8803 Machine Learning Theory

Maria-Florina Balcan Lecture 26: April 15th, 2010

```
* Active Learning Overview (see also slides)
* Margin Based Learning of Linear Separators
```

There has recently been substantial interest in using unlabeled data together with labeled data for machine learning. The motivation is that unlabeled data can often be much cheaper and more plentiful than labeled data, and so if useful information can be extracted from it that reduces dependence on labeled examples, this can be a significant benefit.

There are currently two settings that have been considered to incorporate unlabeled data in the learning process. The first one is the so-called *Semi-supervised Learning* [2, 4], where, in addition to a set of labeled examples drawn at random from the underlying data distribution, the learning algorithm can also use a (usually larger) set of unlabeled examples from the same distribution. In this setting, unlabeled data becomes informative under *additional* assumptions and beliefs about the learning problem. Examples of such assumptions are the one used by Transductive SVM (namely, that the target function should cut through low density regions of the space), or by Co-training (namely, that the target should be self-consistent in some way). Unlabeled data is then potentially useful in this setting because it allows one to reduce search space from the whole set of hypotheses, down to the set of *a-priori* reasonable with respect to the underlying distribution.

The second setting, an increasingly popular one for the past few years, is *Active Learning* [1, 5, 7]. Here, the learning algorithm has both the capability of drawing random unlabeled examples from the underlying distribution and that of asking for the labels of *any* of these examples, and the hope is that a good classifier can be learned with significantly fewer labels by *actively* directing the queries to *informative* examples. As opposed to the Semi-supervised learning setting, and similarly to the classical supervised learning settings (PAC and Statistical Learning Theory settings) the only prior belief about the learning problem in the Active Learning setting is that the target function (or a good approximation of it) belongs to a given concept class. Luckily, it turns out that for simple concept classes such as linear separators on the line one can achieve an *exponential* improvement (over the usual supervised learning setting) in the labeled data sample complexity, under no additional assumptions about the learning problem [1, 5].[1] In general, however, for more complicated concept classes, the speed-ups achievable in the active learning setting depend on the match between the distribution over example-label pairs and the hypothesis class, and therefore on the target hypothesis in the class. Furthermore, there are simple examples where active learning does not help at all, even if there in the realizable case (see, for example, [7]). Recent work of Dasgupta [7] gives a generic characterization of the sample complexity aspect of active learning in the realizable case.

[1] For this simple concept class one can achieve a pure exponential improvement [5] in the realizable case, while in the agnostic case the improvement depends upon the noise rate [1].

1 Margin Based Learning of Linear Separators

We analyze here the Margin-based algorithm in Figure 1. A general version of the type of algorithm we analyze here (with extensions to certain types of noise) appears in [3].

We denote by $\mathrm{d}(f,g)$ the probability that the two classifiers f and g predict differently on an example coming at random from P. Furthermore, for $\alpha \in [0,1]$ we denote by $\mathrm{B}(f,\alpha)$ the set $\{g \mid \mathrm{d}(f,g) \leq \alpha\}$.

1.1 The Realizable Case and the Uniform Distributions

We consider here a commonly studied setting in the active learning literature [6, 7, 8]. Specifically, we assume that the data instances are drawn uniformly from the the unit ball in R^d, and that the labels are consistent with a linear separator w^* going through the origin (that is $P(w^* \cdot xy \leq 0) = 0$). We assume that $\|w^*\|_2 = 1$. It is worth noting that even in this seemingly simple looking scenario, there exists an $\Omega\left(\frac{1}{\epsilon}\left(d + \log\frac{1}{\delta}\right)\right)$ lower bound on the PAC learning sample complexity [9].

Note that given our assumption about the data distribution the error rate of any given separator w is $\mathrm{err}(w) = \frac{\theta(w,w^*)}{\pi}$, where $\theta(w,w^*) = \arccos(w \cdot w^*)$.

Input: allowed error rate ϵ, sampling oracle for D, labeling oracle O
a sequence of sample sizes $m_k > 0$, $k \in Z^+$;
a sequence of cut-off values $b_k > 0$, $k \in Z^+$
a sequence of hypothesis space radii $r_k > 0$, $k \in Z^+$;
Output:classifier w_s of error at most ϵ

First use $O(d)$ examples to find an hypothesis w_1 of error at most $\frac{1}{8}$.
iterate $k = 1, \ldots, s$
Rejection sample m_k samples x from D satisfying $|w_k^T \cdot x| \leq b_k$
Ask for labels and find a separator $w_{k+1} \in \mathrm{B}(w_k, r_k)$ consistent with all these examples.
end iterate

Note In at each iteration k, we can apply our favorite algorithm for finding a consistent linear separator (SVM for the realizable case, linear programming, etc).

Figure 1: Margin-based Active Learning

Theorem 1.1 *For any $\epsilon, \delta > 0$, using Procedure 1 with $b_k = O\left(\frac{\log\left(2^k\right)}{2^k\sqrt{d}}\right)$, $r_k = \frac{1}{2^k}$ and $m_k = O\left(\ln\frac{1}{\epsilon}\left(d\ln\ln\frac{1}{\epsilon} + d\ln\frac{\ln\frac{1}{\epsilon}}{\delta}\right)\right)$, after $s = \log\frac{1}{\epsilon}$ iterations, we find a separator of error at most ϵ with probability $1-\delta$.*

Proof: We prove by induction on k that at the k'th iteration we have $\mathrm{err}(w_k) \leq 2^{-k}$ with probability $1 - \frac{\delta}{k}$. For $k = 1$, by standard VC-bounds, we only need $m_1 = O(d + \ln(1/\delta))$ examples to obtain the desired result.

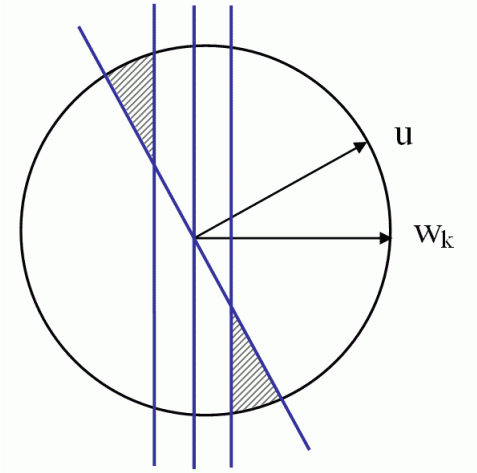

Figure 2: If the angle between u and w_k is at most $\pi\beta$, then at most a $\frac{\beta}{8}$ fraction of the region of disagreement between u and w_k is outside a band of size $\gamma = \frac{C\sin(\beta)\log\left(\frac{1}{\beta}\right)}{\sqrt{d}}$ along the hyperplane specified by w_k.

Assume that w_k has error at most α. We are done if we can show that for our choice of b_k the separator w_{k+1} we find in round $k+1$ has error at most $\frac{\alpha}{2}$ with probability $1-\frac{\delta}{k}$, and we only need $O\left(\ln\frac{1}{\epsilon}\left(d\ln\ln\frac{1}{\epsilon}+d\ln\frac{\ln\frac{1}{\epsilon}}{\delta}\right)\right)$ examples to transition between w_k and w_{k+1}.

Let us denote by $\mathrm{err}(b_k, w_k, \alpha)$ the quantity $\sup_{w''\in B(w_k,\alpha)}\Pr(w'' \text{ errs on x } \mid |w_k\cdot x|\geq\gamma_k)$.

The key point is that for the uniform distribution for b_k as small as $O\left(\frac{\alpha\log\frac{1}{\alpha}}{\sqrt{d}}\right)$ we have $\mathrm{err}(b_k, w_k, \alpha)\leq\frac{\alpha}{4}$ – for a proof see Lemma 1.2. This then implies that in order to find w_{k+1} that has error $\frac{\alpha}{2}$ we only need to sample and label about $O\left(\ln\frac{1}{\alpha}\left(d\ln\ln\frac{1}{\alpha}+d\ln\frac{\ln\frac{1}{\epsilon}}{\delta}\right)\right)$ examples from D satisfying $|w_{k-1}^T\cdot x|\leq b_k$. This follows from the fact that we can decompose

$$\begin{aligned}\mathrm{err}(w_{k+1}) = &\Pr(w_{k+1}\text{ errs on }x\mid |w_k\cdot x|\geq b_k)\Pr(|w_k\cdot x|\geq b_k)\\ &+\Pr(w_{k+1}\text{ errs on }x\mid |w_k\cdot x|\leq b_k)\Pr(|w_k\cdot x|\leq b_k)\end{aligned}\tag{1}$$

By our choice of b_k we made sure that the first term Equation (1) is at most $\frac{\alpha}{4}$, and we also know (from Lemma A.1) that $\Pr(|w_k\cdot x|\leq b_k)\leq\alpha C\log\frac{1}{\alpha}$. So, it's enough to find w_{k+1} with the property that $\Pr(w_{k+1}\text{ errs on x }\mid |w_k\cdot x|\leq\gamma_k)\leq\frac{C_2}{\log\frac{1}{\alpha}}$. By standard VC-bounds, we can do so by using only $O\left(\ln\frac{1}{\epsilon}\left(d\ln\ln\frac{1}{\epsilon}+d\ln\frac{\ln\frac{1}{\epsilon}}{\delta}\right)\right)$ examples. ■

Note: This algorithm is more *aggressive* than the version space based algorithms of [5] and [1]. Indeed, we do not necessarily sample from the *entire* region of uncertainty – but we sample just from a subregion *carefully* chosen.

Lemma 1.1 *There exists C s.t. $\forall \beta \leq C_1$, if $\gamma = \frac{C \sin(\beta) \log\left(\frac{1}{\beta}\right)}{\sqrt{d}}$ we have $\forall w_k$, $\forall u \in B(w_k, \beta)$*

$$\Pr_x \left[(u \cdot x)(w_k \cdot x) < 0, |w_k \cdot x| \geq \gamma\right] \leq \frac{\beta}{8}.$$

Proof: For simplicity, we present here a somewhat informal argument. For a formal proof see [3]. Let C be such that $\forall \beta \leq C_1$ we have

$$\Pr\left[X \geq \frac{C \log\left(\frac{1}{\beta}\right)}{\pi}\right] \leq \frac{\beta}{8},$$

where X is a standard normal random variable. Let us fix $\beta \leq C_1$, and let $\gamma = \frac{C \sin(\beta) \log\left(\frac{1}{\beta}\right)}{\sqrt{d}}$. Let's also fix w_k, and let $u \in \mathrm{B}(w_k, \beta)$. So, $\mathrm{d}(u, w_k) \leq \beta$ and therefore $\theta(u, w_k) = \arccos(u \cdot w_k) = \tilde{\beta} \leq \pi\beta$. Assume now without loosing generality that $u = (1, 0, 0, ..., 0)$ and $w_k = \Big(\cos(\tilde{\beta}), \sin(\tilde{\beta}), 0, 0, ..., 0\Big)$. So, $u \cdot w_k = \cos(\tilde{\beta})$ and for $x = (x_1, x_2, ..., x_d)$ we have $u \cdot x = x_1$ and $w_k \cdot x = \cos(\tilde{\beta})x_1 + \sin(\tilde{\beta})x_2$. It is enough to upper bound $\Pr_{x \sim S^d}\left[w_k \cdot x \geq \gamma \mid (u \cdot x) < 0\right]$ since the desired probability is

$$\begin{aligned}\Pr_{x \sim S^d}\left[(u \cdot x)(w_k \cdot x) < 0, |w_k \cdot x| \geq \gamma\right] =& 2 \Pr_{x \sim S^d}\left[(u \cdot x) < 0, w_k \cdot x \geq \gamma\right] \\ =& \Pr_{x \sim S^d}\left[w_k \cdot x \geq \gamma | (u \cdot x) < 0\right]\end{aligned}$$

But

$$\begin{aligned}&\Pr_{x \sim S^d}\left[w_k \cdot x \geq \gamma | (u \cdot x) < 0\right] = \Pr_{x \sim S^d}\left[\cos(\tilde{\beta})x_1 + \sin(\tilde{\beta})x_2 \geq \gamma | x_1 < 0\right] \\ \leq &\Pr_{x \sim S^d}\left[\sin(\tilde{\beta})x_2 \geq \gamma | x_1 < 0\right] = \Pr_{x \sim S^d}\left[\sin(\tilde{\beta})x_2 \geq \gamma\right] \leq \Pr_{x \sim S^d}\left[\sin(\pi\beta)x_2 \geq \gamma\right].\end{aligned}$$

This implies

$$\Pr_{x \sim S^d}\left[w_k \cdot x \geq \gamma | (u \cdot x) < 0\right] \leq \Pr_{x \sim S^d}\left[\sin(\beta)x_2 \geq \frac{C \sin(\beta) \log\left(\frac{1}{\beta}\right)}{\pi\sqrt{d}}\right],$$

which by our choice of C is at most $\frac{\beta}{8}$, as desired.[2] For an illustration of this lemma see also Figure 1.1. ■

Lemma 1.2 *There exists $\tilde{C}$ s.t. $\forall \beta \leq C_1$, if $\gamma = \frac{\tilde{C} \sin(\beta) \log\left(\frac{1}{\beta}\right)}{\sqrt{d}}$ we have $\forall w_k$, $\forall u \in B(w_k, \beta)$ and $\forall w_{k+1} \in B(w_k, \beta)$*

$$\Pr_x \left[(u \cdot x)(w_{k+1} \cdot x) < 0, |w_k \cdot x| \geq \gamma\right] \leq \frac{\beta}{4}.$$

[2] We have used the fact that for large d, x_2 looks *almost* like a Gaussian with mean 0 and variance $\frac{1}{\sqrt{d}}$.

Proof: Let us fix $\beta \leq C_1$ and let $\tilde{\gamma} = \frac{C \sin(2\beta) \log\left(\frac{1}{2\beta}\right)}{\sqrt{d}}$, where C is the constant specified in Lemma 1.2. Let's also fix w_k, and let $u \in \mathrm{B}(w_k, \beta)$ and $w_{k+1} \in \mathrm{B}(w_k, \beta)$. We have (from Lemma 1.2) both

$$\Pr_x \left[(u \cdot x)(w_k \cdot x) < 0, |w_k \cdot x| \geq \tilde{\gamma}\right] \leq \frac{\beta}{8} \quad \text{and}$$

$$\Pr_x \left[(u \cdot x)(w_{k+1} \cdot x) < 0, |w_k \cdot x| \geq \tilde{\gamma}\right] \leq \frac{\beta}{8},$$

and a simple union bound implies:

$$\Pr_x \left[(u \cdot x)(w_{k+1} \cdot x) < 0, |w_k \cdot x| \geq \tilde{\gamma}\right] \leq \frac{\beta}{4}.$$

Finally note that $\frac{C \sin(2\beta) \log\left(\frac{1}{2\beta}\right)}{\sqrt{d}} \leq \frac{2C \sin(\beta) \log\left(\frac{1}{\beta}\right)}{\sqrt{d}}$, which clearly implies that the probability

$$\Pr_x \left[(u \cdot x)(w_{k+1} \cdot x) < 0, |w_k \cdot x| \geq \frac{C \sin(2\beta) \log\left(\frac{1}{2\beta}\right)}{\sqrt{d}}\right]$$

is at least as large as the probability

$$\Pr_x \left[(u \cdot x)(w_{k+1} \cdot x) < 0, |w_k \cdot x| \geq \frac{2C \sin(\beta) \log\left(\frac{1}{\beta}\right)}{\sqrt{d}}\right].$$

This implies that it's enough to choose $\tilde{C} = 2C$. ∎

References

[1] M. F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[2] M.-F. Balcan and A. Blum. A PAC-style model for learning from labeled and unlabeled data. In *Proceedings of the 18th Annual Conference on Computational Learning Theory (COLT)*, 2005.

[3] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*, 2007.

[4] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[5] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 201–221, 1994.

[6] S. Dasgupta. Analysis of a greedy active learning strategy. *Advances in Neural Information Processing Systems*, 17, 2004.

[7] S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

[8] S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the 18th Annual Conference on Computational Learning Theory (COLT)*, 2005.

[9] P. M. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

A Useful Facts

A well known result concerning the behavior of the uniform distribution on the sphere is the following:

Lemma A.1 *For any fixed unit vector w and any $0 < \gamma \leq 1$,*

$$\frac{\gamma}{4} \leq \Pr_x \left[|w \cdot x| \leq \frac{\gamma}{\sqrt{d}} \right] \leq \gamma,$$

where x is drawn uniformly from the unit sphere.

We state now a useful property of the normal distribution.

Lemma A.2 *Let X be a standard normal random variable. Let $\tilde{\phi}(t) = \Pr[X \geq t]$ for $t \in R$. Then: $\tilde{\phi}(t) \leq \frac{1}{\sqrt{2\pi}t} e^{-\frac{t^2}{2}}$.*

Notice Lemma A.2 implies the following:

Lemma A.3 *Let X be a standard normal random variable and let $C_1 = \frac{1}{16}$. There exists C such that for all $\beta \leq C_1$ we have:*

$$\Pr \left[X \geq \frac{\tilde{C} \log\left(\frac{1}{\beta}\right)}{\pi} \right] \leq \frac{\beta}{8}.$$