Optimal learners for multiclass problems

Amit Daniely
Joint work with Shai Shalev-Shwartz
The Hebrew University of Jerusalem

June 15, 2014
Basic problem: Statistical learning of a hypothesis class $\mathcal{H} \subset \mathcal{Y}^\mathcal{X}$
Basic problem: Statistical learning of a hypothesis class $\mathcal{H} \subset \mathcal{Y}^X$

- Capture a variety of problems (Speech recognition, Object categorization, ...)
- Many methods (One vs All, Multiclass SVM, Error Correcting Output Codes, Structured output prediction,...)
- Extensive theoretical and non-theoretical study, yet, not sufficiently understood.
Multiclass classification – what is learnable? and how?

**Basic problem:** *Statistical* learning of a hypothesis class $\mathcal{H} \subset \mathcal{Y}^X$

- Capture a variety of problems (Speech recognition, Object categorization, ...)
- Many methods (One vs All, Multiclass SVM, Error Correcting Output Codes, Structured output prediction, ...)
- Extensive theoretical and non-theoretical study, yet, not sufficiently understood.

**Basic questions**

- When $\mathcal{H}$ is learnable?
- What is the *sample complexity* of $\mathcal{H}$?
- How to learn $\mathcal{H}$ optimally?
The fundamental theorem for binary classification (VC, 71)

- When $\mathcal{H}$ is learnable? $\text{VC}(\mathcal{H}) < \infty$
- What is the sample complexity of $\mathcal{H}$? $\tilde{\Theta}\left(\frac{\text{VC}(\mathcal{H})}{\epsilon}\right)$
- How to learn $\mathcal{H}$ optimally? Use ERM
The fundamental theorem for binary classification (VC, 71)

- When $\mathcal{H}$ is learnable? $\text{VC}(\mathcal{H}) < \infty$
- What is the sample complexity of $\mathcal{H}$? $\tilde{\Theta}\left(\frac{\text{VC}(\mathcal{H})}{\epsilon}\right)$
- How to learn $\mathcal{H}$ optimally? Use ERM

- So, what about multiclass classification?
The fundamental theorem for binary classification (VC, 71)

- When $\mathcal{H}$ is learnable? $VC(\mathcal{H}) < \infty$
- What is the sample complexity of $\mathcal{H}$? $\tilde{\Theta}\left(\frac{VC(\mathcal{H})}{\epsilon}\right)$
- How to learn $\mathcal{H}$ optimally? Use ERM

So, what about multiclass classification?

- Gaps between ERMs – Some ERMs are suboptimal! (DSSB 2011)
The fundamental theorem for binary classification (VC, 71)

- When $\mathcal{H}$ is learnable? $\text{VC}(\mathcal{H}) < \infty$
- What is the sample complexity of $\mathcal{H}$? $\tilde{\Theta}\left(\frac{\text{VC}(\mathcal{H})}{\epsilon}\right)$
- How to learn $\mathcal{H}$ optimally? Use ERM

So, what about multiclass classification?

- Gaps between ERMs – Some ERMs are suboptimal! (DSSB 2011)
- Optimal learning cannot be proper. Must output $h \notin \mathcal{H}$!
The fundamental theorem for binary classification (VC, 71)

- When $\mathcal{H}$ is learnable? $\text{VC}(\mathcal{H}) < \infty$
- What is the sample complexity of $\mathcal{H}$? $\tilde{\Theta}\left(\frac{\text{VC}(\mathcal{H})}{\epsilon}\right)$
- How to learn $\mathcal{H}$ optimally? Use ERM

So, what about multiclass classification?

- Gaps between ERMs – Some ERMs are suboptimal! (DSSB 2011)
- Optimal learning cannot be proper. Must output $h \notin \mathcal{H}$!
- Gaps even in structured output prediction!
The fundamental theorem for binary classification (VC, 71)

- When $\mathcal{H}$ is learnable? $\text{VC}(\mathcal{H}) < \infty$
- What is the sample complexity of $\mathcal{H}$? $\tilde{\Theta}\left(\frac{\text{VC}(\mathcal{H})}{\epsilon}\right)$
- How to learn $\mathcal{H}$ optimally? Use ERM

- So, what about multiclass classification?

Gaps between ERMs – Some ERMs are suboptimal! (DSSB 2011)
- Optimal learning cannot be proper. Must output $h \notin \mathcal{H}$!
- Gaps even in structured output prediction!

- “Reopen” the basic questions.
The fundamental theorem for binary classification (VC, 71)

- When $\mathcal{H}$ is learnable? $VC(\mathcal{H}) < \infty$
- What is the sample complexity of $\mathcal{H}$? $\tilde{\Theta}\left(\frac{VC(\mathcal{H})}{\epsilon}\right)$
- How to learn $\mathcal{H}$ optimally? Use ERM

So, what about multiclass classification?

- Gaps between ERMs – Some ERMs are suboptimal! (DSSB 2011)
- Optimal learning cannot be proper. Must output $h \notin \mathcal{H}$!
- Gaps even in structured output prediction!

- “Reopen” the basic questions.

- The one inclusion algorithm (RBR, 06) is optimal!
The fundamental theorem for binary classification (VC, 71)

- When $\mathcal{H}$ is learnable? $VC(\mathcal{H}) < \infty$
- What is the sample complexity of $\mathcal{H}$? $\tilde{\Theta}\left(\frac{VC(\mathcal{H})}{\epsilon}\right)$
- How to learn $\mathcal{H}$ optimally? Use ERM

So, what about multiclass classification?

- Gaps between ERMs – Some ERMs are suboptimal! (DSSB 2011)
- Optimal learning cannot be proper. Must output $h \notin \mathcal{H}$!
- Gaps even in structured output prediction!

- “Reopen” the basic questions.

- The one inclusion algorithm (RBR, 06) is optimal!
- The sample complexity is characterized by a sequence $\mu_\mathcal{H}(m)$. 
The fundamental theorem for binary classification (VC, 71)

- When $\mathcal{H}$ is learnable? $\text{VC}(\mathcal{H}) < \infty$
- What is the sample complexity of $\mathcal{H}$? $\tilde{\Theta}\left(\frac{\text{VC}(\mathcal{H})}{\epsilon}\right)$
- How to learn $\mathcal{H}$ optimally? Use ERM

- So, what about multiclass classification?

Gaps between ERMs – Some ERMs are suboptimal! (DSSB 2011)
- Optimal learning cannot be proper. Must output $h \notin \mathcal{H}$!
- Gaps even in structured output prediction!

- “Reopen” the basic questions.

The one inclusion algorithm (RBR, 06) is optimal!
- The sample complexity is characterized by a sequence $\mu_{\mathcal{H}}(m)$.
- New dimension is conjectured to characterize the sample complexity.
1. Optimal learner must be improper!

2. An optimal multiclass learner

3. Characterizing multiclass learnability
Goal: learn $h^* \in \mathcal{H}$ based on $S_m = \{(x_i, h^*(x_i))\}^m_{i=1}$ where $x_i \sim \mathcal{D}$

Error of $h$: $\text{Err}(h) = \Pr[h(x) \neq h^*(x)]$

Learner: $\mathcal{A} : \bigcup_m (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{Y}^\mathcal{X}$

- ERM learner: always return a consistent hypothesis $h \in \mathcal{H}$.
- Proper learner: always return $h \in \mathcal{H}$. 
Goal: learn \( h^* \in \mathcal{H} \) based on \( S_m = \{(x_i, h^*(x_i))\}_{i=1}^m \) where \( x_i \sim \mathcal{D} \)

Error of \( h \): \( \text{Err}(h) = \Pr[h(x) \neq h^*(x)] \)

Learner: \( \mathcal{A} : \bigcup_m (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{Y}^\mathcal{X} \)

- ERM learner: always return a consistent hypothesis \( h \in \mathcal{H} \).
- Proper learner: always return \( h \in \mathcal{H} \).

(PAC) Sample complexity of \( \mathcal{A} \): \( m_{\mathcal{A}}(\epsilon) \) is the minimal number \( m \) such that, w.p. \( \geq 1/2 \), \( \text{Err}(\mathcal{A}(S_m)) \leq \epsilon \).
Setting and notation

- **Goal:** learn $h^* \in \mathcal{H}$ based on $S_m = \{(x_i, h^*(x_i))\}_{i=1}^{m}$ where $x_i \sim D$
- **Error of $h$:** $\text{Err}(h) = \Pr[h(x) \neq h^*(x)]$
- **Learner:** $\mathcal{A} : \cup_{m}(X \times Y)^m \rightarrow Y^X$
  - **ERM learner:** always return a consistent hypothesis $h \in \mathcal{H}$.
  - **Proper learner:** always return $h \in \mathcal{H}$.
- **(PAC) Sample complexity of $\mathcal{A}$:** $m_\mathcal{A}(\epsilon)$ is the minimal number $m$ such that, w.p. $\geq 1/2$, $\text{Err}(\mathcal{A}(S_m)) \leq \epsilon$.
- **(PAC) Sample complexity of $\mathcal{H}$:** $m_\mathcal{H}(\epsilon) = \min_\mathcal{A} m_\mathcal{A}(\epsilon)$. 
Optimal learner must be improper

- **Improper** learners are often used for computational reasons.
- Surprisingly, we show that in multiclass classification, optimal learner (even computationally unbounded) **must** be improper.
The Cantor Class

The Cantor class

- $\mathcal{X}$ – an arbitrary set, $\mathcal{Y} = 2^\mathcal{X} \cup \{\ast\}$
- For $T \subset \mathcal{X}$, let,
  \[
  h_T(x) = \begin{cases} 
  \ast & x \notin T \\
  T & x \in T 
  \end{cases}
  \]
- $\mathcal{H} = \{h_T : |T| = \frac{|\mathcal{X}|}{2}\}$
The Cantor Class

The Cantor class

- \( \mathcal{X} \) – an arbitrary set, \( \mathcal{Y} = 2^{\mathcal{X}} \cup \{ \ast \} \)
- For \( T \subset \mathcal{X} \), let,
  \[
  h_T(x) = \begin{cases} 
  \ast & x \notin T \\
  T & x \in T
  \end{cases}
  \]
- \( \mathcal{H} = \{ h_T : |T| = \frac{|\mathcal{X}|}{2} \} \)

Suppose that a learner gets a sample \( \{(x_i, y_i)\}_{i=1}^m \) labelled by some (unknown) \( h_T \in \mathcal{H} \).
The Cantor Class

The Cantor class

- $\mathcal{X}$ – an arbitrary set, $\mathcal{Y} = 2^\mathcal{X} \cup \{\ast\}$
- For $T \subset \mathcal{X}$, let,
  
  $$h_T(x) = \begin{cases} 
  \ast & x \notin T \\
  T & x \in T 
  \end{cases}$$

- $\mathcal{H} = \{h_T : |T| = \frac{\mathcal{X}}{2}\}$

- Suppose that a learner gets a sample $\{(x_i, y_i)\}_{i=1}^{m}$ labelled by some (unknown) $h_T \in \mathcal{H}$.
- If $y_i = T$ for some $i$, it knows that the learnt hypothesis is $h_T$. 
The Cantor Class

The Cantor class

- $\mathcal{X}$ – an arbitrary set, $\mathcal{Y} = 2^\mathcal{X} \cup \{\ast\}$
- For $T \subset \mathcal{X}$, let,

$$h_T(x) = \begin{cases} 
\ast & x \notin T \\
T & x \in T 
\end{cases}$$

- $\mathcal{H} = \{h_T : |T| = \frac{\mathcal{X}}{2}\}$

Suppose that a learner gets a sample $\{(x_i, y_i)\}_{i=1}^m$ labelled by some (unknown) $h_T \in \mathcal{H}$.

If $y_i = T$ for some $i$, it knows that the learnt hypothesis is $h_T$.

Therefore, a learning algorithm is fully determined by its output on samples of the form

$$(x_1, \ast), \ldots, (x_m, \ast)$$
Optimal learners must be improper

Theorem

- $m_H \leq \frac{1}{\epsilon}$.
- For every proper algorithm, $m_A \geq \frac{|X|}{\epsilon}$.
Optimal learners must be improper

Theorem

- \( m_H \leq \frac{1}{\epsilon} \).
- For every proper algorithm, \( m_A \geq \frac{|X|}{\epsilon} \).

- Similar phenomenon (slightly weaker, namely, gaps between ERMs) happens in classes that are used in practice.
Gaps between ERMs

**Theorem**

- \( m_{\mathcal{H}} \leq \frac{1}{\epsilon} \).
- For every proper algorithm, \( m_{\mathcal{A}} \geq \frac{|X|}{\epsilon} \).

**Proof. (sketch).**

- **Claim:** \( m_{\mathcal{H}} \leq \frac{1}{\epsilon} \)
Gaps between ERMs

Theorem

- \( m_H \leq \frac{1}{\epsilon} \).
- For every proper algorithm, \( m_A \geq \frac{|X|}{\epsilon} \).

Proof. (sketch).

- **Claim:** \( m_H \leq \frac{1}{\epsilon} \)
  - Suppose \( A \) return \( h_\emptyset \) on the a sample \((x_1, \ast), \ldots, (x_m, \ast)\).
Theorem

- \( m_\mathcal{H} \leq \frac{1}{\epsilon} \).
- For every proper algorithm, \( m_\mathcal{A} \geq \frac{|X|}{\epsilon} \).

Proof. (sketch).

- **Claim:** \( m_\mathcal{H} \leq \frac{1}{\epsilon} \)
  - Suppose \( \mathcal{A} \) return \( h_\emptyset \) on the a sample \((x_1, \ast), \ldots, (x_m, \ast)\).
  - Let \( h_T \) be the target hypothesis
Gaps between ERMs

Theorem

- $m_{\mathcal{H}} \leq \frac{1}{\epsilon}$.
- For every proper algorithm, $m_{\mathcal{A}} \geq \frac{|X|}{\epsilon}$.

Proof. (sketch).

- **Claim:** $m_{\mathcal{H}} \leq \frac{1}{\epsilon}$
  - Suppose $\mathcal{A}$ return $h_0$ on the sample $(x_1, \ast), \ldots, (x_m, \ast)$.
  - Let $h_T$ be the target hypothesis.
  - $\mathcal{A}$ will return either $h_T$ or $h_0$. 
**Theorem**

- \( m_H \leq \frac{1}{\epsilon} \).
- For every proper algorithm, \( m_A \geq \frac{|X|}{\epsilon} \).

**Proof. (sketch).**

- **Claim:** \( m_H \leq \frac{1}{\epsilon} \)
  - Suppose \( A \) return \( h_0 \) on the a sample \( (x_1,*) , \ldots , (x_m,*) \).
  - Let \( h_T \) be the target hypothesis
  - \( A \) will return either \( h_T \) or \( h_0 \).
  - If \( \text{Err}(h_0) \geq \epsilon \), it will be rejected w.h.p. using \( \frac{1}{\epsilon} \) examples.
Gaps between ERMs

Proof. (sketch, for $\epsilon = \frac{1}{10}$.)

...
Proof. (sketch, for $\epsilon = \frac{1}{10}$).

Claim: For any proper $\mathcal{A}$, $m_{\mathcal{A}} \geq |\mathcal{X}|$. 

Let $D$ be uniform on some $E \subset \mathcal{X}$, $|E| = \frac{|\mathcal{X}|}{2}$, and let the target classifier be $h_{\mathcal{X} \setminus E}$. $A$ will choose some $h_{T}$ with $|T| = \frac{|\mathcal{X}|}{2}$. Therefore, to have a small error, $T$ should almost coincide with $E^c$. Requires $\Omega(|\mathcal{X}|)$ examples.
Gaps between ERMs

Proof. (sketch, for $\epsilon = \frac{1}{10}$).

- **Claim**: For any proper $\mathcal{A}$, $m_A \geq |\mathcal{X}|$.
  - Let $\mathcal{D}$ be uniform on some $E \subset \mathcal{X}$, $|E| = \frac{|\mathcal{X}|}{2}$, and let the target classifier be $h_{\mathcal{X}\setminus E}$.
Proof. (sketch, for $\epsilon = \frac{1}{10}$).

**Claim:** For any proper $\mathcal{A}$, $m_{\mathcal{A}} \geq |\mathcal{X}|$.

- Let $\mathcal{D}$ be uniform on some $E \subset \mathcal{X}$, $|E| = |\mathcal{X}|/2$, and let the target classifier be $h_{\mathcal{X}\setminus E}$.
- $\mathcal{A}$ will choose some $h_T$ with $|T| = |\mathcal{X}|/2$.

```latex
\begin{figure}[h]
\centering
\begin{tikzpicture}
  \draw[fill=black!30] (0,0) circle (1cm) node {E};
  \draw[fill=white] (2,0) circle (1cm) node {T};
  \draw[ultra thick] (0,0) rectangle (2,2);
  \draw[fill=black!30] (0.5,0.5) circle (0.5cm) node {X};
\end{tikzpicture}
\end{figure}
```
Proof. (sketch, for $\epsilon = \frac{1}{10}$).

**Claim:** For any proper $\mathcal{A}$, $m_{\mathcal{A}} \geq |\mathcal{X}|$.

- Let $\mathcal{D}$ be uniform on some $E \subset \mathcal{X}$, $|E| = |\mathcal{X}|/2$, and let the target classifier be $h_{\mathcal{X}\setminus E}$
- $\mathcal{A}$ will choose some $h_T$ with $|T| = |\mathcal{X}|/2$
- $\text{Err}_D(h_T) = |T \cap E|/|E|$. Therefore, to have a small error, $T$ should almost coincide with $E^c$. 

![Diagram](image-url)
Gaps between ERMs

Proof. (sketch, for $\epsilon = \frac{1}{10}$).

- **Claim:** For any proper $\mathcal{A}$, $m_\mathcal{A} \geq |\mathcal{X}|$.
  - Let $\mathcal{D}$ be uniform on some $E \subset \mathcal{X}$, $|E| = |\mathcal{X}|/2$, and let the target classifier be $h_{\mathcal{X}\setminus E}$
  - $\mathcal{A}$ will choose some $h_T$ with $|T| = |\mathcal{X}|/2$
  - $\text{Err}_\mathcal{D}(h_T) = |T \cap E|/|E|$. Therefore, to have a small error, $T$ should almost coincide with $E^c$.
  - Requires $\Omega(|\mathcal{X}|)$ examples.
Outline

1. Optimal learner must be improper!

2. An optimal multiclass learner

3. Characterizing multiclass learnability
An optimal multiclass learner

- Haussler, Littlestone, Warmuth (1994) proposed an (improper) learner for binary classification based on the “one inclusion graph”
- Rubinstein, Bartlett, Rubinstein (2006) generalized it to multiclass problems using a “one inclusion hyper-graph”
- The analysis of RBR showed optimality up to a factor of $\log(|\mathcal{Y}|)$.
- By a new analysis, we show optimality up to a constant factor.
The one-inclusion algorithm

- The **mistake bound**, $\epsilon_A(m)$, of an algorithm $A$ is the probability that $A$ errs on a new example after observing $m$ examples.
The one-inclusion algorithm

- The **mistake bound**, $\epsilon_A(m)$, of an algorithm $A$ is the probability that $A$ errs on a new example after observing $m$ examples.

**Theorem**

Let $I$ be the one-inclusion algorithm. For every $A$, $\epsilon_A(m) \geq \frac{1}{2e} \epsilon_I(m)$
The one-inclusion algorithm

The **mistake bound**, $\epsilon_A(m)$, of an algorithm $A$ is the probability that $A$ errs on a new example after observing $m$ examples.

**Theorem**

Let $I$ be the one-inclusion algorithm. For every $A$, $\epsilon_A(m) \geq \frac{1}{2e} \epsilon_I(m)$

- Improves on RBR who had a factor of $\log(|Y|)$ instead of $2e \approx 5.44$. 

Amit Daniely (Hebrew U) Multiclass Learning June 15, 2014 13 / 21
The one-inclusion algorithm

- The **mistake bound**, $\epsilon_A(m)$, of an algorithm $A$ is the probability that $A$ errs on a new example after observing $m$ examples.

**Theorem**

Let $I$ be the one-inclusion algorithm. For every $A$, $\epsilon_A(m) \geq \frac{1}{2e} \epsilon_I(m)$

- Improves on RBR who had a factor of $\log(|Y|)$ instead of $2e \approx 5.44$.
- By a standard argument the one inclusion algorithm is optimal in the PAC model as well, up to a factor of $\log \left( \frac{1}{\epsilon} \right)$.
The one-inclusion algorithm

- The **mistake bound**, $\epsilon_A(m)$, of an algorithm $A$ is the probability that $A$ errrs on a new example after observing $m$ examples.

**Theorem**

Let $I$ be the one-inclusion algorithm. For every $A$, $\epsilon_A(m) \geq \frac{1}{2e}\epsilon_I(m)$

- Improves on RBR who had a factor of $\log(|Y|)$ instead of $2e \approx 5.44$.
- By a standard argument the one inclusion algorithm is optimal in the PAC model as well, up to a factor of $\log\left(\frac{1}{\epsilon}\right)$.
- We derive **efficient** algorithms for *linear classes*.
Basic questions

- When $\mathcal{H}$ is learnable?
- What is the sample complexity of $\mathcal{H}$?
- How to learn $\mathcal{H}$ optimally? Use the one inclusion algorithm.
1. Optimal learner must be improper!

2. An optimal multiclass learner

3. Characterizing multiclass learnability
The density function

- We define the **degree** of \( h \in \mathcal{H} \) w.r.t. \( \mathcal{H} \) as the number of points \( x \in \mathcal{X} \) for which there exists \( g \in \mathcal{H} \) such that \( g \) disagree with \( h \) only on \( x \).
The density function

- We define the **degree** of $h \in \mathcal{H}$ w.r.t. $\mathcal{H}$ as the number of points $x \in \mathcal{X}$ for which there exists $g \in \mathcal{H}$ such that $g$ disagree with $h$ **only** on $x$.

- The **density** of $\mathcal{H}$ is the average degree of a hypothesis in $\mathcal{H}$.
The density function

- We define the **degree** of $h \in \mathcal{H}$ w.r.t. $\mathcal{H}$ as the number of points $x \in \mathcal{X}$ for which there exists $g \in \mathcal{H}$ such that $g$ disagree with $h$ **only on** $x$.

- The **density** of $\mathcal{H}$ is the average degree of a hypothesis in $\mathcal{H}$.

- The **density function** of $\mathcal{H}$ is

$$
\mu_{\mathcal{H}}(m) = \max \{ \text{density}(\mathcal{F}|_S) \mid |S| = m, \mathcal{F} \subset \mathcal{H} \text{ is finite} \} 
$$
The density function

- We define the **degree** of \( h \in \mathcal{H} \) w.r.t. \( \mathcal{H} \) as the number of points \( x \in \mathcal{X} \) for which there exists \( g \in \mathcal{H} \) such that \( g \) disagree with \( h \) only on \( x \).
- The **density** of \( \mathcal{H} \) is the average degree of a hypothesis in \( \mathcal{H} \).
- The **density function** of \( \mathcal{H} \) is

\[
\mu_{\mathcal{H}}(m) = \max\{\text{density}(\mathcal{F}|S) \mid |S| = m, \mathcal{F} \subset \mathcal{H} \text{ is finite}\}
\]

**Theorem**

\[
\frac{1}{2e} \frac{\mu_{\mathcal{H}}(m)}{m} \leq \epsilon_{\mathcal{H}}(m) \leq \frac{\mu_{\mathcal{H}}(m)}{m}
\]
The density function

- We define the **degree** of \( h \in \mathcal{H} \) w.r.t. \( \mathcal{H} \) as the number of points \( x \in \mathcal{X} \) for which there exists \( g \in \mathcal{H} \) such that \( g \) disagree with \( h \) *only on* \( x \).
- The **density** of \( \mathcal{H} \) is the average degree of a hypothesis in \( \mathcal{H} \).
- The **density function** of \( \mathcal{H} \) is

\[
\mu_{\mathcal{H}}(m) = \max\{\text{density}(\mathcal{F}|_S) \mid |S| = m, \mathcal{F} \subset \mathcal{H} \text{ is finite}\}
\]

**Theorem**

\[
\frac{1}{2e} \frac{\mu_{\mathcal{H}}(m)}{m} \leq \epsilon_{\mathcal{H}}(m) \leq \frac{\mu_{\mathcal{H}}(m)}{m}
\]

- \( \Rightarrow \) The sample complexity is characterized by the density function \( \mu_{\mathcal{H}}(m) \).
- \( \mathcal{H} \) is learnable if and only if \( \lim_{m \to \infty} \frac{\mu_{\mathcal{H}}(m)}{m} = 0 \)
Basic questions

- When $\mathcal{H}$ is learnable? When $\lim_{m \to \infty} \frac{\mu_{\mathcal{H}}(m)}{m} = 0$.
- What is the sample complexity of $\mathcal{H}$? $\epsilon_{\mathcal{H}}(m) = \frac{\mu_{\mathcal{H}}(m)}{m}$
- How to learn $\mathcal{H}$ optimally? Use the one inclusion algorithm.
Basic questions

- When $\mathcal{H}$ is learnable? When $\lim_{m \to \infty} \frac{\mu_{\mathcal{H}}(m)}{m} = 0$.
- What is the sample complexity of $\mathcal{H}$? $\epsilon_{\mathcal{H}}(m) = \frac{\mu_{\mathcal{H}}(m)}{m}$
- How to learn $\mathcal{H}$ optimally? Use the one inclusion algorithm.

- End of story?
Basic questions

- When $\mathcal{H}$ is learnable? When $\lim_{m \to \infty} \frac{\mu_{\mathcal{H}}(m)}{m} = 0$.
- What is the sample complexity of $\mathcal{H}$? $\epsilon_{\mathcal{H}}(m) = \frac{\mu_{\mathcal{H}}(m)}{m}$
- How to learn $\mathcal{H}$ optimally? Use the one inclusion algorithm.

- End of story?
- We would like to characterize the growth of $\mu_{\mathcal{H}}(m)$ by a single number (a la the VC dimension).
The complexity of ERM algorithms is analysed using the growth function:

$$\pi_H(m) = \max\{|H|_S : |S| = m\}$$

However, the sample complexity is governed by the density function

$$\mu_H(m) = \max\{\text{density}(F|S) : |S| = m, F \subset H\}$$

Instead of analyse growth, we should analyse density!

Amit Daniely (Hebrew U)
The moral implication: density instead of growth

- The complexity of ERM algorithms is analysed using the growth function:
  \[ \pi_H(m) = \max\{|H|_S : |S| = m\} \]

- However, the sample complexity is governed by the density function
  \[ \mu_H(m) = \max\{\text{density}(F|S) : |S| = m, F \subset H\} \]
The moral implication: density instead of growth

- The complexity of ERM algorithms is analysed using the growth function:
  \[ \pi_H(m) = \max\{|H|_S : |S| = m\} \]

- However, the sample complexity is governed by the density function
  \[ \mu_H(m) = \max\{\text{density}(\mathcal{F}|_S) : |S| = m, \mathcal{F} \subset H\} \]

- Instead of analyse growth, we should analyse density!
For, $\mathcal{H} \subset \{0, 1\}^X$, there are two ways to define the VC dimension:

$$\text{VC}(\mathcal{H}) = \max\{m \mid \pi_{\mathcal{H}}(m) = 2^m\}$$

$$\text{VC}(\mathcal{H}) = \max\{m \mid \mu_{\mathcal{H}}(m) = m\}$$

No longer equivalent if $|Y| > 2$!

More natural to adapt the second!
For, $\mathcal{H} \subset \{0, 1\}^X$, there are two ways to define the VC dimension:

$$VC(\mathcal{H}) = \max\{m \mid \pi_{\mathcal{H}}(m) = 2^m\}$$

$$VC(\mathcal{H}) = \max\{m \mid \mu_{\mathcal{H}}(m) = m\}$$

No longer equivalent if $|\mathcal{Y}| > 2$!

More natural to adapt the second!
A new dimension and the density function

**Definition**

The **dimension** of $\mathcal{H}$ is $\text{Dim}(\mathcal{H}) = \max\{m \mid \mu_\mathcal{H}(m) = m\}$.
A new dimension and the density function

**Definition**

The **dimension** of $\mathcal{H}$ is $\text{Dim}(\mathcal{H}) = \max \{ m \mid \mu_{\mathcal{H}}(m) = m \}$

- Consider the case $|\mathcal{Y}| = 2$.

**Theorem (HLW, 94)**

$$\text{VC}(\mathcal{H}) \leq \mu_{\mathcal{H}}(m) \leq 2\text{VC}(\mathcal{H})$$
A new dimension and the density function

**Definition**

The **dimension** of $\mathcal{H}$ is $\text{Dim}(\mathcal{H}) = \max\{m \mid \mu_\mathcal{H}(m) = m\}$

- Consider the case $|\mathcal{Y}| = 2$.

**Theorem (HLW, 94)**

$$\text{VC}(\mathcal{H}) \leq \mu_\mathcal{H}(m) \leq 2\text{VC}(\mathcal{H})$$

**Conjecture**

$$\text{Dim}(\mathcal{H}) \leq \mu_\mathcal{H}(m) \leq 2 \cdot \text{Dim}(\mathcal{H})$$
A new dimension and the density function

Definition

The **dimension** of $\mathcal{H}$ is $\text{Dim}(\mathcal{H}) = \max\{m \mid \mu_\mathcal{H}(m) = m\}$

- Consider the case $|\mathcal{Y}| = 2$.

Theorem (HLW, 94)

$$VC(\mathcal{H}) \leq \mu_\mathcal{H}(m) \leq 2VC(\mathcal{H})$$

Conjecture

$$\text{Dim}(\mathcal{H}) \leq \mu_\mathcal{H}(m) \leq 2 \cdot \text{Dim}(\mathcal{H})$$

In particular $\epsilon_\mathcal{H}(m) = \Theta\left(\frac{\text{Dim}(\mathcal{H})}{m}\right)$ and $m_\mathcal{H}(\epsilon) = \tilde{\Theta}\left(\frac{\text{Dim}(\mathcal{H})}{\epsilon}\right)$
ERMs are not necessarily optimal (not even for linear classes).

Optimal learners must be improper.

Basic questions

- When $\mathcal{H}$ is learnable? When $\lim_{m \to \infty} \frac{\mu_{\mathcal{H}}(m)}{m} = 0$.
- What is the sample complexity of $\mathcal{H}$? $\epsilon_{\mathcal{H}}(m) = \frac{\mu_{\mathcal{H}}(m)}{m}$
- How to learn $\mathcal{H}$ optimally? Use the one inclusion algorithm.

Conjecture

$$\text{Dim}(\mathcal{H}) \leq \mu_{\mathcal{H}}(m) \leq 2 \cdot \text{Dim}(\mathcal{H})$$

- What about the agnostic case?