Compression-Schemes for Real-Valued Learners

Meni Sadigurschi
March 24, 2018

Joint work with
Prof. Aryeh Kontorovich BGU
Dr. Steve Hanneke TTIC

Ben-Gurion University of the Negev

What are compression schemes?

A *k-sample compression scheme* (κ, ρ) for a hypothesis class $\mathcal{C}$

sample

What are compression schemes?

A *k-sample compression scheme* (κ, ρ) for a hypothesis class $\mathcal{C}$

$$\text{sample} \xrightarrow{\kappa} \text{sub-sample, information}$$

What are compression schemes?

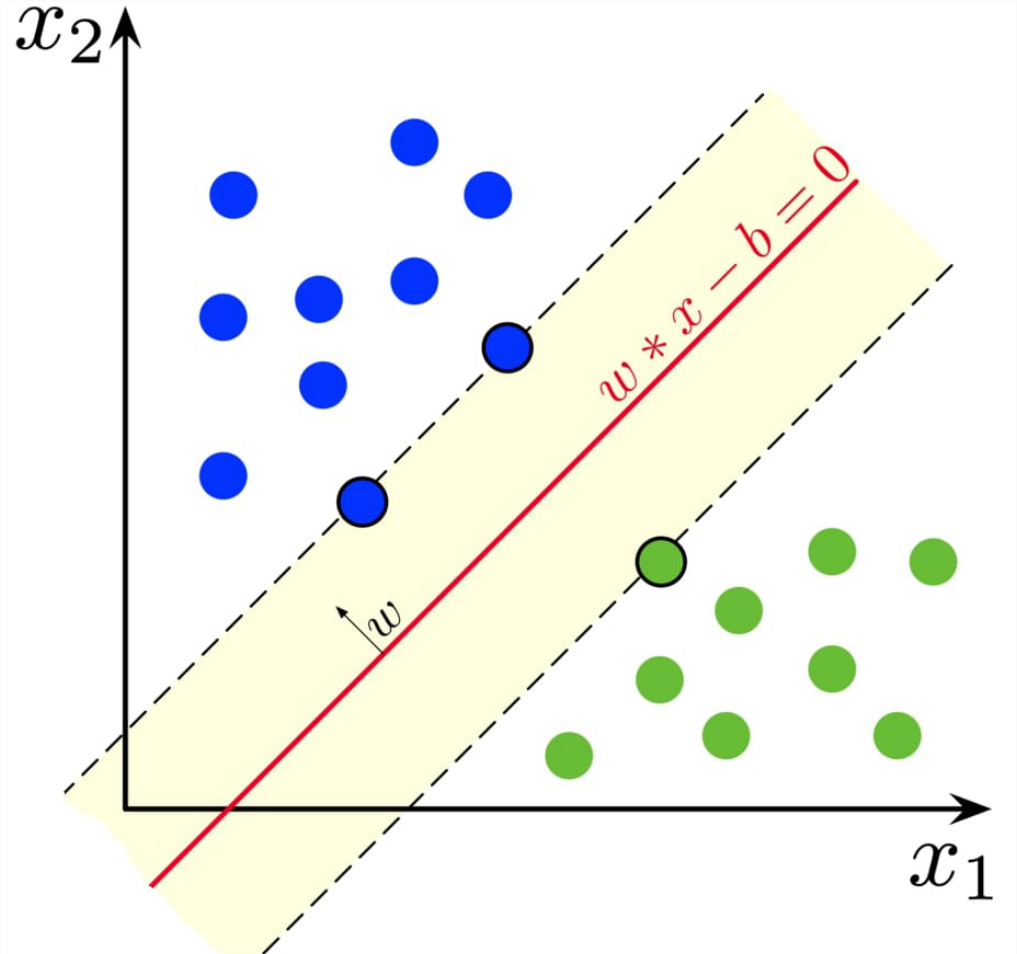
A *k-sample compression scheme* (κ, ρ) for a hypothesis class $\mathcal{C}$

$$\text{sample} \xrightarrow{\kappa} \text{sub-sample, information} \xrightarrow{\rho} \text{hypothesis}$$

What are compression schemes?

A *k-sample compression scheme* (κ, ρ) for a hypothesis class $\mathcal{C}$

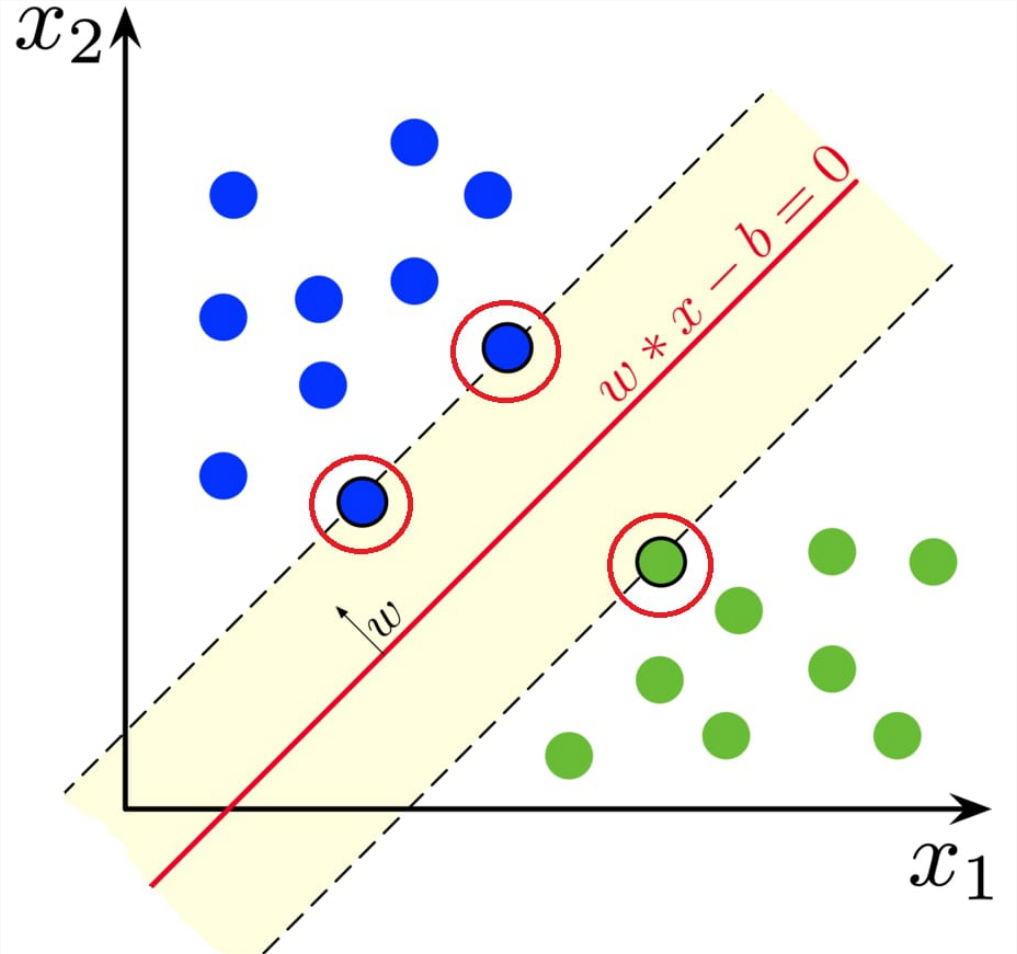
$$\text{sample} \xrightarrow{\kappa} \text{sub-sample, information} \xrightarrow{\rho} \text{hypothesis}$$



What are compression schemes?

A *k-sample compression scheme* (κ, ρ) for a hypothesis class $\mathcal{C}$

$$\text{sample} \xrightarrow{\kappa} \text{sub-sample, information} \xrightarrow{\rho} \text{hypothesis}$$



Milestones

Theorem (Littlestone and Warmuth [1986])

Compressibility $\Rightarrow$ *Learnability.*

Milestones

Theorem (Littlestone and Warmuth [1986])

Compressibility $\Rightarrow$ *Learnability.*

Conjecture (Floyd and Warmuth [1995])

Compressibility $\Leftarrow$ *Learnability.*

Milestones

Theorem (Littlestone and Warmuth [1986])

Compressibility $\Rightarrow$ *Learnability.*

Conjecture (Floyd and Warmuth [1995])

Compressibility $\Leftarrow$ *Learnability.*

Conjecture (Warmuth [2003])

The size of the compression is linear in the VC-dimension of the class.

And finally...

Theorem (Moran and Yehudayoff [2016])

For a binary function class $\mathcal{C}$ of $VC(\mathcal{C}) = d$, there exist a $\mathcal{O}\left(d2^d\right)$-compression scheme.

And finally...

Theorem (Moran and Yehudayoff [2016])

For a binary function class $\mathcal{C}$ of $VC(\mathcal{C}) = d$, there exist a $\mathcal{O}\left(d2^d\right)$-compression scheme.

What about regression?

Do finite Fat-Shattering dimension implies the class is compressible into constant size?

What about regression?

Do finite Fat-Shattering dimension implies the class is compressible into constant size?

Boosting

But how do you even boost real-valued functions?

But how do you even boost real-valued functions?

Definition ("Standard"-Weak-Hypothesis)

For $\gamma \in [0, 1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is a γ-*weak hypothesis* (with respect to distribution $\mathcal{D}$ and target $f^* \in \mathcal{C}$) if

$$\mathbb{E}_{X \sim \mathcal{D}} [\ell(f_S(x), f^*(x))] \leq \frac{1}{2} - \gamma.$$

But how do you even boost real-valued functions?

Definition ("Standard"-Weak-Hypothesis)

For $\gamma \in [0, 1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is a γ-*weak hypothesis* (with respect to distribution $\mathcal{D}$ and target $f^* \in \mathcal{C}$) if

$$\mathbb{E}_{X \sim \mathcal{D}} [\ell(f_S(x), f^*(x))] \leq \frac{1}{2} - \gamma.$$

A better notion

A better notion

Definition ((η, γ)-Weak-Hypothesis)

For $\eta \in [0, 1]$ and $\gamma \in [0, 1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is an (η, γ)-*weak hypothesis* (with respect to distribution $\mathcal{D}$ and target $f^* \in \mathcal{C}$) if

$$\Pr_{X \sim \mathcal{D}}(|f(X) - f^*(X)| > \eta) \leq \frac{1}{2} - \gamma.$$

A better notion

Definition ((η, γ)-Weak-Hypothesis)

For $\eta \in [0, 1]$ and $\gamma \in [0, 1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is an (η, γ)-*weak hypothesis* (with respect to distribution $\mathcal{D}$ and target $f^* \in \mathcal{C}$) if

$$\Pr_{X \sim \mathcal{D}}(|f(X) - f^*(X)| > \eta) \leq \frac{1}{2} - \gamma.$$

Where can Good-Models be found?

Where can Good-Models be found?

Theorem (Hanneke, Kontorovich, S)

For any $\eta, \delta, \beta \in (0,1)$, *letting* $X_1, \ldots, X_m$ *be i.i.d.* $\mathcal{D}$*-distributed, where*

$$m = \mathcal{O}\left(\frac{Fat_{c\eta\beta}(\mathcal{C})}{\beta} \ln \frac{1}{\eta\beta} + \frac{1}{\beta} \ln \frac{1}{\delta}\right)$$

with probability at least $1 - \delta$, *every* $f \in \mathcal{C}$ *with*

$$\max_{i \in [m]} |f(X_i) - f^*(X_i)| \leq \eta/2$$

satisfies

$$\Pr(x : |f(x) - f^*(x)| > \eta) \leq \beta$$

Now we boost!

Algorithm 1 - `MedBoost` [Kégl]

1: Input: $\mathcal{A}$, S, T, γ, η
2: **for** $t = 0, \ldots, T$ **do**
3: Sample sub sample S'
4: Learn h_t using S' and the weak-learner $\mathcal{A}$
5: Update the distribution on S
6: Assign weight α_t to h_t
7: Return *ensemble* $(h_1, \ldots, h_T)$ and weights $(\alpha_1, \ldots, \alpha_T)$

So....

So....

1. Use the ERM on $\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C})\right)$ samples to get weak-learners.

So....

1. Use the ERM on $\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C})\right)$ samples to get weak-learners.
2. Use `MedBoost` with those weak-learners to achieve approximate uniform consistency using only ...
   $\log(m)$ learners

So....

1. Use the ERM on $\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C})\right)$ samples to get weak-learners.
2. Use `MedBoost` with those weak-learners to achieve approximate uniform consistency using only ...
   $\log(m)$ learners

We got compression of size

$$\widetilde{\mathcal{O}}\left(log(m)Fat_{c\eta}(\mathcal{C})\right)!$$

Wait!

We said we are aiming for size which is **independent** on the sample size!!!

Lets make it sparse

Algorithm 2 - Sparsify

Run $\texttt{MedBoost}(\{(x_i, y_i)\}_{i\in[m]}, T, \gamma, \eta)$
Let $h_1, \ldots, h_T$ and $\alpha_1, \ldots, \alpha_T$ be its return values
Denote $\alpha'_t = \alpha_t / \sum_{t'=1}^{T} \alpha_{t'}$ for each $t \in [T]$
repeat
    Sample $(J_1, \ldots, J_n) \sim Cat(\alpha'_1, \ldots, \alpha'_T)^n$
    Let $F = \{h_{J_1}, \ldots, h_{J_n}\}$
until $\max_{1\le i\le m} |\{f \in F : |f(x_i) - y_i| > \eta\}| < n/2$
Return F

Lets make it sparse

Theorem (Hanneke, Kontorovich, S)

Choosing

$$n = \Theta\left(\frac{1}{\gamma^2} Fat_{c\eta}(\mathcal{C}^*) \log^2 \frac{Fat_{c\eta}(\mathcal{C}^*)}{\eta}\right)$$

suffices for the `Sparsify` *procedure to return* $\{f_1, \ldots, f_n\}$ *with*

$$\max_{1 \le i \le m} |Med(f_1(x_i), \ldots, f_n(x_i)) - y_i| \le \eta.$$

Summing up

To sum up

Summing up

To sum up

1. Use the ERM on $\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C})\right)$ samples to get weak-learners.

Summing up

To sum up

1. Use the ERM on $\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C})\right)$ samples to get weak-learners.
2. Use `MedBoost` with those weak-learners to achieve approximate uniform consistency

Summing up

To sum up

1. Use the ERM on $\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C})\right)$ samples to get weak-learners.
2. Use `MedBoost` with those weak-learners to achieve approximate uniform consistency
3. Sparsify the ensemble to one of size $\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C}^*)\right)$

Summing up

To sum up

1. Use the ERM on $\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C})\right)$ samples to get weak-learners.
2. Use `MedBoost` with those weak-learners to achieve approximate uniform consistency
3. Sparsify the ensemble to one of size $\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C}^*)\right)$

We got compression of constant size

$$\widetilde{\mathcal{O}}\left(Fat_{c\eta}(\mathcal{C})Fat_{c\eta}(\mathcal{C}^*)\right)$$

Any Questions?

References

Sally Floyd and Manfred K. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Balázs Kégl. Robust regression by boosting the median. In *COLT*, 2003.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. Technical report, Department of Computer and Information Sciences, Santa Cruz, CA, Ju, 1986.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3):21:1–21:10, 2016. doi: 10.1145/2890490. URL `http://doi.acm.org/10.1145/2890490`.

Manfred K. Warmuth. Compressing to VC dimension many points. In *Proceedings of the 16th Conference on Learning Theory*, 2003.