How can a Machine Learn: Passive, Active, and Teaching

Xiaojin Zhu

jerryzhu@cs.wisc.edu
Department of Computer Sciences
University of Wisconsin-Madison

NLP&CC 2013

Item

$$x \in \mathbb{R}^d$$

Class Label

$$y \in \{-1, 1\}$$

Learning

- You see a training set $(x_1, y_1), \ldots, (x_n, y_n)$
- You must learn a good function $f : \mathbb{R}^d \mapsto \{-1, 1\}$
- f must predict the label y on test item x, which may not be in the training set
- You need some assumptions

The World

$$p(x, y)$$

Independent and Identically-Distributed

$$(x_1, y_1), \ldots, (x_n, y_n), (x, y) \overset{iid}{\sim} p(x, y)$$

Example: Noiseless 1D Threshold Classifier

$$
\begin{aligned}
p(x) &= \text{uniform}[0,1] \\
p(y=1 \mid x) &= \begin{cases} 0, & x < \theta \\ 1, & x \geq \theta \end{cases}
\end{aligned}
$$

Example: Noisy 1D Threshold Classifier

$$
\begin{aligned}
p(x) &= \text{uniform}[0,1] \\
p(y=1 \mid x) &= \begin{cases} \epsilon, & x < \theta \\ 1-\epsilon, & x \geq \theta \end{cases}
\end{aligned}
$$

Generalization Error

- $R(f) = \mathbb{E}_{(x,y)\sim p(x,y)}(f(x) \neq y)$
- Approximated by test set error

$$\frac{1}{m} \sum_{i=n+1}^{n+m} (f(x_i) \neq y_i)$$

on test set $(x_{n+1}, y_{n+1}) \ldots (x_{n+m}, y_{n+m}) \stackrel{iid}{\sim} p(x,y)$

Zero Generalization Error is a Dream

- Speed limit #1: Bayes error

$$R(\text{Bayes}) = \mathbb{E}_{x \sim p(x)} \left(\frac{1}{2} - \left| p(y = 1 \mid x) - \frac{1}{2} \right| \right)$$

- Bayes classifier

$$\text{sign} \left(p(y = 1 \mid x) - \frac{1}{2} \right)$$

- All learners are no better than the Bayes classifier

Hypothesis Space

$$f \in \mathcal{F} \subset \{g : \mathbb{R}^d \mapsto \{-1, 1\} \text{ measurable}\}$$

Approximation Error

- $\mathcal{F}$ may include the Bayes classifier

$$\text{e.g. } \mathcal{F} = \left\{g(x) = \mathrm{sign}(x \geq \theta') : \theta' \in [0,1]\right\}$$

- ... or not

$$\text{e.g. } \mathcal{F} = \{g(x) = \mathrm{sign}(\sin(\alpha x)) : \alpha > 0\}$$

- Speed limit #2: approximation error

$$\inf_{g \in \mathcal{F}} R(g) - R(\text{Bayes})$$

Estimation Error

- Let $f^* = \arg\inf_{g \in \mathcal{F}} R(g)$. Can we at least learn f^*?
- No. You see a training set $(x_1, y_1), \ldots, (x_n, y_n)$, not $p(x, y)$
- You learn $\hat{f}_n$
- Speed limit #3: Estimation error

$$R(\hat{f}_n) - R(f^*)$$

Estimation Error

- As training set size n increases, estimation error goes down
- But how quickly?

Paradigm 1: Passive Learning

- $(x_1, y_1), \ldots, (x_n, y_n) \stackrel{iid}{\sim} p(x, y)$
- 1D example: $O(\frac{1}{n})$

Paradigm 2: Active Learning

- In iteration t
  1. the learner picks a query x_t
  2. the world (oracle) answers with a label $y_t \sim p(y \mid x_t)$
- Pick x_t to maximally reduce the hypothesis space
- 1D example:

$$O(\frac{1}{2^n})$$

Paradigm 3: Teaching

- A teacher designs the training set
- 1D example:

$$\begin{aligned} x_1 &= \theta - \epsilon/2, \quad y_1 = -1 \\ x_2 &= \theta + \epsilon/2, \quad y_2 = 1 \end{aligned}$$

$n = 2$ suffices to drive estimation error to ϵ (teaching dimension [Goldman & Kearns'95])

Teaching as an Optimization problem

$$\min_{\mathcal{D}} \quad \text{loss}(\widehat{f_{\mathcal{D}}}, \theta) + \text{effort}(\mathcal{D})$$

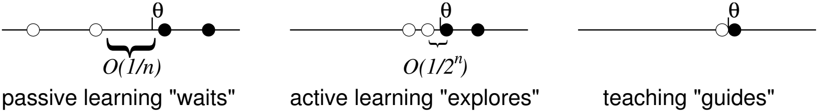
Teaching Bayesian Learners

$$\min_{n,x_1,\ldots,x_n} \; -\log p(\theta^* \mid x_1,\ldots,x_n) + cn$$

if we choose

- $\text{loss}(\widehat{f_{\mathcal{D}}}, \theta^*) = KL\left(\delta_{\theta^*} \| p(\theta \mid \mathcal{D})\right)$
- $\text{effort}(\mathcal{D}) = cn$

Example 1: Teaching a 1D threshold classifier



- $p_0(\theta) = 1$
- $p(y = 1 \mid x, \theta) = 1$ if $x \geq \theta$ and 0 otherwise
- $\text{effort}(\mathcal{D}) = c|\mathcal{D}|$
- For any $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,
  $p(\theta \mid \mathcal{D}) = \text{uniform}\left[\max_{i:y_i=-1}(x_i), \min_{i:y_i=1}(x_i)\right]$
- The optimal teaching problem becomes

$$\min_{n, x_1, y_1, \ldots, x_n, y_n} \; -\log\left(\frac{1}{\min_{i:y_i=1}(x_i) - \max_{i:y_i=-1}(x_i)}\right) + cn.$$

- One solution: $\mathcal{D} = \{(\theta^* - \epsilon/2, -1), (\theta^* + \epsilon/2, 1)\}$ as $\epsilon \to 0$ with
  $TI = \log(\epsilon) + 2c \to -\infty$

Example 2: Learner with poor perception

- Same as Example 1 but the learner cannot distinguish similar items
- Encode this in $\mathrm{effort}()$

$$\mathrm{effort}(\mathcal{D}) = \frac{c}{\min_{x_i, x_j \in \mathcal{D}} |x_i - x_j|}$$

- With $\mathcal{D} = \{(\theta^* - \epsilon/2, -1), (\theta^* + \epsilon/2, 1)\}$, $TI = \log(\epsilon) + c/\epsilon$ with minimum at $\epsilon = c$.
- $\mathcal{D} = \{(\theta^* - c/2, -1), (\theta^* + c/2, 1)\}$.

Example 3: Teaching to pick a model out of two

- $\Theta = \{\theta_A = N(-\frac{1}{4}, \frac{1}{2}), \theta_B = N(\frac{1}{4}, \frac{1}{2})\}$, $p_0(\theta_A) = p_0(\theta_B) = \frac{1}{2}$. $\theta^* = \theta_A$.
- Let $\mathcal{D} = \{x_1, \ldots, x_n\}$. $\mathrm{loss}(\mathcal{D}) = \log\left(1 + \prod_{i=1}^n \exp(x_i)\right)$ minimized by $x_i \to -\infty$.
- But suppose box constraints $x_i \in [-d, d]$:

$$\min_{n, x_1, \ldots, x_n} \log\left(1 + \prod_{i=1}^{n} \exp(x_i)\right) + cn + \sum_{i=1}^{n} \mathbb{I}(|x_i| \leq d)$$

- Solution: all $x_i = -d$, $n = \max\left(0, \left[\frac{1}{d}\log\left(\frac{d}{c} - 1\right)\right]\right)$.
- Note $n = 0$ for certain combinations of c, d (e.g., when $c \geq d$): the effort of teaching outweighs the benefit. The teacher may choose to not teach at all and maintain the status quo (prior p_0) of the learner!

Teaching Dimension is a Special Case

- Given concept class $C = \{c\}$, define $P(y = 1 \mid x, \theta_c) = [c(x) = +]$ and $P(x)$ uniform.
- The world has $\theta^* = \theta_{c^*}$
- The learner has $\Theta = \{\theta_c \mid c \in C\}$, $p_0(\theta) = \frac{1}{|C|}$.
- $P(\theta_c \mid \mathcal{D}) = \frac{1}{|c \in C \text{ consistent with } \mathcal{D}|}$ or 0.
- Teaching dimension [Goldman & Kearns'95] $TD(c^*)$ is the minimum cardinality of $\mathcal{D}$ that uniquely identifies the target concept:

$$\min_{\mathcal{D}} -\log P(\theta_{c^*} \mid \mathcal{D}) + \gamma|\mathcal{D}|$$

  where $\gamma \leq \frac{1}{|C|}$.
- The solution $\mathcal{D}$ is a minimum teaching set for c^*, and $|\mathcal{D}| = TD(c^*)$.

Teaching Bayesian Learners in the Exponential Family

- So far, we solved the examples by inspection.
- Exponential family $p(x \mid \theta) = h(x) \exp\left(\theta^\top T(x) - A(\theta)\right)$
  - $T(x) \in \mathbb{R}^D$ sufficient statistics of x
  - $\theta \in \mathbb{R}^D$ natural parameter
  - $A(\theta)$ log partition function
  - $h(x)$ base measure
- For $\mathcal{D} = \{x_1, \ldots, x_n\}$ the likelihood is

$$p(\mathcal{D} \mid \theta) = \prod_{i=1}^{n} h(x_i) \exp\left(\theta^\top \mathbf{s} - A(\theta)\right)$$

  with aggregate sufficient statistics $\mathbf{s} \equiv \sum_{i=1}^{n} T(x_i)$
- Two-step algorithm: finding aggregate sufficient statistics + unpacking

Step 1: Aggregate Sufficient Statistics from Conjugacy

- The conjugate prior has natural parameters $(\lambda_1, \lambda_2) \in \mathbb{R}^D \times \mathbb{R}$:

$$p(\theta \mid \lambda_1, \lambda_2) = h_0(\theta) \exp\left(\lambda_1^\top \theta - \lambda_2 A(\theta) - A_0(\lambda_1, \lambda_2)\right)$$

- The posterior $p(\theta \mid \mathcal{D}, \lambda_1, \lambda_2) =$

$$h_0(\theta) \exp\left((\lambda_1 + \mathbf{s})^\top \theta - (\lambda_2 + n) A(\theta) - A_0(\lambda_1 + \mathbf{s}, \lambda_2 + n)\right)$$

- $\mathcal{D}$ enters the posterior only via s and n
- Optimal teaching problem

$$\min_{n,\mathbf{s}} \quad -{\theta^*}^\top (\lambda_1 + \mathbf{s}) + A(\theta^*)(\lambda_2 + n) + A_0(\lambda_1 + \mathbf{s}, \lambda_2 + n) + \text{effort}(n, \mathbf{s})$$

- Convex relaxation: $n \in \mathbb{R}$ and s $\in \mathbb{R}^D$ (assuming $\text{effort}(n, \mathbf{s})$ convex)

Step 2: Unpacking

- Cannot teach with the aggregate sufficient statistics
- $n \leftarrow \max(0, [n])$
- Find n teaching examples whose aggregate sufficient statistics is $\mathbf{s}$.
  - exponential distribution $T(x) = x$, $x_1 = \ldots = x_n = \mathbf{s}/n$.
  - Poisson distribution $T(x) = x$ (integers), round $x_1, \ldots, x_n$
  - Gaussian distribution $T(x) = (x, x^2)$, harder. $n = 3, \mathbf{s} = (3, 5)$:
    - ★ $\{x_1 = 0, x_2 = 1, x_3 = 2\}$
    - ★ $\{x_1 = \frac{1}{2}, x_2 = \frac{5+\sqrt{13}}{4}, x_3 = \frac{5-\sqrt{13}}{4}\}$.
- An approximate unpacking algorithm:
  1. initialize $x_i \stackrel{iid}{\sim} p(x \mid \theta^*)$, $i = 1 \ldots n$.
  2. solve $\min_{x_1,\ldots,x_n} \|\mathbf{s} - \sum_{i=1}^n T(x_i)\|^2$ (nonconvex)

Example 4: Teaching the mean of a univariate Gaussian

- The world is $N(x; \mu^*, \sigma^2)$, σ^2 is known to the learner
- $T(x) = x$
- Learner's prior in standard form $\mu \sim N(\mu \mid \mu_0, \sigma_0^2)$
- Optimal aggregate sufficient statistics $s = \frac{\sigma^2}{\sigma_0^2}(\mu^* - \mu_0) + \mu^* n$
  - $\frac{s}{n} \neq \mu^*$: compensating for the learner's initial belief μ_0.
- n is the solution to $n - \frac{1}{2\,\mathrm{effort}'(n)} + \frac{\sigma^2}{\sigma_0^2} = 0$
  - e.g. when $\mathrm{effort}(n) = cn$, $n = \frac{1}{2c} - \frac{\sigma^2}{\sigma_0^2}$
- Not to teach if the learner initially had a "narrow mind": $\sigma_0^2 < 2c\sigma^2$.
- Unpacking s is trivial, e.g. $x_1 = \ldots = x_n = s/n$

Example 5: Teaching a multinomial distribution

- The world multinomial $\pi^* = (\pi_1^*, \ldots, \pi_K^*)$
- The learner Dirichlet prior $p(\pi \mid \beta) = \frac{\Gamma(\sum \beta_k)}{\prod \Gamma(\beta_k)} \prod_{k=1}^K \pi_k^{\beta_k - 1}$.
- Step 1: find aggregate sufficient statistics $\mathbf{s} = (s_1, \ldots, s_K)$

$$\begin{aligned}\min_{\mathbf{s}} \quad & -\log \Gamma\left(\sum_{k=1}^K (\beta_k + s_k)\right) + \sum_{k=1}^K \log \Gamma(\beta_k + s_k) \\ & - \sum_{k=1}^K (\beta_k + s_k - 1) \log \pi_k^* + \text{effort}(\mathbf{s})\end{aligned}$$

  Relax $\mathbf{s} \in \mathbb{R}_{\geq 0}^K$
- Step 2: unpack $s_k \leftarrow [s_k]$ for $k = 1 \ldots K$.

Examples of Example 5

- world $\pi^* = (\frac{1}{10}, \frac{3}{10}, \frac{6}{10})$
- learner's "wrong" Dirichlet prior $\beta = (6, 3, 1)$
- If effortless $\mathrm{effort}(\mathbf{s}) = 0$,
  - $\mathbf{s} = (317, 965, 1933)$ (fmincon)
  - The MLE from $\mathcal{D}$ is $(0.099, 0.300, 0.601)$, very close to π^*.
  - "brute-force teaching": using big data to overwhelm the learner's prior
- If costly $\mathrm{effort}(\mathbf{s}) = 0.3 \sum_{k=1}^{K} s_k$,
  - $\mathbf{s} = (0, 2, 8)$, $TI = 2.65$.
  - Not $\mathbf{s} = (1, 3, 6)$: the wrong prior. $TI = 4.51$
  - Not $\mathbf{s} = (317, 965, 1933)$, $TI = 956.25$

Example 6: Teaching a multivariate Gaussian

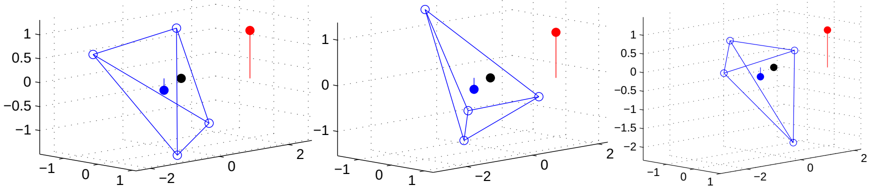
- world: $\mu^* \in \mathbb{R}^D$ and $\Sigma^* \in \mathbb{R}^{D\times D}$
- learner likelihood $N(x \mid \mu, \Sigma)$, Normal-Inverse-Wishart (NIW) prior
- Given $x_1, \ldots, x_n \in \mathbb{R}^D$, the aggregate sufficient statistics are $s = \sum_{i=1}^n x_i, \quad \mathbb{S} = \sum_{i=1}^n x_i x_i^\top$
- Step 1: optimal aggregate sufficient statistics via SDP

$$
\begin{aligned}
\min_{n,s,\mathbb{S}} \quad & \frac{D\log 2}{2}\nu_n + \sum_{i=1}^{D} \log\Gamma\left(\frac{\nu_n + 1 - i}{2}\right) - \frac{\nu_n}{2}\log|\Lambda_n| \\
& -\frac{D}{2}\log\kappa_n + \frac{\nu_n}{2}\log|\Sigma^*| + \frac{1}{2}\mathrm{tr}(\Sigma^{*-1}\Lambda_n) \\
& +\frac{\kappa_n}{2}(\mu^* - \mu_n)^\top\Sigma^{*-1}(\mu^* - \mu_n) + \mathrm{effort}(n, s, \mathbb{S}) \\
\text{s.t.} \quad & \mathbb{S} \succeq 0; \quad \mathbb{S}_{ii} \geq s_i^2/2,\ \forall i.
\end{aligned}
$$

- Step 2: unpack $s, \mathbb{S}$
  - initializing $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu^*, \Sigma^*)$
  - solve $\min \|vec(s, \mathbb{S}) - \sum_{i=1}^n vec(T(x_i))\|^2$

Examples of Example 6

- The target Gaussian is $\mu^* = (\mathbf{0}, \mathbf{0}, \mathbf{0})$ and $\Sigma^* = I$
- The learner's NIW prior
  $\mu_0 = (1, 1, 1), \kappa_0 = 1, \nu_0 = 2 + 10^{-5}, \Lambda_0 = 10^{-5} I.$
- "expensive" $\text{effort}(n, s, \mathbb{S}) = n$
- Optimal $\mathcal{D}$ with $n = 4$, unpacked into a tetrahedron



- $TI(\mathcal{D}) = 1.69$. Four points $\sim N(\mu^*, \Sigma^*)$ have $\text{mean}(TI) = 9.06 \pm 3.34$, $\min(TI) = 1.99$, $\max(TI) = 35.51$ (100,000 trials)