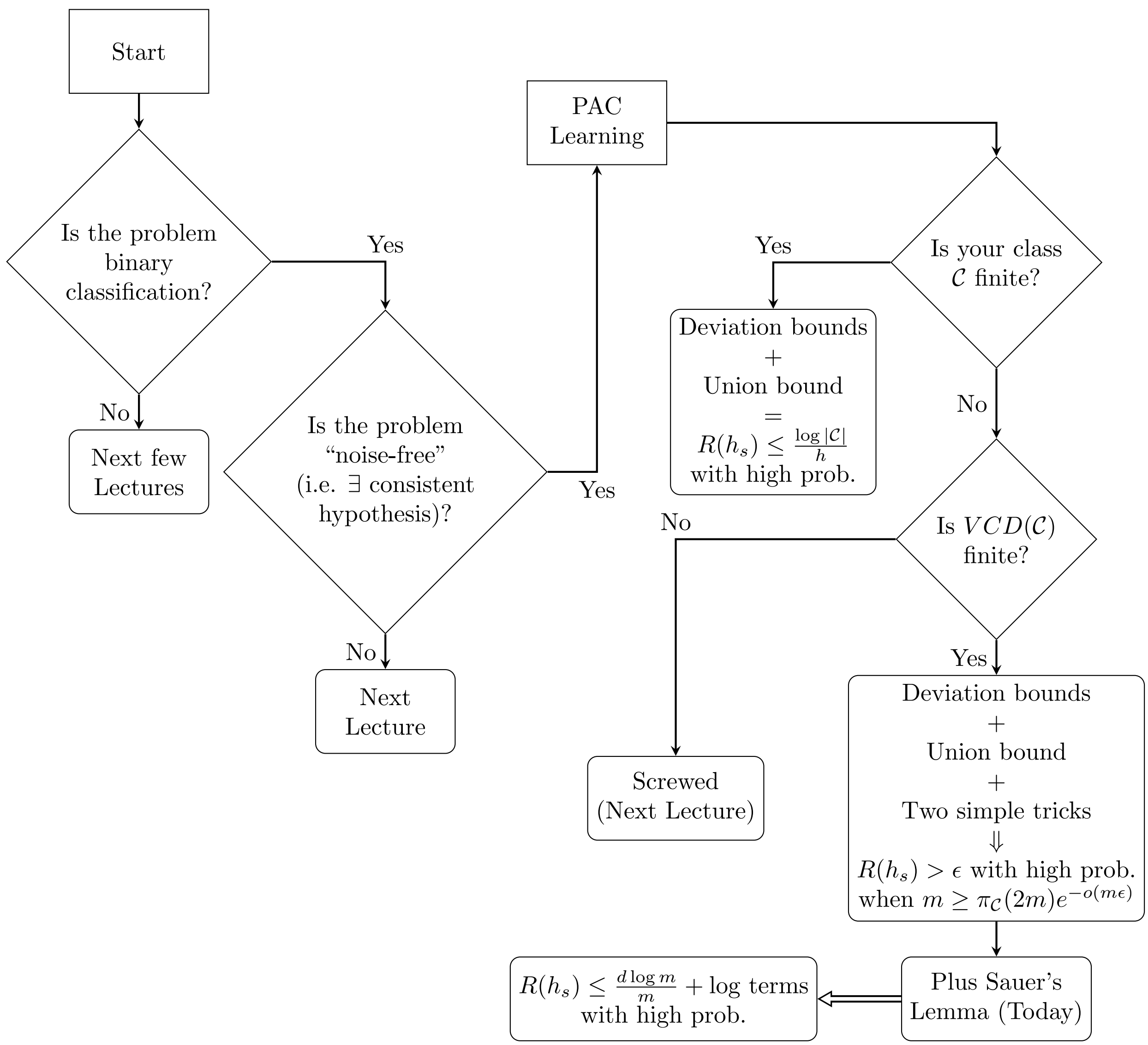
EECS 598-005: Theoretical Foundations of Machine Learning **Fall 2015**

Lecture 9: Relations between Lectures and Sauer's Lemma

Lecturer: Jacob Abernethy *Scribes: Hsu Kao, Editor: Jinqi Shen*

9.1 Relations between Lectures

Here is the Big Question: When does learning work? Which conclusions can we expect generalized from i.i.d. data samples? (dealing with non-i.i.d. samples will be the second half of the course)



9.2 Sauer's Lemma

Recall $\mathcal{C}$ is the set of functions $h : \mathcal{X} \to \{0,1\}$. Given any $S = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$, we may define the following.

Definition 9.1. *The concept class $\mathcal{C}$ restricted to set $\mathcal{S}$ is*

$$\mathcal{C}|_{\mathcal{S}} := \{(h(x_1), \ldots, h(x_m)), h \in \mathcal{C}\}.$$

Definition 9.2. *The growth function for $\mathcal{C}$ is*

$$\Pi_{\mathcal{C}}(m) = \sup\{|\mathcal{C}|_{\mathcal{S}}| : \mathcal{S} \subseteq \mathcal{X}, |\mathcal{S}| = m\}.$$

Definition 9.3. *The VC-dimension of $\mathcal{C}$ is the size of largest set that can be fully shattered by $\mathcal{C}$*

$$VCD(\mathcal{C}) = \sup\left\{d : \pi_{\mathcal{C}}(d) = 2^d\right\}.$$

Sauer's Lemma will tell us how fast $\Pi_{\mathcal{C}}(m)$ grows. Consider $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{C}$ be the class of linear threshold functions (LTFs)

$$\mathcal{C} := \left\{h_{\mathbf{w},b} : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\right\} \text{ where } h_{\mathbf{w},b}(\mathbf{x}) = sign(\mathbf{w}^\top \mathbf{x} + b).$$

Claim 9.4. $\Pi_{\mathcal{C}}(m) \leq \binom{m}{d+1} 2^{d+1}$ *when $\mathcal{C}$ is the class of LTF in dimension d.*

Proof: (a non-rigorous one) Every labelling of m points in $\mathbb{R}^d$ can be compressed to a labelling of $d+1$ points in $\mathbb{R}^d$. That is, for any set with m points, there exist certain $d+1$ points, s.t. we can infer the labels of the rest based on the certain $d+1$ points. Therefore, the number of possible labellings is less than or equal to the number of $d+1$ sized labelled subsets which is equal to $\binom{m}{d+1} 2^{d+1}$. ■

Fact 9.5. $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k = O\left(n^k\right)$.

From the above claim and fact, it follows that for the class of LTF we have $\Pi_{\mathcal{C}}(m) \leq \binom{m}{d+1} 2^{d+1} = O\left(m^{d+1}\right)$.

Theorem 9.6 (Sauer's Lemma)**.** *Assume $VCD(\mathcal{C}) = d$, then*

$$\Pi_{\mathcal{C}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} = O\left(m^d\right),$$

which says that

(i). $\Pi_{\mathcal{C}}(m) = 2^m$ *for $m \leq d$, and*

(ii). $\Pi_{\mathcal{C}}(m)$ *is polynomial on m for $m > d$.*

Proof: Take $\mathcal{S} \subseteq \mathcal{X}$, $|\mathcal{S}| = m$, $\mathcal{S} = \{x_1, \ldots, x_m\}$.

Algorithm 9.7 (Shifting Algorithm)**.**

---

for $i = 1, \ldots, m$ **do**
  for $j = 1, \ldots, |\mathcal{C}|_{\mathcal{S}}|$ **do**
    if The entry (x_i, h_j) is 1 **and** replacing it with 0 doesn't duplicate a row **then**
      Replace the entry with 0
    end if
  end for
end for

Example Consider the following table:

$T =$

	x_1	x_2	x_3	x_4	x_5
h_1	0	1	0	1	1
h_2	1	0	0	1	1
h_3	1	1	1	0	1
h_4	0	1	1	0	0
h_5	0	0	0	1	0

After performing the Shifting Algorithm, it will become:

$S =$

	x_1	x_2	x_3	x_4	x_5
h_1	0	1	0	0	0
h_2	0	0	0	1	1
h_3	0	0	0	0	1
h_4	0	0	0	0	0
h_5	0	0	0	1	0

We now claim the following:

1. All rows are unique.
   This is obvious through the algorithm.

2. If there exist a row with 1's in every columns of Q, where Q is a subset of all columns, then R shatters Q.
   If there is a row l with 1's in $i_1, ..., i_k$, there must be other $2^k - 1$ rows with all the combinations of 1 and 0 for the column $i_1, ..., i_k$, otherwise, for example, we don't have the row containing 0,1,1,...,1 for column $i_1, ..., i_k$, then at some point, the row l will be reduced to 0,1,1,...,1 in the algorithm, which is a contradiction to the fact that they are all 1. Thus, Q will be shattered.

3. $VCD(T) \leq VCD(R)$. The subclaim is that after every shift of a column, I haven't shattered any new subsets (of columns).
   To see this, take a subset of columns U containing column i. Consider shifting column i and suddenly a new row j restricted to U appears. I claim that all rows equivalent to row j restricted to U will be shifted as well. Hence, the number of unique rows on U can't increase.

Statement 2 says the number of 1's in any row is less than or equal to d. Rows are unique (by 1). Hence total number of rows is less than or equal to number of subsets of size d or fewer elements of $\mathcal{S}$, which is in fact $\sum_{i=0}^{d} \binom{m}{i}$. ■