（1）Statistical Learning Framework：依赖传统i.i.d.和固定抽样分布D
（2）Online Learning Framework：不假设数据的来源是什么固定的分布（相比于传统SL Frame），模型可以视为learner和adversary之间的game博弈。流程是：t=1，2，3....回合；adv选择一个实例x发送给learner，learner预测表现yt，adv揭示正确的标签yt'。
学习者目标：尽可能减少mistake
（3）mistake bound model：如果学习者A在任何H中某个目标函数f*一致的例子序列上，都只犯M次错误，则称A以错误bound M学习了假设类别H
（4）littlestone dimension, lit(H)：可以刻画哪些假设类别H是可以在线学习的。H的littlestone维度lit(H)定义——H村在littlestone tree的最大深度d
f*：目标函数
y：正确标签
H：假设类别，Hypothesis Class。
H realizable：假设正确标签y，由H中的某个目标函数f*决定，即y=f*(H)，如果有序列((x,y))满足条件，则可以称为：可被H实现，H realizable

Statistical Learning Theory

Omar Montasser

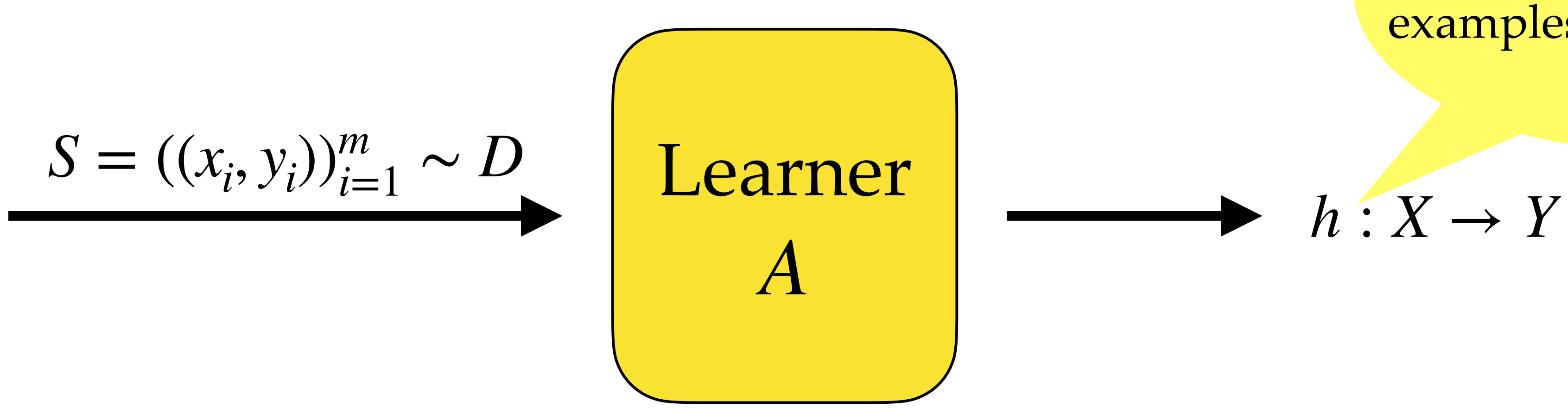
Lecture 9

Online Learning

Statistical Learning Framework

- Unknown source distribution D over $X \times Y$.
- Goal: find a predictor $h : X \to Y$ achieving small *expected error* $L_D(h) = \mathbb{P}_{(x,y)\sim D}\{h(x) \neq y\}$.
- Based on i.i.d. training samples $S = ((x_i, y_i))_{i=1}^m$ drawn from D (each (x_i, y_i) … $S \sim D^m$).

We use h on future examples drawn from D.



$$A : (X \times Y)^\star \to Y^X$$

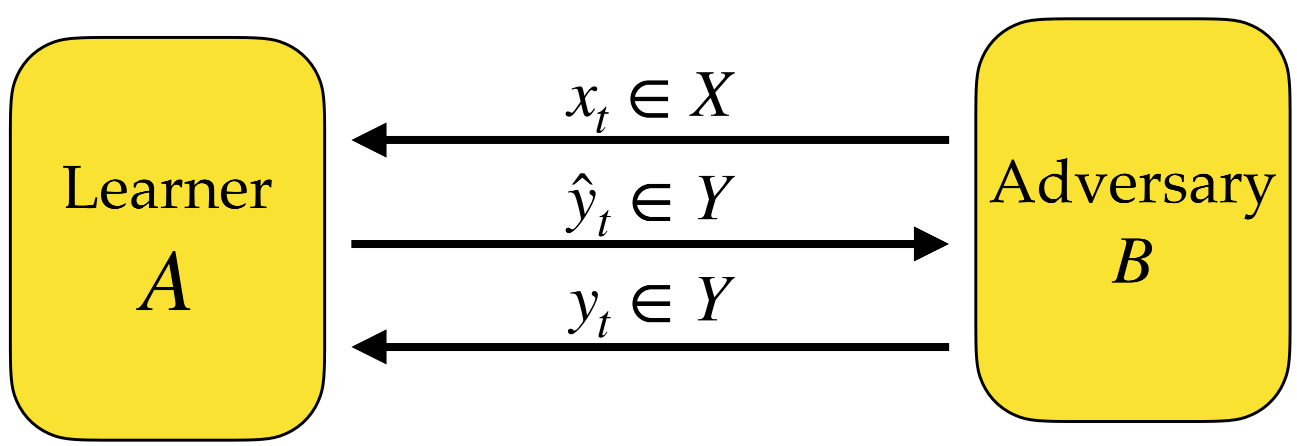
- **Main Assumptions:**
  - We observe i.i.d. training samples from (unknown) distribution D.
  - Future (*unseen*) examples are drawn from the same distribution D.

Online Learning Framework

- We do not assume that data is coming from some (fixed) distribution.
- Can we hope to say anything interesting?
- It can be viewed as a game between a Learner and an Adversary.

For rounds $t = 1,2,\ldots$

- The Adversary chooses an instance $x_t \in X$ and sends it the Learner.
- The Learner predicts a label $\hat{y}_t \in Y$ for x_t.
- The Adversary reveals the correct label $y_t \in Y$ for x_t.



Goal for the Learner is to make as few mistakes as possible.

- The Learner A can be viewed as a map $\cup_{t \geq 1} (X \times Y)^{t-1} \times X \to Y$, or $\cup_{t \geq 1} (X \times Y)^{t-1} \to Y^X$.
  - $h_t = A((x_1, y_1), \ldots, (x_{t-1}, y_{t-1}))$.

在没有任何限制的条件下，对手可以构造一个目标函数f*，使得学习者在N个实例上犯N个错误

No Free Lunch in Online Learning

- Is online learning possible without further restrictions?
- Let's play a game.
  - $X = \{\text{students in class}\}$ and $Y = \{\pm 1\}$.
  - Try to learn my labels …

Claim. For any finite domain $X = \{x_1, \ldots, x_N\}$ and any Learner A, there exists a target function $f^\star : X \to \{\pm 1\}$ such that A makes N mistakes on the sequence $x_1, x_2, \ldots, x_N$.

Proof. Present the instances $x_1, x_2, \ldots, x_N$ to A, and define $f^\star(x_i) = -\hat{y}_i$ where $\hat{y}_i$ is the label predicted by A.

Corollary. For any *infinite* domain X and any Learner A, there exists a target function $f^\star : X \to \{\pm 1\}$ such that A makes a mistake in each round.

Prior Knowledge

- Assume $y = f^\star(x)$ for some $f^\star \in H$.
- $H \subseteq Y^X$ is a "hypothesis class" or a "concept class".
    - Learner knows H but not $f^\star$.
- H represents our "prior knowledge" or "expert knowledge".
- We say sequence $((x_i, y_i))_{t \geq 1}$ is realizable by H.
- What if assumption is wrong?
    - More on this soon …

Mistake Bound Model

Definition. Algorithm A learns a hypothesis class H with a mistake bound M if Learner A makes at most M mistakes on *any sequence of examples* consistent with some $f^\star \in H$.

- We make no assumptions on order of examples $(x_t)_{t \geq 1}$.
- We only assume that the target function $f^\star \in H$.
- Goal is to bound number of mistakes.

Finite Hypothesis Classes

- Are *finite* hypothesis classes learnable in the Mistake Bound model?

CONSISTENT$_H$.
Initialize the version space $V_1 = H$.
For rounds $t = 1,2,\ldots$
- Upon receiving $x_t \in X$, choose a predictor $h_t \in V_t$ and predict $\hat{y}_t = h_t(x_t)$.
- Upon receiving true label y_t, update the version space $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$.

Theorem. On any sequence $((x_t, y_t))_{t\geq 1}$ realizable by H, CONSISTENT$_H$ makes $\leq |H| - 1$ mistakes.

Proof. If $y_t \neq \hat{y}_t$, then h_t will be removed from V_t and thus $|V_{t+1}| \leq |V_t| - 1$. Since true $f^\star$ always remains in the version space $(V_1, V_2, \ldots, V_t, \ldots)$, it holds that for any round t, $|V_t| \geq 1$. Thus, the total number of mistakes is at most $|V_1| - 1$.

Halving

- Can we do better than the $|H| - 1$ mistake bound of CONSISTENT?

HALVING$_H$.
Initialize the version space $V_1 = H$.
For rounds $t = 1,2,\ldots$
- Upon receiving $x_t \in X$, predict $\hat{y}_t = \text{MAJORITY}(h_t(x_t) : h_t \in V_t)$.
- Upon receiving true label y_t, update the version space $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$.

Theorem. On any sequence $((x_t, y_t))_{t\geq 1}$ realizable by H, HALVING$_H$ makes $\leq \log_2(|H|)$ mistakes.

Proof. If $y_t \neq \hat{y}_t$, then at least half of the predictors $h_t \in V_t$ are wrong and will be removed from V_t, thus $|V_{t+1}| \leq |V_t|/2$. Since true $f^\star$ always remains in the version space $(V_1, V_2, \ldots, V_t, \ldots)$, it holds that for any round t, $|V_t| \geq 1$. Thus, the total number of mistakes is at most $\log_2(|V_1|)$.

Mistake Bound Model Properties

Definition. An online learning algorithm A is *conservative* if it only changes its state when it makes a mistake.

Claim. If a hypothesis class H is online learnable with a Mistake Bound M, then it is online learnable by a conservative algorithm with a Mistake Bound M.

Proof Sketch. For any generic online learning algorithm A, we construct a new *conservative* algorithm A' by running algorithm A and rewinding its state when no mistake is made. A' still makes at most M mistakes because A still sees a legal sequence of examples, which are filtered to include only the mistakes.

Thresholds

- $X = [0,1]$ and $H = \{x \mapsto \text{sign}(x - \theta) \mid \theta \in [0,1]\}$.

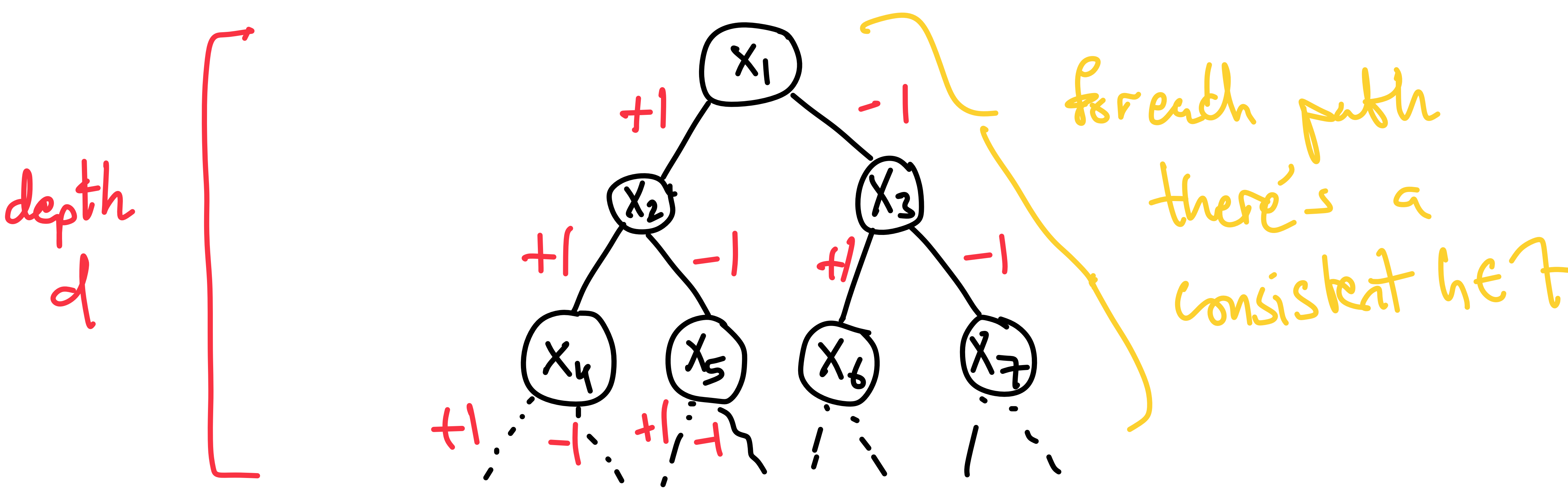
Claim. For any Learner A, there exists a sequence $((x_t, y_t))_{t \geq 1}$ that is realizable by H, on which A makes a mistake on every round.

Proof.
- Start with $l_1 = 0, r_1 = 1$.
- For rounds $t = 1,2,\ldots$
  - Present $x_t = l_t + (r_t - l_t)/2$.
  - If A predicts $\hat{y}_t = +1$, set $y_t = -1$ and update $l_{t+1} = x_t$.
  - If A predicts $\hat{y}_t = -1$, set $y_t = +1$ and update $r_{t+1} = x_t$.
- Observe that for all rounds t, any threshold $\theta \in (l_{t+1}, r_{t+1})$ is consistent with $((x_t, y_t))_{t'=1}^{t}$.

- We *can not* learn Thresholds in the Mistake Bound model!
- This implies that we *can not* learn halfspaces (linear predictors) in higher dimensions!
    - $H = \{x \mapsto \mathrm{sign}(\langle w, x \rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$.
- We can learn *finite* classes H with a Mistake Bound of at most $\log_2(|H|)$.
- Are there examples of *infinite* classes that are online learnable?
- Can we have a characterization of which classes H are online learnable?
- How can we learn optimally in the Mistake Bound model?

Littlestone Trees and Dimension

Definition (Littleton trees). A Littlestone tree of depth d is a complete binary tree whose internal nodes are labeled by instances from X, and whose two edges connecting a node to its children are labeled with +1 and −1 such that every finite path emanating from the root is consistent with some concept in H. That is, a Littlestone tree is a collection $\{x_u : 0 \leq k < d, u \in \{\pm 1\}\} \subseteq X$ such that for every $y \in \{\pm 1\}^d$, there exists $h \in H$ such that $h(x_{y_{1:k}}) = y_{k+1}$ for $0 \leq k < d$.



Littlestone Trees and Dimension

Definition (Littleton trees). A Littlestone tree for H of depth d is a complete binary tree whose internal nodes are labeled by instances from X, and whose two edges connecting a node to its children are labeled with +1 and −1 such that every finite path emanating from the root is consistent with some concept in H. That is, a Littlestone tree is a collection $\{x_u : 0 \leq k < d, u \in \{\pm 1\}\} \subseteq X$ such that for every $y \in \{\pm 1\}^d$, there exists $h \in H$ such that $h(x_{y_{1:k}}) = y_{k+1}$ for $0 \leq k < d$.

Definition (Littleton Dimension). The Littlestone dimension of H, denoted $\mathrm{lit}(H)$, is defined as the largest integer d such that there exists a Littlestone tree for H of depth d.

Characterizing Online Learnability

Theorem. For any class H and any Learner A, the Mistake Bound of A for learning H is $\geq \mathrm{lit}(H)$. [*The Littlestone dimension of H*]

Theorem. For any class H, there exists a Learner A that learns H with a Mistake Bound of $\leq \mathrm{lit}(H)$. [*The Littlestone dimension of H*]

Corollary. A class H is learnable in the Mistake Bound model if and only if the Littlestone dimension of H, $\mathrm{lit}(H)$, is finite.

Lower bound proof

Theorem. For any class H and any Learner A, the Mistake Bound of A for learning H is $\geq \text{lit}(H)$. [*The Littlestone dimension of H*]

Proof.

- Let $T = \text{lit}(H)$ and consider a Littlestone tree for H of depth T.
- Start with x_1 being root of the tree.
- For $1 \leq t \leq T$:
    - Present the root x_t of current subtree to learner A.
    - If A predicts $\hat{y}_t$, recurse to opposite subtree which labels x_t with $-y_t$.

Note that by definition of Littlestone tree, each possible path is realizable by H.

Upper bound proof

Theorem. For any class H, there exists a Learner A that learns H with a Mistake Bound of $\leq \mathrm{lit}(H)$. [*The Littlestone dimension of H*]

Standard Optimal Algorithm (SOA).
Initialize the version space $V_1 = H$.
For rounds $t = 1,2,\ldots$

- Receive $x_t \in X$.
- For $r \in \{\pm 1\}$, let $V_t^{(r)} = \{h \in V_t : h(x_t) = r\}$.
- Predict $\hat{y}_t = \arg\max_{r \in \{\pm 1\}} \mathrm{lit}\left(V_t^{(r)}\right)$. [*i.e., predict the label that maximizes the Littlestone dimension.*]
- Upon receiving true label y_t, update the version space $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$.

Proof. If $y_t \neq \hat{y}_t$, then this implies that $\mathrm{lit}(V_{t+1}) \leq \mathrm{lit}(V_t) - 1$. Because if $\mathrm{lit}(V_{t+1}) = \mathrm{lit}(V_t)$, then by definition of SOA, $\mathrm{lit}\left(V_t^{(+1)}\right) = \mathrm{lit}\left(V_t^{(-1)}\right) = \mathrm{lit}\left(V_t\right)$. This implies that we can construct a Littletone tree of depth $\mathrm{lit}\left(V_t\right) + 1$ for the class V_t, which contradicts the definition of Littlestone dimension. Thus, the total number of mistakes is at most $\mathrm{lit}(V_1)$.

More on Littlestone classes

- What is the Littlestone dimension of Thresholds?
  - $X = [0,1]$ and $H = \{x \mapsto \text{sign}(x - \theta) \mid \theta \in [0,1]\}$.
  - $\text{lit}(H) = \infty$, as we can construct Littlestone trees of infinite depth!

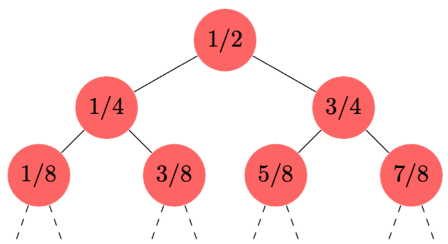


Figure from S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms

  - Remember, $\text{vc}(H) = 1$.
- Similarly for halfspaces (linear predictors) in higher dimensions.
  - $H = \{x \mapsto \text{sign}(\langle w, x \rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$.
  - $\text{lit}(H) = \infty$, but $\text{vc}(H) = d + 1$.
- In general, for any class H, it holds that $\text{vc}(H) \leq \text{lit}(H)$.
  - Why? Construct a Littlestone tree using a VC-shattered set.
- Are there examples of *infinite* classes that are online learnable?
  - $X = [0,1]$ and $H = \{x \mapsto \mathbf{1}[x = \theta] \mid \theta \in [0,1]\}$.
  - We claim that $\text{lit}(H) = 1$. Why?
    - Consider an online learner that always predicts the label 0.

Online-to-Batch Conversions

If a hypothesis class H is learnable in Mistake Bound model, does that imply H is PAC-learnable?

Longest Running Survivor Technique.
Input: a (conservative) online learner A with mistake bound M, training samples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \sim D$.

- Run online learner A on sequence $(x_1, y_1), \ldots, (x_m, y_m)$ until it produces a hypothesis h that survives $\geq (1/\varepsilon)\ln(M/\delta)$ many examples.

Claim. For any class H, let A be an online learning algorithm with a Mistake Bound of $M(H)$. Then, the Longest Running Survivor technique halts after seeing $O\left(\frac{M \log(M/\delta)}{\varepsilon}\right)$ examples, and with probability at least $1 - \delta$, produces a hypothesis with error at most ε.

Analysis. $\mathbb{P}$(any single survived h has error $> \varepsilon$) $\leq (1-\epsilon)^{\ln(\delta/M)/\varepsilon} \leq \delta/M$. Since A is conservative, there are at most M hypotheses. So, we take a union bound.

Summary

- Online Learning: Mistake Bound model.
  - No assumptions on data, except realizability by concept class H.
  - We have a complete characterization which classes H that are learnable in the Mistake Bound model.
  - Namely, classes H with finite Littlestone dimension.
- Next time:
  - Beyond realizability? The notion of minimizing regret.
  - Characterizing what is learnable.
  - Special casses.

Readings

- Chapter 21 of S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms.