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

- World: test items \( x \sim p(x \mid \theta^*) \).
- Learner: hypothesis space \( \Theta \).
- Teacher: knows \( \theta^*, \Theta \), learning algorithm, teaches by creating a training set \( D \).

Optimal Teaching Key Idea

\[
\min_{\theta^*} \text{loss}(f_D, \theta^*) + \text{effort}(D)
\]

- effort() of the teacher/learner to work with \( D \).
- Not regularized estimation: \( \theta^* \) given.
- Hard combinatorial optimization.
- Objective called Teaching Impedance \( T_I(D) \).

Step 1: Sufficient Statistics

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

Step 2: Unpacking

- Round \( n \leftarrow \max(0, \lfloor n \rfloor) \).
- Find \( D \) teaching examples whose aggregate sufficient statistics is approximately \( s \):
  - Initialize \( x_i \sim p(x \mid \theta') \).
  - Solve \( \min_{x_i \in X} \| s - \sum_{i=1}^n x_i \|^2 \) (non-convex).
- Some unpacking examples:
  - Exponential: \( T(x) = x \).
  - Poisson: \( T(x) = x \) (integers): rounding.
  - Gaussian: \( T(x) = (x, x^2) \).
  - Minimize constraints on \( x \) minimize \( \exp(\theta^T x) \). Example 1

Teaching Bayesian Learners in the Exponential Family

- Exponential family \( p(x \mid \theta) = h(x) \exp(\theta^T T(x) - A(\theta)) \).
- For \( D = \{x_1, \ldots, x_n\} \), the likelihood is
  \[
  p(D \mid \theta) = \prod_{i=1}^n h(x_i) \exp(\theta^T x_i - A(\theta))
  \]
  with aggregate sufficient statistics
  \[
  s = \sum_{i=1}^n T(x_i)
  \]
  - Two-step algorithm:
    - Finding aggregate sufficient statistics
    - Unpacking

Example 1

Teaching a 1D threshold classifier.
- Learner \( p(\theta | e) = 1 \) \( p(y = 1 | x, \theta) = 1_{x \geq \theta} \).
- \( p(\theta \mid D) \) uniform in \([\min_{x_i \in X} x_i, \max_{x_i \in X} x_i]\).
- \( \text{effort}(D) = c |D| \).
- The optimal teaching problem becomes
  \[
  \min_{\theta, n, x_i \in x} -\log \left( \min_{n, x_i \in x} \frac{1}{\max_{x_i \in X} x_i - \min_{x_i \in X} x_i} \right) + cn.
  \]
- One solution: \( D = \{ (\theta^* - \epsilon/2, -1), (\theta^* + \epsilon/2, 1) \} \) as \( \epsilon \to 0 \) with \( T_I = \log(c) + 2c \to -\infty \).

Example 2

Learner can’t tell similar items
\[
\text{effort}(D) = \min_{x_i \in x} \log |x_i - x_j|
\]
- With \( D = \{ (\theta^* - \epsilon/2, -1), (\theta^* + \epsilon/2, 1) \} \).
- \( T_I = \log(c) + c \epsilon \) with minimum at \( \epsilon = c \).
- \( D = \{ (\theta^* - c/2, -1), (\theta^* + c/2, 1) \} \).

Example 3

Teaching to pick a Gaussian out of two.
- \( \Theta = \{ \theta_A = N(-\frac{1}{2}, \frac{1}{2}), \theta_B = N(\frac{1}{2}, \frac{1}{2}) \} \).
- \( \text{effort}(\theta_A) = p(\theta_A) = \frac{1}{2} \).
- \( \text{effort}(D) = \log(1 + \frac{1}{\sqrt{n}} \exp(x_i)) \) minimized by \( x_i \to -\infty \), weird items.
- Box constraints \( x_i \in [-d, d] \):
  \[
  \min_{x_i \in [-d, d]} \log \left( 1 + \frac{1}{\sqrt{n}} \exp(x_i) \right) + cn + \frac{1}{n} \sum_{i=1}^n |x_i| \leq d
  \]
- Solution: \( n = \max \left( 0, \left( 2 \log \left( \frac{2}{d} \right) \right) \right) \), \( x_i = -d \).
- Note \( n = 0 \) when \( c > \frac{2}{d} \), the effort of teaching outweighs the benefit. The teacher will choose not to teach, leaving learner with its prior \( p(\theta) \).

Example 4

Teaching the mean of a univariate Gaussian.
- The world is \( N(x, \mu^*, \sigma^2) \).
- Learner’s prior \( p(\mu) = N(\mu | \mu_0, \sigma_0^2) \), knows \( \sigma^2 \).
- \( T(x) = x \).
- Aggregate sufficient statistics solution
  \[
  \begin{align*}
  s &= \frac{\sigma^2}{\sigma_0^2} (\mu^* - \mu_0) + \mu^* n \\
  \text{Note} \quad \frac{n}{\sigma_0^2} &\mu^* \quad \text{compensating for the learner’s (wrong) prior belief.}
  \end{align*}
  \]
- \( n \) is the solution to
  \[
  n = \frac{1}{2} \text{effort}(n) + \frac{\sigma^2}{\sigma_0^2} = 0
  \]
  \[
  \text{When effort}(n) = cn, n = \frac{c}{2} - \frac{\sigma^2}{\sigma_0^2}
  \]
- 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”: \( \sigma_0^2 < 2c^2 \).

Example 5

Teaching a multinomial distribution.
\[
\min_{\beta} \log \left( \sum_{k=1}^K (\beta_k + s_k) \right) + \frac{1}{K} \log |\text{effort}(s)|
\]
- Example: world \( n^* = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}) \).
- Learner “wrong” Dirichlet prior \( \beta = (6, 3, 1) \).
- If \( \text{effort}(s) = 0 \), “brute-force teaching” \( \text{effort}(s) = (317, 965, 1933) \).
- If \( \text{effort}(s) = 0.3 \gamma^N_0 \), \( \gamma = (0.2, 8) \). \( T_I = 2.65 \).
- Not \( s = (1, 3, 0) \), \( T_I = 4.51 \), doesn’t correct prior.
- Not \( s = (317, 965, 1933) \), \( T_I = 956.25 \).

Example 6

Teaching a multivariate Gaussian.
- World \( N(\mu^* = (0, 0, 0), \Sigma = I) \).
- Learner Normal-Inverse-Wishart prior \( p(\mu) = N(1, 1), \alpha_0 = 1, \nu_0 = 2 + 10^{-3}, \lambda_0 = 10^{-3} I \).
- “Expensive” effort \( D = n \).
- Optimal \( D \) with \( n = 4 \), unpacked into a tetrahedron.

Teaching Dimension is a Special Case

- Given concept class \( C = \{ c \} \), define \( P(y = 1 | x, \theta) = c(x) = 1 \) and \( P(x) \) uniform.
- The world has \( \theta^* = \theta_c \).
- The learner has \( \Theta = \{ \theta_c | c \in C \} \).
- \( p(\theta_c \mid D) = \frac{1}{|C|} \) except consistent with \( \theta_c \).
- Teaching dimension (Goldman & Kearns’95) \( T_D(c^*) \) is the minimum cardinality of \( D \) that uniquely identifies the target concept.
\[
\min_{\theta} \min_{p(\theta \mid D)} T(\theta, D)
\]
where \( \gamma \leq \frac{\mu^*}{10} \).
- The solution \( D \) is a minimum teaching set for \( c^* \) and \( |D| = T_D(c^*) \).