Teaching a black-box learner

Sanjoy Dasgupta, Daniel Hsu, Stefanos Poulis, Jerry Zhu

Teaching

Three models of learning:

- The statistical learning model
- Online learning
- Teaching

Teaching

Three models of learning:

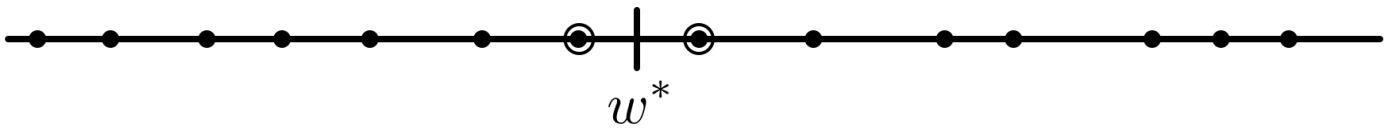
- The statistical learning model
- Online learning
- Teaching

Teacher	Learner
Human	Human
Human	Machine
Machine	Human
Machine	Machine

Minimum teaching sets

Teacher chooses informative examples [Kearns-Goldman, Shinohara-Miyano]:

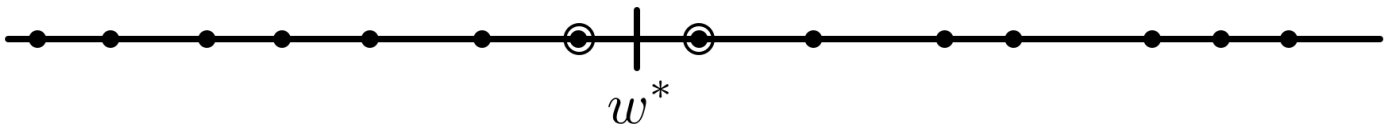
- Finite instance space $\mathcal{X}$
- Learner is using finite concept class C
- Target concept $c^* \in C$
- Teaching set: a set of labeled examples that uniquely identifies c^* in C
- What is the smallest teaching set?



Minimum teaching sets

Teacher chooses informative examples [Kearns-Goldman, Shinohara-Miyano]:

- Finite instance space $\mathcal{X}$
- Learner is using finite concept class C
- Target concept $c^* \in C$
- Teaching set: a set of labeled examples that uniquely identifies c^* in C
- What is the smallest teaching set?



Problem: Teacher needs to know learner's concept class

Teaching a black-box learner

Setting: Learner is using some concept class C (say with VC dimension d, teaching set size t) but teacher has no idea what it is.

Teaching a black-box learner

Setting: Learner is using some concept class C (say with VC dimension d, teaching set size t) but teacher has no idea what it is.

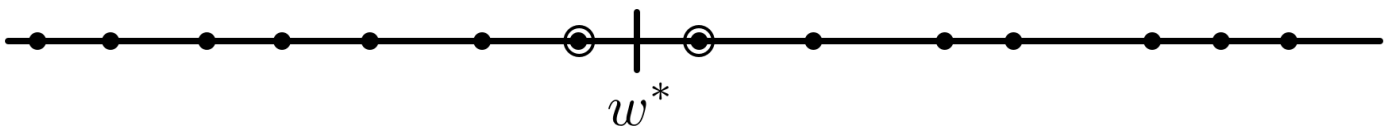
Without interaction: If teaching examples are supplied in advance, can do no better in general than providing all of $\mathcal{X}$.

Teaching a black-box learner

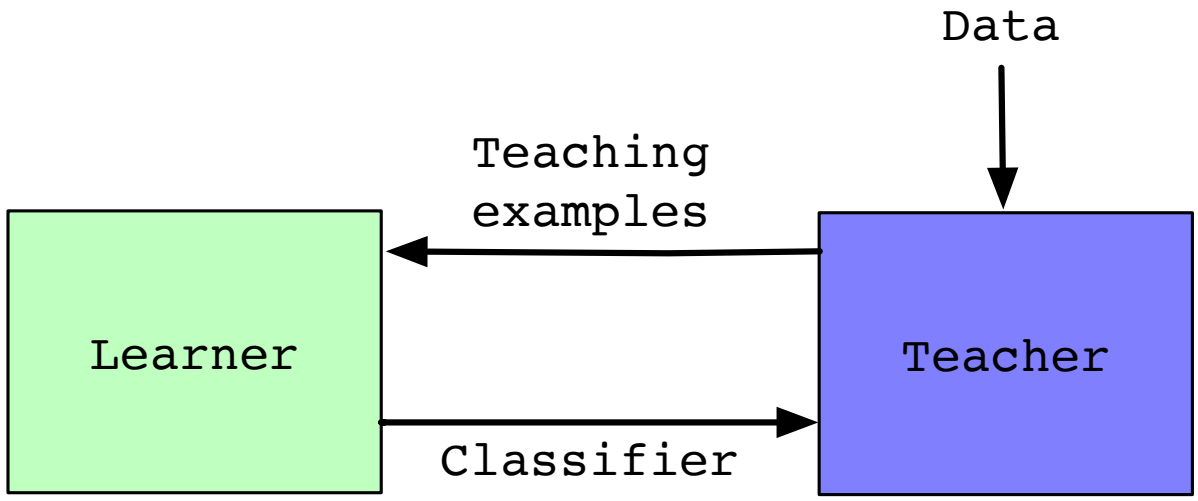
Setting: Learner is using some concept class C (say with VC dimension d, teaching set size t) but teacher has no idea what it is.

Without interaction: If teaching examples are supplied in advance, can do no better in general than providing all of $\mathcal{X}$.

Construction: data in $\mathbb{R}^k$, learner's hypothesis class consists of thresholds along one of the k dimensions:



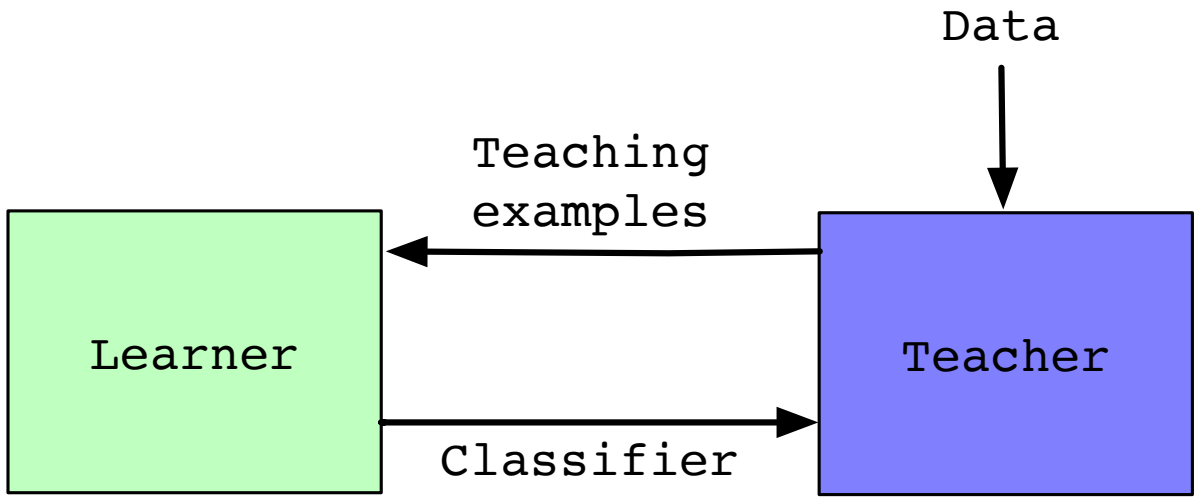
Teaching with interaction



Teaching occurs in rounds:

- The teacher gets to *probe* learner's current concept before choosing which example to provide next.

Teaching with interaction



Teaching occurs in rounds:

- The teacher gets to *probe* learner's current concept before choosing which example to provide next.

Positive result: Efficiently find teaching set of size $O(td \log^2 |\mathcal{X}|)$.

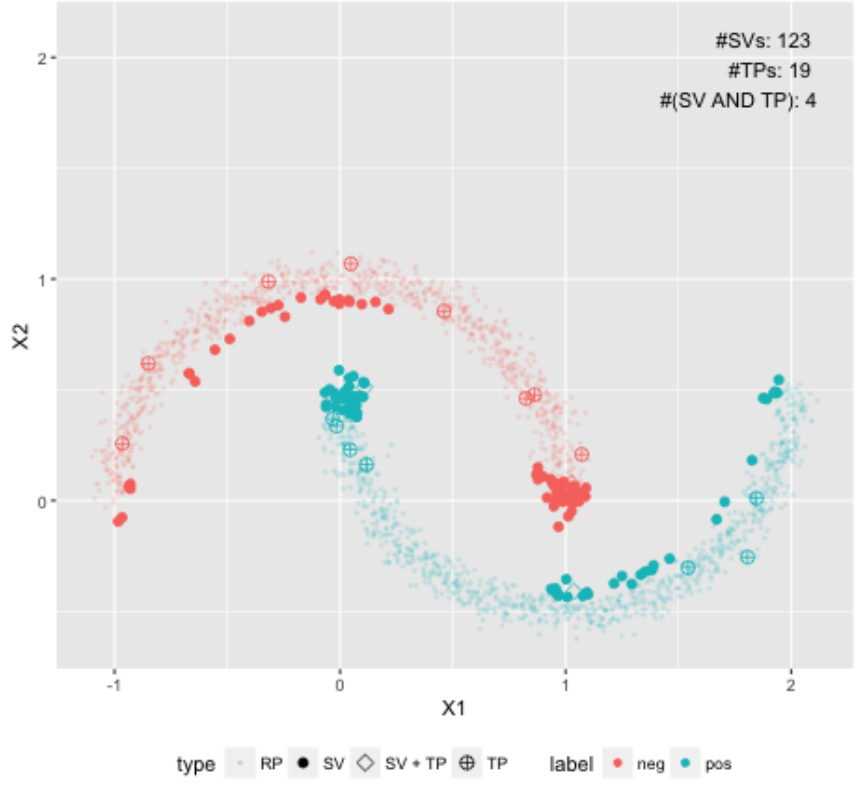
Teaching algorithm

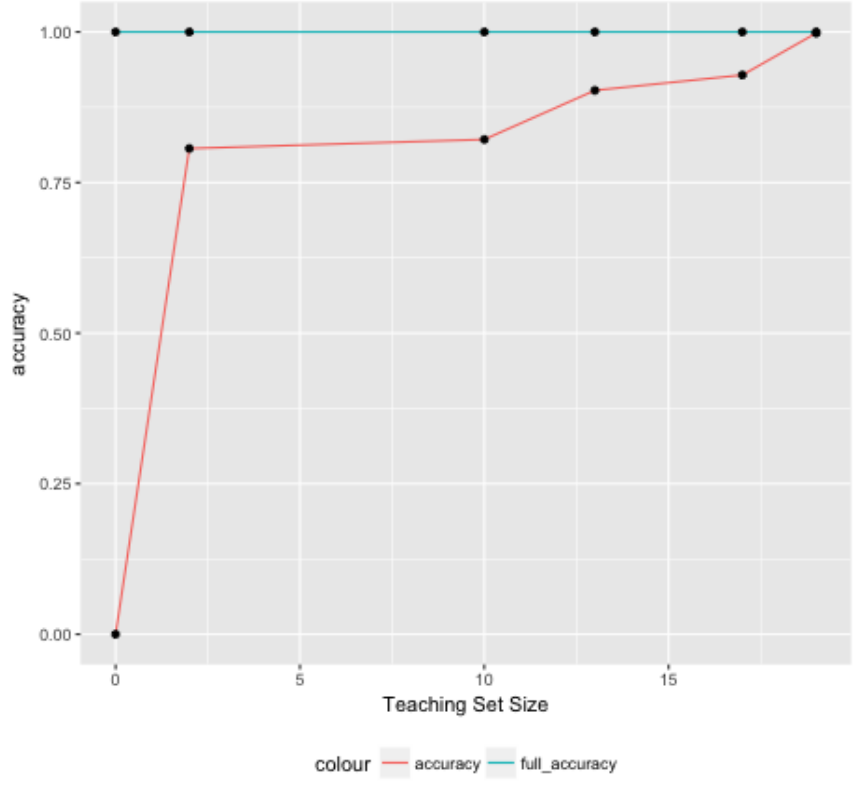
1. Let $S = \emptyset$ (teaching set)

2. For each $x \in \mathcal{X}$:
   - Initialize weight $w(x) = 1/m$
   - Draw T_x from an exponential distribution, rate $\ln(N/\delta)$

3. Repeat until done:
   - Learner provides $h : \mathcal{X} \to \{0, 1\}$ as a black box
   - Let $\Delta(h) = \{x \in \mathcal{X} : h(x) \neq h^*(x)\}$
   - If $\Delta(h) = \emptyset$: halt and accept h
   - While $w(\Delta(h)) < 1$:
     - Double each $w(x)$, for $x \in \Delta(h)$
     - If this causes some $w(x)$ to exceed T_x for the first time, add x to S and provide as a teaching example

Example


#SVs: 123
#TPs: 19
#(SV AND TP): 4
X2
X1
type RP SV SV + TP TP
label neg pos


accuracy
Teaching Set Size
colour accuracy full_accuracy

Open problem in teaching

Teacher	Learner
Human	Human
Human	Machine
Machine	Human
Machine	Machine

Psychological finding: Human learners treat teaching examples differently from random examples.