forming an estimate of θ, written as $\hat{\theta}$ based on a training set.

Formular

- If the task is to classify a test document di into a single class, simply select the class with the highest posterior probability: argmax$_j$ P(c$_j$|d$_j$; $\hat{\theta}$)

$$\mathrm{P}(c_j|d_i;\hat{\theta}) = \frac{\mathrm{P}(c_j|\hat{\theta})\prod_{k=1}^{|d_i|}\mathrm{P}(w_{d_{ik}}|c_j;\hat{\theta})}{\sum_{r=1}^{|C|}\mathrm{P}(c_r|\hat{\theta})\prod_{k=1}^{|d_i|}\mathrm{P}(w_{d_{ik}}|c_r;\hat{\theta})}.$$

EM and Unlabeled data

- problem:
- When naïve bayes is given just a small set of labeled training data, classifiction accuracy will suffer because variance in the parameter estimates of the generative model will be high.

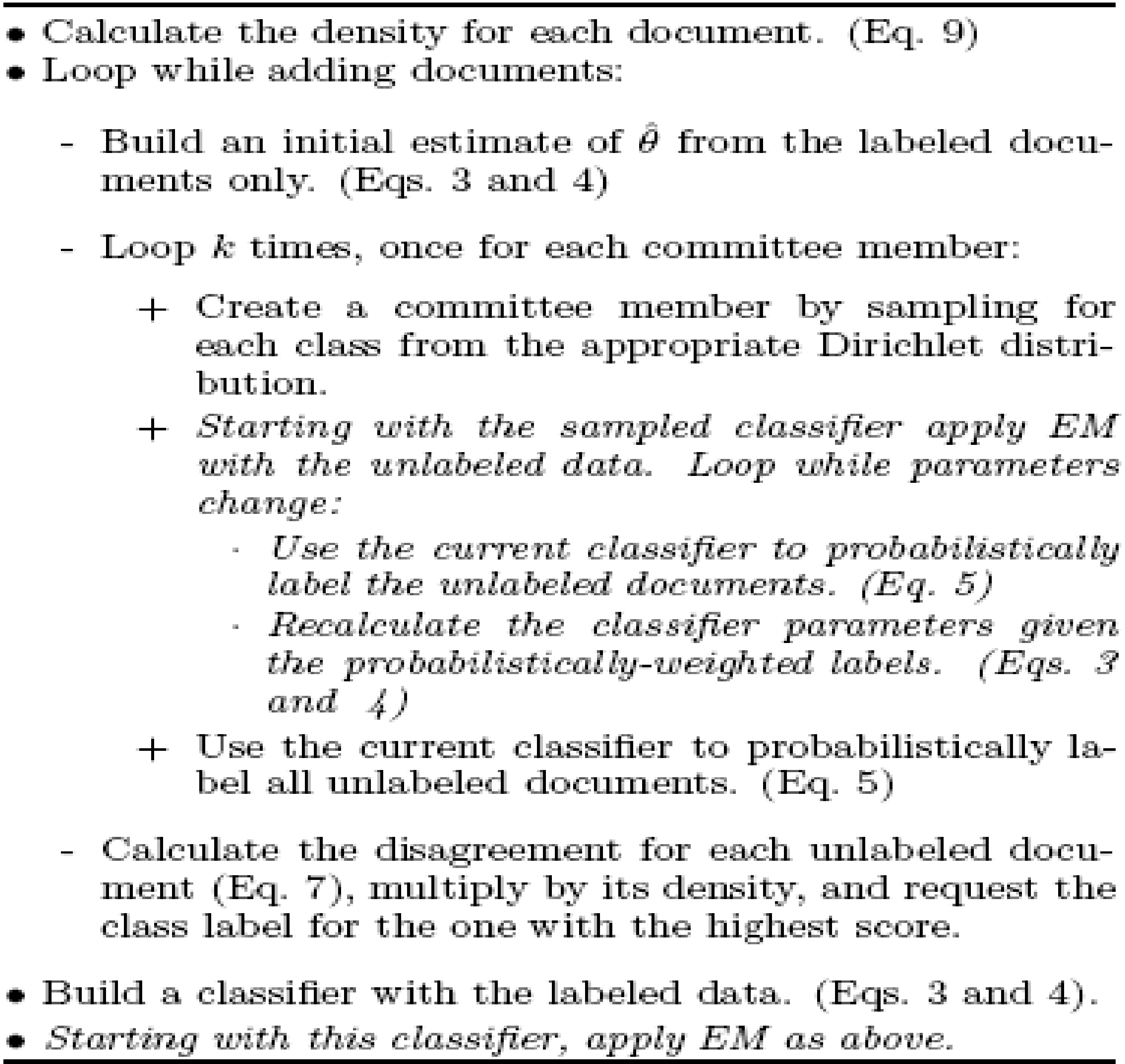
EM and Unlabeled data

- Motivation:
- By augmenting this small labeled set with a large set of unlabeled data and combining the two pools with EM, we can improve the parameter estimates.

implementation of EM

- Initialize just using labeled data.
- E-step:
- Calculate probabilistically-weighted class labels, P(cj | dj;$\hat{\theta}$), for every unlabeled document.
- M-step:
- Calculate a new maximum likelihood estimate for θ using all the labeled data.
- The process iterate until $\hat{\theta}$ reaches a fixed point

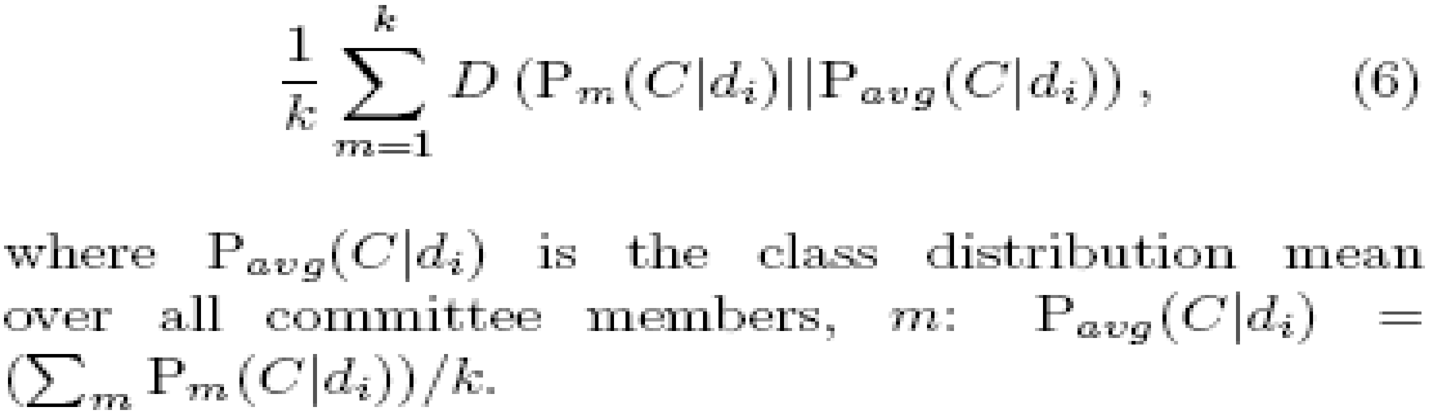
Active learning with EM



```
• Calculate the density for each document. (Eq. 9)
• Loop while adding documents:
    - Build an initial estimate of $\hat{\theta}$ from the labeled documents only. (Eqs. 3 and 4)
    - Loop $k$ times, once for each committee member:
        + Create a committee member by sampling for each class from the appropriate Dirichlet distribution.
        + Starting with the sampled classifier apply EM with the unlabeled data. Loop while parameters change:
            · Use the current classifier to probabilistically label the unlabeled documents. (Eq. 5)
            · Recalculate the classifier parameters given the probabilistically-weighted labels. (Eqs. 3 and 4)
        + Use the current classifier to probabilistically label all unlabeled documents. (Eq. 5)
    - Calculate the disagreement for each unlabeled document (Eq. 7), multiply by its density, and request the class label for the one with the highest score.
• Build a classifier with the labeled data. (Eqs. 3 and 4).
• Starting with this classifier, apply EM as above.
```

Disagreement creteria

- To measure committee disagreement for each document using Kullback-Leibler divergence to the mean.
- KL divergence to the mean is an average of the KL divergence between each distribution and the mean of all the distributions:



$$\frac{1}{k}\sum_{m=1}^{k} D\left(\mathrm{P}_m(C|d_i)||\mathrm{P}_{avg}(C|d_i)\right), \qquad (6)$$

where $\mathrm{P}_{avg}(C|d_i)$ is the class distribution mean over all committee members, m: $\mathrm{P}_{avg}(C|d_i) = (\sum_m \mathrm{P}_m(C|d_i))/k$.

END

Thank you