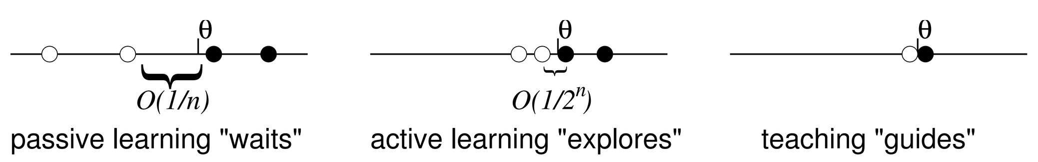
Machine Teaching for Bayesian Learners in the Exponential Family

Xiaojin Zhu

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

Machine Teaching

Machine teaching: finding the best training set.



- **World**: test items $x \overset{iid}{\sim} p(x \mid \theta^*)$.
- **Learner**: hypothesis space Θ
- **Teacher**: knows θ^*, Θ, learning algorithm, teaches by creating a training set $\mathcal{D}$.

Optimal Teaching Key Idea

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

- effort() of the teacher/learner to work with $\mathcal{D}$.
- Not regularized estimation: θ^* given.
- Hard combinatorial optimization
- Objective called Teaching Impedance $TI(\mathcal{D})$

Teaching Bayesian Learners

- Teacher knows learner prior $p_0(\theta)$ and likelihood $p(\mathcal{D} \mid \theta)$, can design non-*iid* $\mathcal{D}$
- $\text{loss}(\widehat{f_{\mathcal{D}}}, \theta^*) = KL\left(\delta_{\theta^*} \| p(\theta \mid \mathcal{D})\right)$
- Teaching is to
$$\min_{\mathcal{D}} -\log p(\theta^* \mid \mathcal{D}) + \text{effort}(\mathcal{D}).$$
- Not MAP estimate! Still hard.

Teaching Bayesian Learners in the Exponential Family

- Exponential family
$p(x \mid \theta) = h(x) \exp\left(\theta^\top T(x) - A(\theta)\right)$
- 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:
  1. finding aggregate sufficient statistics
  2. unpacking

Step 1: Sufficient Statistics

- Conjugate prior $p(\theta \mid \lambda_1, \lambda_2) =$
$$h_0(\theta) \exp\left(\lambda_1^\top \theta - \lambda_2 A(\theta) - A_0(\lambda_1, \lambda_2)\right)$$
- $\mathcal{D}$ enters the posterior only via $\mathbf{s}$ and n:
$$\exp\left((\lambda_1 + \mathbf{s})^\top \theta - (\lambda_2 + n) A(\theta) - A_0(\lambda_1 + \mathbf{s}, \lambda_2 + n)\right)$$
- Optimal teaching problem
$$\min_{n, \mathbf{s}} \ -\theta^{*\top}(\lambda_1 + \mathbf{s}) + A(\theta^*)(\lambda_2 + n) + A_0(\lambda_1 + \mathbf{s}, \lambda_2 + n) + \text{effort}(n, \mathbf{s})$$
- Convex relaxation: $n \in \mathbb{R}$ and $\mathbf{s} \in \mathbb{R}^D$

Step 2: Unpacking

1. Round $n \leftarrow \max(0, [n])$
2. Find n teaching examples whose aggregate sufficient statistics is approximately $\mathbf{s}$:
   - initialize $x_i \overset{iid}{\sim} p(x \mid \theta^*)$, $i = 1 \ldots n$.
   - solve $\min_{x_1, \ldots, x_n} \|\mathbf{s} - \sum_{i=1}^n T(x_i)\|^2$ (nonconvex)

Some unpacking examples:

- Exponential dist $T(x) = x$: $x_i = \frac{\mathbf{s}}{n}$
- Poisson dist $T(x) = x$ (integers): rounding
- Gaussian dist $T(x) = (x, x^2)$, $n = 3, \mathbf{s} = (3, 5)$: $\{x_1 = 0, x_2 = 1, x_3 = 2\}$ or $\{x_1 = \frac{1}{2}, x_2 = \frac{5+\sqrt{13}}{4}, x_3 = \frac{5-\sqrt{13}}{4}\}$

Example 1

Teaching a 1D threshold classifier.

- Learner $p_0(\theta) = 1$, $p(y = 1 \mid x, \theta) = 1_{x \ge \theta}$
- $p(\theta \mid \mathcal{D})$ uniform in $[\max_{i: y_i = -1}(x_i), \min_{i: y_i = 1}(x_i)]$
- $\text{effort}(\mathcal{D}) = c|\mathcal{D}|$
- The optimal teaching problem becomes
$$\min_{n, (x_i, y_i)_{1:n}} \ -\log\left(\frac{1}{\min_{i: y_i = 1}(x_i) - \max_{i: y_i = -1}(x_i)}\right) + cn.$$
- 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 can't tell similar items
$$\text{effort}(\mathcal{D}) = \frac{c}{\min_{x_i, x_j \in \mathcal{D}} |x_i - x_j|}$$

- 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 Gaussian out of two

- $\Theta = \{\theta_A = N(-\frac{1}{4}, \frac{1}{2}), \theta_B = N(\frac{1}{4}, \frac{1}{2})\}$, $\theta^* = \theta_A$, $p_0(\theta_A) = p_0(\theta_B) = \frac{1}{2}$
- $\text{loss}(\mathcal{D}) = \log\left(1 + \prod_{i=1}^n \exp(x_i)\right)$ minimized by $x_i \to -\infty$, weird items.
- Box constraints $x_i \in [-d, d]$:
$$\min_{n, x_{1:n}} \ \log\left(1 + \prod_{i=1}^n \exp(x_i)\right) + cn + \sum_{i=1}^n \mathbb{I}(|x_i| \le d)$$
- Solution: $n = \max\left(0, \left[\frac{1}{d} \log\left(\frac{d}{c} - 1\right)\right]\right)$, $x_{1:n} = -d$
- Note $n = 0$ when $c \ge \frac{d}{2}$: the effort of teaching outweighs the benefit. The teacher will choose not to teach, leaving learner with its prior p_0!

Example 4

Teaching the mean of a univariate Gaussian.

- The world is $N(x; \mu^*, \sigma^2)$
- Learner's prior $p_0(\mu) = N(\mu \mid \mu_0, \sigma_0^2)$, knows σ^2
- $T(x) = x$
- Aggregate sufficient statistics solution
$$s = \frac{\sigma^2}{\sigma_0^2}(\mu^* - \mu_0) + \mu^* n$$
Note $\frac{s}{n} \ne \mu^*$: compensating for the learner's (wrong) prior belief μ_0.
- n is the solution to
$$n - \frac{1}{2\,\text{effort}'(n)} + \frac{\sigma^2}{\sigma_0^2} = 0$$
When $\text{effort}(n) = cn$, $n = \frac{1}{2c} - \frac{\sigma^2}{\sigma_0^2}$
- Unpacking s is trivial, e.g. $x_1 = \ldots = x_n = s/n$
- Teacher will choose not to teach if the learner initially had a "narrow mind": $\sigma_0^2 < 2c\sigma^2$.

Example 5

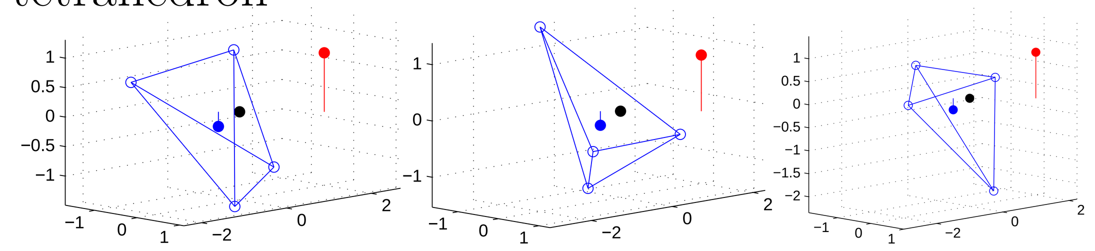
Teaching a multinomial distribution.
$$\min_{\mathbf{s}} \ -\log\Gamma\left(\sum_{k=1}^K (\beta_k + s_k)\right) + \sum_{k=1}^K \log\Gamma(\beta_k + s_k) - \sum_{k=1}^K (\beta_k + s_k - 1)\log \pi_k^* + \text{effort}(\mathbf{s})$$

- Example: world $\pi^* = (\frac{1}{10}, \frac{3}{10}, \frac{6}{10})$
- Learner "wrong" Dirichlet prior $\beta = (6, 3, 1)$
- If $\text{effort}(\mathbf{s}) = 0$, "brute-force teaching" $\mathbf{s} = (317, 965, 1933)$
- If $\text{effort}(\mathbf{s}) = 0.3 \sum_{k=1}^K s_k$,
  - $\mathbf{s} = (0, 2, 8)$, $TI = 2.65$.
  - Not $\mathbf{s} = (1, 3, 6)$, $TI = 4.51$. doesn't correct prior
  - Not $\mathbf{s} = (317, 965, 1933)$, $TI = 956.25$

Example 6

Teaching a multivariate Gaussian.

- World $N(\mu^* = (\mathbf{0}, \mathbf{0}, \mathbf{0}), \Sigma^* = I)$
- Learner Normal-Inverse-Wishart prior $\mu_0 = (1, 1, 1)$, $\kappa_0 = 1$, $\nu_0 = 2 + 10^{-5}$, $\Lambda_0 = 10^{-5} I$.
- "Expensive" $\text{effort}(\mathcal{D}) = n$
- Optimal $\mathcal{D}$ with $n = 4$, unpacked into a tetrahedron



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