CSE522, Winter 2011, Learning Theory	Lecture 18 - 03/03/2011
Sample compression schemes	
Lecturer: Ofer Dekel	*Scribe: Aniruddh Nath*

Definition 1. *An* ***unlabeled compression scheme*** *of size k is defined by a pair of functions:*

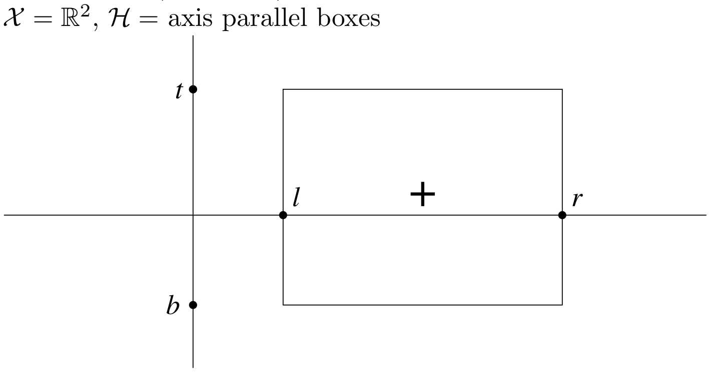
- *Compression function $c : (x \times y)^m \to \mathcal{X}^{\leq k}$*
- *Reconstruction function: $r : \mathcal{X}^{\leq k} \to \mathcal{H}$*

Definition 2. *A* ***labeled compression scheme*** *of size k is defined by a pair of functions:*

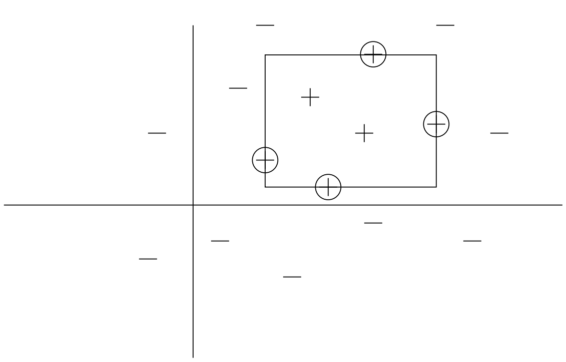
- *Compression function $c : (x \times y)^m \to (x \times y)^{\leq k}$*
- *Reconstruction function: $r : (x \times y)^{\leq k} \to \mathcal{H}$*

Definition 3. *An algorithm A is called a* ***sample compression algorithm*** *if $\exists c, r$ such that $A(S) = r(c(S))$.*

Example 1 (unlabeled)
$\mathcal{X} = \mathbb{R}^2$, $\mathcal{H} =$ axis parallel boxes

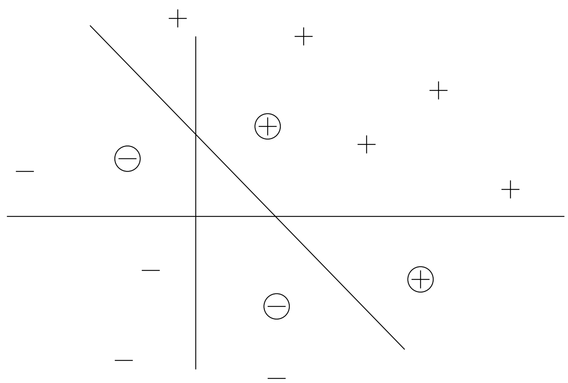


$h_{t,b,l,r}(x) = +1$ if $(l \leq x_1 \leq r \wedge b \leq x_2 \leq t)$; $\quad -1$ otherwise.



$c(S) \to \{x_l, x_r, x_t, x_b\}$
$r(x_l, x_r, x_t, x_b) \to$ smallest enclosing box

Example 2 (labeled)
Support vector machines:



$c(S) \rightarrow$ support vectors
$r(c(S)) \rightarrow$ max-margin hyperplane

Theorem 4. *Consider a sample compression algorithm A of size k; $\ell \in [0, c]$.*

$$\ell(A(S); \mathcal{D}) \leq \frac{m}{m-k} \ell(A(S); S) + c\sqrt{\frac{\log \frac{1}{\delta} + k \log \frac{em}{k}}{2m}}$$

Proof. Let $I \subseteq \{1, \ldots, m\}$, with $|I| \leq k$.
Let $h_I = r(S_I)$ where $S_I = \{(x_i, y_i)\}_{i \in I}$
Note: h_I is independent of $S_{\bar{I}}$, where $\bar{I} = \{1, \ldots, m\} \backslash I$

$$\text{By Hoeffding,} \qquad \ell(h_I; \mathcal{D}) \leq \ell(h_I; S_{\bar{I}}) + c\sqrt{\frac{log \frac{1}{\delta}}{2m}} \qquad \text{(with probability} \geq 1 - \delta) \tag{1}$$

The number of candidate output hypotheses of A is:

$$\sum_{i=0}^{k} \binom{m}{i} \leq \left(\frac{em}{k}\right)^k$$

Using union bound, equation 1 holds for all $|I| \leq k$ uniformly with probability $\geq 1 - \delta \left(\frac{em}{k}\right)^k = 1 - \delta'$.

$$\text{Therefore, with probability} \geq 1 - \delta',\ \forall |I| \leq k, \quad \ell(h_I; \mathcal{D}) \leq \ell(h_I; S_{\bar{I}}) + c\sqrt{\frac{log \frac{1}{\delta'} + k \log \frac{em}{k}}{2m}}$$

Note that $(m-k)\ell(h; S_{\bar{I}}) \leq m\ell(h; S)$. The theorem follows.

□

Definition 5. ***Realizable case:*** *$\exists h \in \mathcal{H}$ with $\ell(h; \mathcal{D}) = 0$.*

Assume: $\forall S, r(c(S))$ is consistent with S.
In other words, if $S = \{(x_i, y_i)\}_{i=1}^m$ and $h = r(c(S))$, then $\forall i\ h(x_i) = y_i$.

Theorem 6. *If ℓ is the error indicator,*

$$\Pr\left(\ell(A(S); \mathcal{D}) > \varepsilon\right) \leq \left(\frac{em}{k}\right)^k . (1-\varepsilon)^{m-k} \leq \left(\frac{em}{k}\right)^k .e^{-\varepsilon(m-k)} \equiv \delta$$

$$\Leftrightarrow \quad \textit{with probability} \geq 1 - \delta, \quad \ell(A(S); \mathcal{D}) \leq O\left(\frac{1}{m}\right)$$

Proof. Let I be a subset of $\{1, \ldots, m\}$

$$\begin{aligned}
\Pr\left((\ell(h_I; \mathcal{D}) > \varepsilon \wedge \ell(h_I; S_{\bar{I}}) = 0)\right) &\leq \Pr\left(\ell(H_I; S_{\bar{I}}) = 0 | \ell(h_I; \mathcal{D}) > \varepsilon)\right) \\
&\leq (1 - \varepsilon)^{|I|} \\
&\leq (1 - \varepsilon)^{m-k}
\end{aligned}$$

Using the union bound on $\sum_{i=0}^{k} \binom{m}{i}$ 'bad events', the theorem follows.

□

Conclusion: the size of the smallest compression scheme for $\mathcal{H}$ ($r(c, S)$ is consistent with S) behaves like $VC(\mathcal{H})$, i.e. it measures the complexity of $\mathcal{H}$.

Connection to VC

$$\text{Growth function:} \quad g_H(m) = \max_{S \subseteq \mathcal{X}^m} |\{(h(x_1), \ldots, h(x_m))\}_{h \in \mathcal{H}}|$$

Therefore, there are at most $g_H(|S|)$ ways to label S.
We also assume that one of them is consistent with S (realizability).
Sauer's lemma: $g_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}$ where $d = VC(\mathcal{H})$.

The six hundred dollar question:
Is it true that $VC(H) = d \Rightarrow \exists$ an unlabeled compression scheme of size d?
This statement is true for maximum classes. Class $\mathcal{H}$ is a maximum class if Sauer's lemma holds with equality.

Connection to PAC-Bayes

Redefine the reconstruction function $r : \mathcal{X}^{\leq k} \times \mathcal{M} \to \mathcal{H}$, where $\mathcal{M}$ is a set of **message strings**.

Case 1: $r(S_I, \sigma)$ ignores $\sigma \to$ back to original definition of compression schemes.
Case 2: $r(S_I, \sigma)$ ignores $S_I \to$ back to standard statistical learning.

We decompose the prior $\Pr(I, \sigma) = \Pr(I) \Pr(\sigma | I)$

- If $|I| = |I'|$, the prior should not differentiate between them, i.e. $\Pr(I) = \frac{\Pr(|I|)}{\binom{m}{|I|}}$
- $\Pr(i) = 0$ for $i > k$.

Theorem 7. *(c, r) is a compression scheme (with message) of size k, for any prior P (expressed as above). $\forall \mathcal{D}$, $\forall \delta > 0$, with probability $\leq 1 - \delta$ over $S \sim \mathcal{D}^m$, $\forall Q$ (posteriors),*

$$KL(\ell(Q; S) || \ell(Q; \mathcal{D})) \leq \frac{KL(Q||P) + \ln \frac{m+1}{\delta}}{m - k}$$