CS/CNS/EE 253: Advanced Topics in Machine Learning
Topic: Active Learning 1: Background and Pool-based learning **Lecturer:** Andreas Krause
Scribe: Nam Nguyen **Date:** Feb. 3, 2010

10.1 Background

10.1.1 Definitions

Hypotheses (ways to make decisions) $\mathbf{H} = h_1, ..., h_m, h_i : \mathbf{X} \mapsto -1, +1$
True concept $c \in \mathbf{H}$
Probability distribution $\mathbf{Pr}[x]$
Generalization error $error_{true}(h) = \mathbf{Pr}[h(x) \neq c(x)]$

10.1.2 Generalization Error Bounds

Want h such that $error_{true}(h) < \epsilon$
Standard learning problem: find best hypothesis by elimination

PAC-Bounds:
$\forall \epsilon, \delta > 0$ if we draw n instances from P *i.i.d.*,
$n \geq \frac{1}{\epsilon}(\log(|\mathbf{H}|) + \log(\frac{1}{\delta}))$ implies that $\mathbf{Pr}[\textit{all bad hypotheses eliminated}] \geq 1 - \delta$

In practice, however, there is typically an infinite number of hypotheses (linear separators, etc.)

Class suggestions for handling an infinite number of hypotheses:

1. Discretization: Enforce a discretization scheme onto the continuous set of hypotheses. The problem with this is that it is unclear how to proceed to get a bound on generalization error.
2. Shattering: finite $\mathbf{X}$, so there is a finite number of hypotheses from the point of view of possible classifications. However, this is $\log(2^n) = n\log(2)$, which is useless as a bound.

The answer lies with the concept of *VC dimension*, which takes into account that hypotheses are not independent.

10.1.3 VC Dimension

Definition 10.1.1 $\mathbf{S} \subset X$ *is called shattered by* $\mathbf{H}$ *if* $\forall \mathbf{S}_+ \subset \mathbf{S} \exists h \in \mathbf{H} : \forall x \in \mathbf{S} : x \in \mathbf{S}_+ \iff h(x) = +1$ $VC_x(\mathbf{H}) =$ *size of largest set* $\mathbf{S}$ *that is shattered by* $\mathbf{H}$

To prove that $VC(\mathbf{H}) \geq n$, one simply needs to find one configuration of n points that $\mathbf{H}$ shatters. However, to prove that $VS(\mathbf{H}) < n$, one needs to show that *all configurations* of n points cannot be shattered by $\mathbf{H}$.

Example: Linear threshold functions
All points on the real line less than a scalar parameter are classified as -1 and all others are +1.//
We can shatter one point, so $VC(\mathbf{H}) \geq 1$.

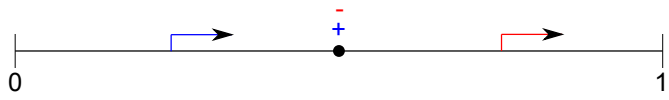

Figure 10.1.1: One spatial configuration, two possible classifications, both of which can be classified.

However, we can't shatter two points so $VC(\mathbf{H}) < 2$. $\therefore VC(\mathbf{H}) = 1$.

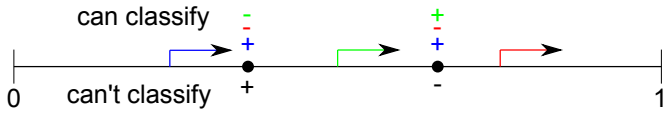

Figure 10.1.2: One spatial configuration, four possible classifications, one of which cannot be classified.

■

Example: Linear separators in 2D
We can find a configuration that can be shattered: three points that are noncollinear, so $VC(\mathbf{H}) \geq 3$.

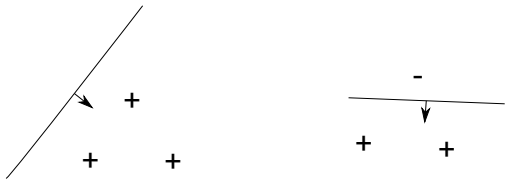

Figure 10.1.3: With three points, there are two types of configurations: collinear or noncollinear. Noncollinear configurations can be shattered.

However, we cannot shatter any configuration of four points, so $VC(\mathbf{H}) < 4$. $\therefore VC(\mathbf{H}) = 3$.

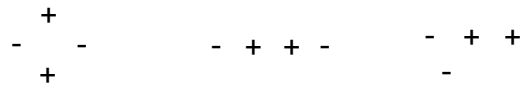

Figure 10.1.4: The three cases of spatial configurations of four points in 2D. All three cannot be shattered.

■

Generalization bounds based on VC dimension

$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\log |\mathbf{H}| + \log \frac{1}{\delta}}{2n}}$ (finite number of hypotheses)

$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{VC(|\mathbf{H}|)(1+\log n)+\log \frac{4}{\delta}}{n}}$

10.2 Active Learning

Labels Imply Other Labels
This means that there is underlying structure in the data X.

10.2.1 Uncertainty Sampling

Strategy consists of sampling points closest to current hypothesis. In the case of a linear separator, sample unlabeled points closest to the line.

D: set of unlabeled training examples
Assign each $x \in \mathbf{D}$ an uncertainty score $u(x) = /frac1|w^T x|$. Note that $|w^T x|$ is the distance from x to the hyperplane defined by w.

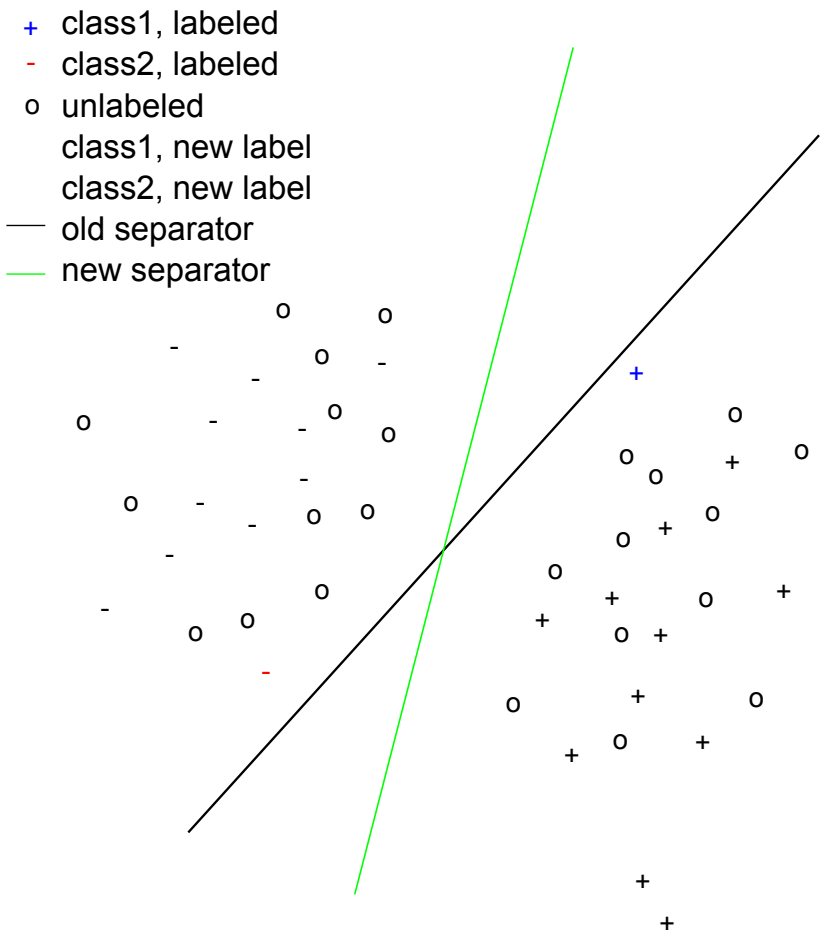

Figure 10.2.5: Intuition: a few extra labels can cause a big change.

However, uncertainty sampling can fail for cases in which $\mathbf{X}$ is very nonuniform:

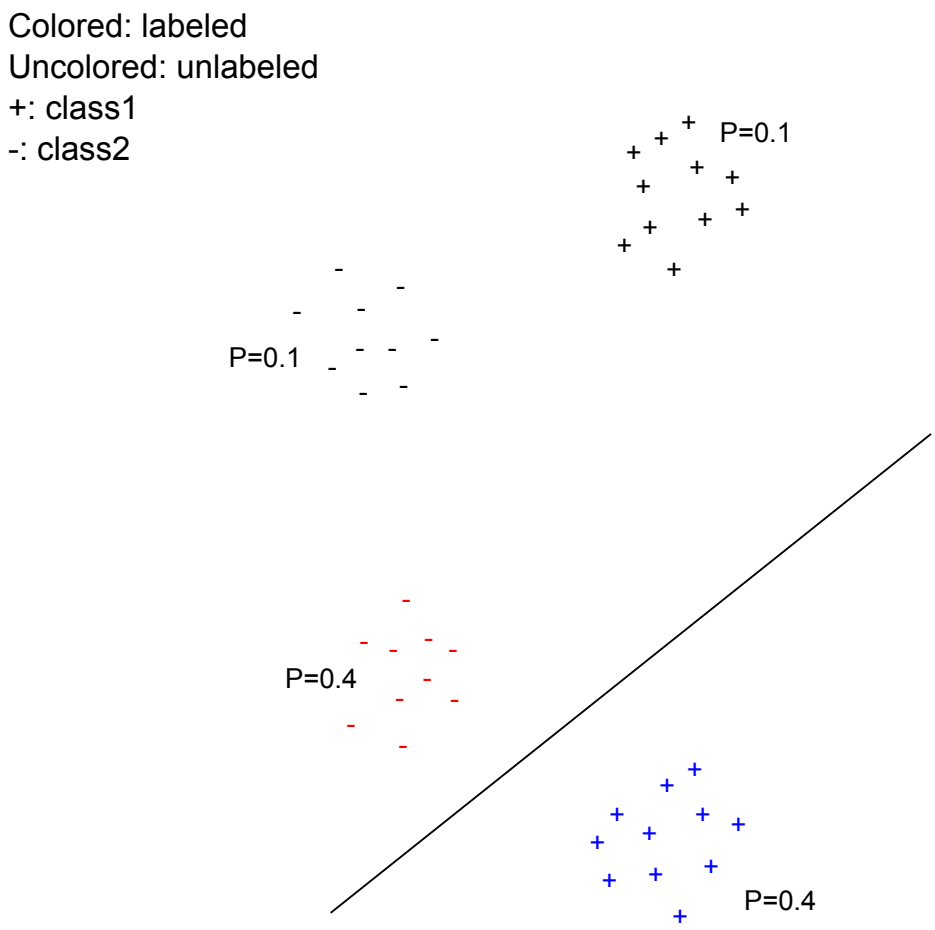

Figure 10.2.6: Large classification error because the learning is too local.

Uncertainty sampling produces an error of $error_{true}(h) \geq 0.1$. It doesn't target informative labels, so we need to develop a notion of informativeness for a given data point based on how much labeling it *would* change the hypothesis.

Active Learning Bias

Active learning bias refers to the degree to which the algorithm's selection of labels affects the distribution of labels you see. This may yield higher generalization error versus passive learning. It is reasonable to expect, however, that any decent active learning algorithm should perform at least as well as passive learning because the situation is the same except with more constraints. The idea of active learning bias leads us to pool-based learning.

10.2.2 Pool-based Learning

$\forall \epsilon, \delta$, classical learning theory gives a bound on n samples from $\mathbf{Pr}[n]$ such that if we had labels for all samples, for every h consistent with those labels,
$\mathbf{Pr}[error_{true}(h) \leq \epsilon] \geq 1 - \delta$

1. sample n unlabeled points
2. selectively query labels until all remaining labels are *implied*

The idea is to use the structure of the hypotheses together with the currently labeled points to infer as many labels of unlabeled points as possible. This only works if the true concept is at least similar to some $h \in \mathbf{H}$. Pool-based learning results in no active learning bias.
(In other words, this is in the noiseless separable case. This algorithm is sensitive to noise; one way to reduce noise is to sample more in the vicinity of a noisy location in $\mathbf{X}$.)

Example: Linear threshold functions
Here, pool-based learning is a binary search for the transition point between - and +. ■

Example: Homogeneous linear separators
We generally need few random samples to find a representative + and - pair. Once this is found, the algorithm reduces to binary search. This yields an exponential improvement in the number of labels needed for a given error. ■

However, if one class is very rare, active learning approaches passive learning because examples from each class are needed.