10-806 Foundations of Machine Learning and Data Science

Lecturer: Maria-Florina Balcan Lecture 22: November 23, 2015

Active Learning of Linear Separators

Preliminaries and Notation

We focus on binary classification problems; that is, we consider the problem of predicting a binary label y based on its corresponding input vector x. As in the standard machine learning formulation, we assume that the data points (x, y) are drawn from an unknown underlying distribution D_{XY} over $X \times Y$; X is called the *instance space* and Y is the *label space*. We assume that $Y = \{\pm 1\}$ and $X = \mathbb{R}^d$; we also denote the marginal distribution over X by D. Let $\mathcal{C}$ be the class of linear separators through the origin, that is $\mathcal{C} = \{\text{sign}(w \cdot x) : w \in \mathbb{R}^d, ||w|| = 1\}$. To keep the notation simple, we sometimes refer to a weight vector and the linear classifier with that weight vector interchangeably. Our goal is to output a hypothesis function $w \in \mathcal{C}$ of small error, where $\text{err}(w) = \text{err}_{D_{XY}}(w) = P_{(x,y)\sim D_{XY}}[\text{sign}(w \cdot x) \neq y]$.

Recall that in (pool-based) active learning, a set of labeled examples $(x_1, y_1) \ldots (x_m, y_m)$ is drawn i.i.d. from D_{XY}; the learning algorithm is permitted direct access to the sequence of x_i values (unlabeled data points), but has to make a label request to obtain the label y_i of example x_i. The hope is that we can output a classifier of small error by using many fewer label requests than in passive learning by actively directing the queries to informative examples (while keeping the number of unlabeled examples polynomial). For added generality, we also consider the selective sampling active learning model, where the algorithm visits the unlabeled data points x_i in sequence, and, for each i, makes a decision on whether or not to request the label y_i based only on the previously-observed x_j values ($j \leq i$) and corresponding requested labels, and never changes this decision once made. Our upper and lower bounds will apply to both selective sampling and pool-based active learning.

In the "realizable case", we assume that the labels are deterministic and generated by a target function that belongs to $\mathcal{C}$. In the non-realizable case we do not make this assumption and instead aim to compete with the best function in $\mathcal{C}$.

Given two vectors u and v and any distribution $\tilde{D}$ we denote by $\text{d}_{\tilde{D}}(u, v) = \mathbb{P}_{x\sim\tilde{D}}(\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x))$; we also denote by $\theta(u, v)$ the angle between the vectors u and v.

Log-Concave Densities

Throughout this lecture we focus on the case where the underlying distribution D is log-concave. Such distributions have played a key role in the past two decades in several areas including sampling, optimization, and integration algorithms [9], and more recently for learning theory as well [6, 7, 11].

We first summarize several results about such distributions that will be useful for our analysis.

Definition 1. *A distribution over R^d is log-concave if $\log f(\cdot)$ is concave, where f is its associated density function. It is isotropic if its mean is the origin and its covariance matrix is the identity.*

Log-concave distributions form a broad class of distributions: for example, the Gaussian, Logistic, and uniform distribution over any convex set are log-concave distributions. The following lemma summarizes known useful facts about isotropic log-concave distributions (most are from [9]; the upper bound on the density is from [7]).

Lemma 2. *Assume that D is log-concave in R^d and let f be its density function.*

(a) If D is isotropic then $\mathbb{P}_{x\sim D}[||x|| \geq \alpha\sqrt{d}] \leq e^{-\alpha+1}$. If $d = 1$ then: $\mathbb{P}_{x\sim D}[x \in [a, b]] \leq |b - a|$.

(b) If D is isotropic, then $f(x) \geq 2^{-7d}2^{9d||x||}$ whenever $0 \leq ||x|| \leq 1/9$. Furthermore, $2^{-7d} \leq f(0) \leq d(20d)^{d/2}$, and $f(x) \leq A(d)\exp(-B(d)||x||)$, where $A(d)$ is $2^{8d}d^{d/2}e$ and $B(d)$ is $\frac{2^{-7d}}{2(d-1)(20(d-1))^{(d-1)/2}}$, for all x of any norm.

(c) All marginals of D are log-concave. If D is isotropic, its marginals are isotropic as well.

(d) If D is isotropic and $d = 1$ we have $f(0) \geq 1/8$ and $f(x) \leq 1$ for all x.

We will use the fact that there exists a universal constant c such that the probability of disagreement of any two homogeneous linear separators is lower bounded by the c times the angle between their normal vectors. This follows by projecting the region of disagreement in the space given by the two normal vectors, and then using properties of log-concave distributions in 2-dimensions.

Lemma 3. *Assume D is an isotropic log-concave in R^d. Then there exists c such that for any two unit vectors u and v in $\mathbb{R}^d$ we have $c\theta(v, u) \leq \mathrm{d}_D(u, v)$.*

Proof. Consider two unit vectors u and v. Let $proj_{u,v}(x)$ denote the projection operator that, given $x \in R^d$, orthogonally projects x onto the plane determined by u and v. That is, if we define an orthogonal coordinate system in which coordinates $1, 2$ lie in this plane and coordinates $3, \ldots, d$ are orthogonal to this plane, then $x' = proj_{u,v}(x_1, \ldots, x_d) = (x_1, x_2)$. Also, given distribution D over R^d, define $proj_{u,v}(D)$ to be the distribution given by first picking $x \sim D$ and then outputting $x' = proj_{u,v}(x)$. That is, $proj_{u,v}(D)$ is just the marginal distribution over coordinates $1, 2$ in the above coordinate system. Notice that if $x' = proj_{u,v}(x)$ then $u \cdot x = u' \cdot x'$ where $u' = proj_{u,v}(u)$ and $v' = proj_{u,v}(v)$. So, if $D_2 = proj_{u,v}(D)$ then $\mathrm{d}_D(u, v) = \mathrm{d}_{D_2}(u', v')$.

By Lemma 2(c), we have that if D is isotropic and log-concave, then D_2 is as well. Let A be the region of disagreement between u' and v' intersected with the ball of radius $1/9$ in R^2. The probability mass of A under D_2 is at least the volume of A times $\inf_{x\in A} D_2(x)$. So, using Lemma 2(b)

$$\mathrm{d}_{D_2}(u', v') \geq \mathrm{vol}(A) \inf_{x\in A} D_2(x) \geq c\theta(u, v),$$

as desired. □

Analysis of the Disagreement Coefficient

Recall the definition of the disagreement coefficient. For $r > 0$, define $\mathbf{B}(w, r) = \{u \in \mathcal{C} : \mathbb{P}_D(\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(w \cdot x)) \leq r\}$. For any $\mathcal{H} \subseteq \mathcal{C}$, define the region of disagreement as $\mathrm{DIS}(\mathcal{H}) = \{x \in X : \exists w, u \in \mathcal{H} \text{ s.t. } \mathrm{sign}(u \cdot x) \neq \mathrm{sign}(w \cdot x))\}$. Define the Alexander capacity function $\mathrm{cap}_{w^*,D}(\cdot)$ for $w^* \in \mathcal{C}$ w.r.t. D as: $\mathrm{cap}_{w^*,D}(r) = \frac{\mathbb{P}_D(\mathrm{DIS}(\mathbf{B}(w^*,r)))}{r}$. Define the disagreement coefficients for $w^* \in \mathcal{C}$ w.r.t. D as: $\Theta_{w^*,D}(\epsilon) = \sup_{r\geq\epsilon}[\mathrm{cap}_{w^*,D}(r)]$.

The following is a bound on the disagreement coefficient.

Theorem 4. *Assume that D is an isotropic log-concave distribution in R^d. For any w^*, for any r, $\text{cap}_{w^*,D}(r)$ is $O(d^{1/2}\log(1/r))$. Thus $\Theta_{w^*,D}(\epsilon) = O(d^{1/2}\log(1/\epsilon))$.*

Proof. Roughly, we will show that almost all x classified by a large enough margin by w^* are not in DIS($\mathbf{B}(w^*, r)$), because all hypotheses agree with w^* about how to classify such x, and therefore all pairs of hypotheses agree with each other. Consider w such that $d(w, w^*) \leq r$; by Lemma 3 we have $\theta(w, w^*) \leq Cr$, for some constant C. For any x such that $||x|| \leq \sqrt{d}\log(1/r)$ we have

$$\begin{aligned}(w \cdot x - w^* \cdot x) &< ||w - w^*|| \times ||x|| \\ &\leq Cr\sqrt{d}\log(1/r). \quad (w, w^* \text{ are unit length so } ||w - w^*|| \leq \theta(w, w^*))\end{aligned}$$

Thus, if x also satisfies $|w^* \cdot x| \geq Cr\sqrt{d}\log(1/r)$ we have $(w^* \cdot x)(w \cdot x) > 0$. Since this is true for all w, any such x is not in DIS($\mathbf{B}(h, r)$). By Lemma 2(a) we have,

$$\mathbb{P}_{x\sim D}\Big(|w^* \cdot x| \leq Cr\sqrt{d}\log(1/r)\Big) = O(r\sqrt{d}\log(1/r)).$$

Moreover, by Lemma 2(a) we also have

$$\mathbb{P}_{x\sim D}[||x|| \geq \sqrt{d}\log(1/r)] = O(r).$$

These both imply $\text{cap}_{w^*,D}(r) = O(\sqrt{d}\log(1/r))$. □

Theorem 4 immediately leads to concrete bounds on the label complexity of several algorithms in the literature, including the one discusses last time(CAL and A^2 [5, 3, 2]) as well as others [8, 4]. For example, by composing it with a result of [4], we obtain a bound of $\tilde{O}(d^{3/2}(\log^2(1/\epsilon) + (\nu/\epsilon)^2))$ for agnostic active learning when D is isotropic log-concave in R^d; that is we only need $\tilde{O}(d^{3/2}(\log^2(1/\epsilon) + (\nu/\epsilon)^2)))$ label requests to output a classifier of error at most $\nu + \epsilon$, where $\nu = \min_{w\in\mathcal{C}} \text{err}(w)$.

Margin-based Active Learning

We now consider a more aggressive margin-based active learning algorithm for the realizable case. First, to motivate and analyze this algorithm we use the following characterization of the region of disagreement of two linear separators under a log-concave measure:

Theorem 5. *For any $c_1 > 0$, there is a $c_2 > 0$ such that the following holds. Let u and v be two unit vectors in R^d, and assume that $\theta(u, v) = \alpha < \pi/2$. If D is isotropic log-concave in R^d, then:*

$$\mathbb{P}_{x\sim D}[\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x) \text{ and } |v \cdot x| \geq c_2\alpha] \leq c_1\alpha. \tag{1}$$

We now present and analyze a margin-based active learning algorithm for the realizable case — the resulting algorithm is computationally efficient (polynomial time and label efficient).

Theorem 6. *Assume D is isotropic log-concave in R^d. There exist constants C_1, C_2, c s.t. for $d \geq 4$, and for any $\epsilon, \delta > 0$, $\epsilon < 1/4$, using Algorithm 1 with $b_k = \frac{C_1}{2^k}$ and $m_k = C_2\left(d + \ln\frac{s}{\delta}\right)$, after $s = \lceil \log_2 \frac{1}{c\epsilon} \rceil$ iterations, we find a separator of error at most ϵ with probability $1 - \delta$. The total number of labeled examples needed is $O((d + \log(1/\delta) + \log\log(1/\epsilon))\log(1/\epsilon))$.*

Algorithm 1 Margin-based Active Learning

Input: a sampling oracle for D, a labeling oracle, sequences $m_k > 0$, $k \in Z^+$ (sample sizes) and $b_k > 0$, $k \in Z^+$ (cut-off values).
Output: weight vector $\hat{w}_s$.

- Draw m_1 examples from D, label them and put them in $W(1)$.
- **iterate** $k = 1, \ldots, s$
  - find a hypothesis $\hat{w}_k$ with $\|\hat{w}_k\|_2 = 1$ consistent with all labeled examples in $W(k)$.
  - let $W(k+1) = W(k)$.
  - until m_{k+1} additional data points are labeled, draw sample x from D
    - if $|\hat{w}_k \cdot x| \geq b_k$, then reject x,
    - else, ask for label of x, and put into $W(k+1)$.

Proof. Let c be the constant from Lemma 3. We will show, using induction, that, for all $k \leq s$, with probability at least $1 - \frac{k\delta}{s}$, any $\hat{w}$ consistent with the data in the working set $W(k)$ has $\mathrm{err}(\hat{w}) \leq c2^{-k}$, so that, in particular, $\mathrm{err}(\hat{w}_k) \leq c2^{-k}$.

The case where $k = 1$ follows from the standard VC bounds (see e.g.,[10]). Assume now the claim is true for $k-1$ ($k > 1$), and consider the kth iteration. Let $S_1 = \{x : |\hat{w}_{k-1} \cdot x| \leq b_{k-1}\}$, and $S_2 = \{x : |\hat{w}_{k-1} \cdot x| > b_{k-1}\}$. By the induction hypothesis, we know that, with probability at least $1 - \frac{(k-1)\delta}{s}$, all $\hat{w}$ consistent with $W(k-1)$, including $\hat{w}_{k-1}$, have errors at most $c2^{-(k-1)}$. Consider an arbitrary such $\hat{w}$. By Lemma 3 we have $\theta(\hat{w}, w^*) \leq 2^{-(k-1)}$ and $\theta(\hat{w}_{k-1}, w^*) \leq 2^{-(k-1)}$, so $\theta(\hat{w}_{k-1}, \hat{w}) \leq 4 \times 2^{-k}$. Applying Theorem 5, there is a choice of C_1 (the constant such that $b_{k-1} = C_1/2^{k-1}$) that satisfies $\mathbb{P}((\hat{w}_{k-1} \cdot x)(\hat{w} \cdot x) < 0, x \in S_2) \leq \frac{c2^{-k}}{4}$ and $\mathbb{P}((\hat{w}_{k-1} \cdot x)(w^* \cdot x) < 0, x \in S_2) \leq \frac{c2^{-k}}{4}$. So

$$\mathbb{P}((\hat{w} \cdot x)(w^* \cdot x) < 0, x \in S_2) \leq \frac{c2^{-k}}{2}. \tag{2}$$

Now let us treat the case that $x \in S_1$. Since we are labeling m_k data points in S_1 at iteration $k-1$, standard passive-learning VC bounds (Lecture 4) imply that, if C_2 is a large enough absolute constant, then with probability $1 - \delta/s$, for all $\hat{w}$ consistent with the data in $W(k)$,

$$\mathrm{err}(\hat{w}|S_1) = \mathbb{P}((\hat{w} \cdot x)(w^* \cdot x) < 0 \mid x \in S_1) \leq \frac{c2^{-k}}{4b_k} = \frac{c}{4C_1}. \tag{3}$$

Finally, since S_1 consists of those points that, after projecting onto the direction $\hat{w}_{k-1}$, fall into an interval of length $2b_k$, Lemma 2 implies that $\mathbb{P}(S_1) \leq 2b_k$. Putting this together with (2) and (3), with probability $1 - \frac{k\delta}{s}$, we have $\mathrm{err}(\hat{w}) \leq c2^{-k}$, completing the proof. □

See [1] for an extension of this approach to obtain a computationally efficient and label efficient algorithm for learning linear separators under log-concave distributions in the agnostic case and [12] for extensions to more general concept spaces (these results are not computationally efficient however).

References

[1] Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. In *Symposium on Theory of Computing (STOC)*, pages 449–458, 2014.

[2] M. F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, 2006.

[3] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. In *ICML*, 1994.

[4] S. Dasgupta, D.J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *Advances in Neural Information Processing Systems*, 20, 2007.

[5] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

[6] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Annual Symposium on the Foundations of Computer Science (FOCS)*, 2005.

[7] A. R. Klivans, P. M. Long, and A. Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *RANDOM*, 2009.

[8] V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

[9] L. Lovasz and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 2007.

[10] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[11] S. Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *JACM*, 57(6), 2010.

[12] Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Neural Information Processing Systems (NIPS)*, 2014.