Mixture Proportion Estimation for Weakly Supervised Learning

Clay Scott

Electrical and Computer Engineering
University of Michigan
Radioactive sources are characterized by distribution of neutron energies

Organic scintillation detectors: prominent technology for neutron detection
Organic Scintillation Detector

- Detects both neutrons and gamma rays
- Need to classify neutrons and gamma rays
Nuclear Particle Classification

- $X \in \mathbb{R}^d$, $d =$ signal length

- Training data:

  $X_1, \ldots, X_m \overset{iid}{\sim} P_0$ (from gamma ray source, e.g. Na-22)

  $X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} P_1$ (from neutron source, e.g. Cf-252)

- $P_0, P_1 =$ class-conditional distributions; don’t want to model
Reality: No Pure Neutron Sources

- Contamination model for training data:

\[ X_1, \ldots, X_m \overset{iid}{\sim} P_0 \]

\[ X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1 - \pi)P_1 + \pi P_0 \]

- \( \pi \) unknown

- \( P_0, P_1 \) may have overlapping supports (nonseparable problem)

- Nonparametric approach desired

- Problem known as “learning with positive and unlabeled examples” (LPUE)
Measuring Performance

- Classifier:
  \[ f : \mathbb{R}^d \rightarrow \{0, 1\} \]

- False positive/negative rates:
  \[ R_0(f) := P_0(f(X) = 1) \]
  \[ R_1(f) := P_1(f(X) = 0) \]
  \[ \tilde{R}_1(f) := \tilde{P}_1(f(X) = 0) \]

- Estimating false negative rate:
  \[ \tilde{P}_1 = (1 - \pi)P_1 + \pi P_0 \]
  \[ \downarrow \]
  \[ \tilde{R}_1(f) = (1 - \pi)R_1(f) + \pi(1 - R_0(f)) \]
  \[ \downarrow \]
  \[ R_1(f) = \frac{\tilde{R}_1(f) - \pi(1 - R_0(f))}{1 - \pi} \]

- Suffices to estimate \( \pi \)
Mixture Proportion Estimation

- Consider

\[ Z_1, \ldots, Z_m \overset{iid}{\sim} H \]

\[ Z_{m+1}, \ldots, Z_{m+n} \overset{iid}{\sim} F = (1 - \kappa)G + \kappa H \]

- Need consistent estimate of \( \kappa \)

- Note: \( \kappa \) not identifiable in general
Mixture Proportion Estimation

- Given two distributions $F, H$, define

$$\kappa^*(F|H) = \max\{\alpha \in [0, 1] : \exists G' \text{ s.t. } F = (1 - \alpha)G' + \alpha H\}$$

- $\kappa^*$ can be estimated – stay tuned

- When is $\kappa = \kappa^*(F|H)$?
Identifiability Condition

• If

\[ F = (1 - \kappa)G + \kappa H \]

then

\[ \kappa = \kappa^*(F \mid H) \iff \kappa^*(G \mid H) = 0 \]

• Apply to LPUE

\[ X_1, \ldots, X_m \overset{iid}{\sim} P_0 \]

\[ X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1 - \pi)P_1 + \pi P_0 \]

• Need

\[ \kappa^*(P_1 \mid P_0) = 0 \]

In words: Can’t write \( P_1 \) as a (nontrivial) mixture of \( P_0 \) and some other distribution
Classification with Label Noise

- Contaminated training data:
  \[ X_1, \ldots, X_m \sim \tilde{P}_0 = (1 - \pi_0)P_0 + \pi_0 P_1 \]
  \[ X_{m+1}, \ldots, X_{m+n} \sim \tilde{P}_1 = (1 - \pi_1)P_1 + \pi_1 P_0 \]

- \( P_0, P_1 \) unknown
- \( P_0, P_1 \), may have overlapping supports
- \( \pi_0, \pi_1 \) unknown
- Asymmetric label noise: \( \pi_0 \neq \pi_1 \)

- Random label noise, as opposed to adversarial, or feature-dependent
Understanding Label Noise

- Assume $P_0, P_1$ have densities $p_0(x), p_1(x)$
- Then $\tilde{P}_0, \tilde{P}_1$ have densities

$$
\tilde{p}_0(x) = (1 - \pi_0)p_0(x) + \pi_0p_1(x)
$$
$$
\tilde{p}_1(x) = (1 - \pi_1)p_1(x) + \pi_1p_0(x)
$$

- Simple algebra:

$$
\frac{p_1(x)}{p_0(x)} > \gamma \iff \frac{\tilde{p}_1(x)}{\tilde{p}_0(x)} > \lambda,
$$

where

$$
\lambda = \frac{\pi_1 + \gamma(1 - \pi_1)}{1 - \pi_0 + \gamma\pi_0}.
$$
Modified Contamination Model

- Recall contamination model:

\[ X_1, \ldots, X_m \overset{iid}{\sim} \tilde{P}_0 = (1 - \pi_0)P_0 + \pi_0 P_1 \]
\[ X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} \tilde{P}_1 = (1 - \pi_1)P_1 + \pi_1 P_0 \]

- Proposition: If \( \pi_0 + \pi_1 < 1 \) and \( P_0 \neq P_1 \), then

\[ \tilde{P}_0 = (1 - \tilde{\pi}_0)P_0 + \tilde{\pi}_0 \tilde{P}_1 \]
\[ \tilde{P}_1 = (1 - \tilde{\pi}_1)P_1 + \tilde{\pi}_1 \tilde{P}_0 \]

where

\[ \tilde{\pi}_0 = \frac{\pi_0}{1 - \pi_1}, \quad \tilde{\pi}_1 = \frac{\pi_1}{1 - \pi_0} \]
MPE for Label Noise

• Modified contamination model

\[ X_1, \ldots, X_m \sim \tilde{P}_0 = (1 - \tilde{\pi}_0)P_0 + \tilde{\pi}_0 \tilde{P}_1 \]

\[ X_{m+1}, \ldots, X_{m+n} \sim \tilde{P}_1 = (1 - \tilde{\pi}_1)P_1 + \tilde{\pi}_1 \tilde{P}_0 \]

• Need consistent estimates of \( \tilde{\pi}_0, \tilde{\pi}_1 \) → MPE

• Identifiability: Need

\[ \kappa^*(P_0 \mid \tilde{P}_1) = 0 \text{ and } \kappa^*(P_1 \mid \tilde{P}_0) = 0 \]

or equivalently (it can be shown)

\[ \kappa^*(P_0 \mid P_1) = 0 \text{ and } \kappa^*(P_1 \mid P_0) = 0 \]
Identifiability Condition

\[ \kappa^*(P_0 \mid P_1) = 0 \text{ and } \kappa^*(P_1 \mid P_0) = 0 \]
Weakly Supervised Learning Problems

That can be reduced to MPE and its extensions

- Learning with positive and unlabeled examples (JMLR ’10)
- Classification with label noise (COLT ’13)
- Multiclass label noise (AISTATS ’14a)
- Various forms of domain adaptation (AISTATS ’14b)
- Co-training (Electronic J. Stat, ’16)
- Classification with partial labels (arxiv ’16)
- Estimating mixed membership models (arxiv ’16)
- Two-sample problem?

Common theme: contamination models: Observations described by

\[ \tilde{P}_j = \sum_i \pi_{ij} P_i \]
Some Related Work


Multiple hypothesis testing: Genovese and Wasserman (2004)
Approaches to Mixture Prop. Est.

- Plug-in
- ROC slope
- Class probability estimation
- Kernel mean embedding
MPE: Density Ratio Formulation

- Key observation: For any $F, H$

\[ \kappa^*(F \mid H) = \inf_{A : H(A) > 0} \frac{F(A)}{H(A)} \]

- Proof: $\kappa^*$ is the largest $\kappa$ such that

\[ G = \frac{F - \kappa H}{1 - \kappa} \]

is a distribution.

- Similarly, if $F$ and $H$ have densities $f$ and $h$, then

\[ \kappa^*(F \mid H) = \text{ess inf}_{x : h(x) > 0} \frac{f(x)}{h(x)} \]

- Universally consistent estimator established by Blanchard et al. (2010)
ROC Method

- Rewrite previous identity as (substituting $A \rightarrow A^c$)

\[
\kappa^*(F \mid H) = \inf_{A : H(A) < 1} \frac{1 - F(A)}{1 - H(A)}
\]

- Slope of ROC at its right endpoint

- Sanderson and Scott (2014), Scott (2015): implementations based on kernel logistic regression
Class Probability Estimation

- Assume joint distribution on $(X, Y)$, $Y = 0, 1$, where

$$X|Y = 1 \sim F$$
$$X|Y = 0 \sim H$$

- Prior / posterior class probabilities

$$\theta := \Pr(Y = 1)$$
$$\eta(x) := \Pr(Y = 1 \mid X = x)$$

- By a simple application of Bayes rule,

$$\eta_{\text{max}} := \sup_x \eta(x) = \frac{1}{1 + \frac{1-\theta}{\theta} \kappa^*(F \mid H)}$$

Kernel Mean Embedding Approach

- Assume $\kappa^*(G \mid H) = 0$

- Consider

\[
x_1, \ldots, x_m \overset{iid}{\sim} H
\]

\[
x_{m+1}, \ldots, x_{m+n} \overset{iid}{\sim} F = (1 - \kappa)G + \kappa H
\]

- Letting $\lambda = \frac{1}{1-\kappa}$, we have

\[
\lambda F + (1 - \lambda)H
\]

Set of valid distributions
Kernel Mean Embedding Approach

- Recall

\[ \kappa^* = \max\{\alpha \in [0, 1] \mid \exists G' \text{ s.t. } F = (1 - \alpha)G' + \alpha H\} \]

- Define

\[ \lambda^* = \sup\{\lambda \geq 1 \mid \lambda F + (1 - \lambda)H \text{ is a valid distribution}\} \]

- Then \( \lambda^* = \frac{1}{1 - \kappa^*} \), and we have

\[
\begin{align*}
G & \quad \lambda = \lambda^* \\
F & \quad \lambda = 1 \\
H & \quad \lambda = 0
\end{align*}
\]

Set of valid distributions
Kernel Mean Embedding

• Let $\mathcal{H}$ denote a reproducing kernel Hilbert space (RKHS) with reproducing kernel $k$.

• The **kernel mean embedding** of a distribution $P$ is

$$
\phi(P) = \mathbb{E}_{X \sim P} \ k(\cdot, X) \in \mathcal{H}
$$

• If $x_1, \ldots, x_\ell \sim P$, an estimate of $\phi(P)$ is

$$
\phi(\hat{P}) = \frac{1}{\ell} \sum_{i=1}^{\ell} k(\cdot, x_i)
$$

where $\hat{P}$ is the empirical distribution.

• If $\phi$ is injective, then $\|\phi(P) - \phi(P')\|_\mathcal{H}$ is a notion of distance between $P$ and $P'$.
Distance to Set of Distributions

- Define

\[ C = \{ w \in \mathcal{H} : w = \phi(P), \text{ for some distribution } P \} , \]

\[ \hat{C} = \{ w \in \mathcal{H} : w = \sum_{i=1}^{n+m} \alpha_i \phi(x_i), \text{ for some } \alpha \in \Delta_{n+m} \} . \]

- For each \( \lambda \geq 0 \), define

\[ d(\lambda) = \inf_{w \in C} \| \lambda \phi(F) + (1 - \lambda) \phi(H) - w \|_{\mathcal{H}} \]

\[ \hat{d}(\lambda) = \inf_{w \in \hat{C}} \| \lambda \phi(\hat{F}) + (1 - \lambda) \phi(\hat{H}) - w \|_{\mathcal{H}} . \]

- Clearly \( d(\lambda) = 0 \) for \( \lambda \leq \lambda^* \).

- Ideally \( d(\lambda) > 0 \) for \( \lambda > \lambda^* \).
Properties of Distance Function

Theorem: If

- $k$ is universal, e.g., Gaussian kernel
- there exists a compact set $A$ such that $A \subseteq \text{supp}(H) \setminus \text{supp}(G)$ and $H(A) > 0$.

then

$$d(\lambda) > 0$$

for all $\lambda > \lambda^*$. 

Other useful properties:

- $d, \hat{d}$ are nondecreasing, convex
- Computing $\hat{d}(\lambda)$ entails solving a quadratic program
Thresholding Estimators

- For $\tau > 0$, define
  \[ \hat{\lambda}_\tau = \inf \{ \lambda \mid \hat{d}(\lambda) \geq \tau \} \]

- Since $d$ is convex and nondecreasing, we can also consider
  \[ \hat{\lambda}_{\nabla}^{\text{grad}} = \inf \{ \lambda \mid \nabla \hat{d}(\lambda) \geq \nu \} \]

- For appropriately chosen $\tau$ and $\nu$, and under the previous assumptions, both estimators converge to $\lambda^*$ at a rate of
  \[ \frac{1}{\sqrt{\min(m, n)}}. \]
Illustration 1

\[ n = 400, \lambda^* = 1.25 \]

\[ n = 1600, \lambda^* = 1.25 \]
Illustration 2

$n = 400, \lambda^* = 1.75$

$n = 1600, \lambda^* = 1.75$
Experimental results
Multiclass Label Noise

- Training distributions:

\[
\tilde{P}_0 = (1 - \pi_{01} - \pi_{02}) P_0 + \pi_{01} P_1 + \pi_{02} P_2 \\
\tilde{P}_1 = \pi_{10} P_0 + (1 - \pi_{10} - \pi_{12}) P_1 + \pi_{12} P_2 \\
\tilde{P}_2 = \pi_{20} P_0 + \pi_{21} P_1 + (1 - \pi_{20} - \pi_{21}) P_2
\]
Maximum Mixture Proportions

- Given distributions $F$ and $H_1, \ldots, H_M$, define

$$\kappa^*(F \mid H_1, \ldots, H_M) = \max \left\{ \sum_{i=1}^M \nu_i \left| \nu_i \geq 0, \sum_{i=1}^M \nu_i \leq 1, \text{ and } \exists \text{ a distribution } G \text{ s.t.} \right. \right.$$ 

$$F = \left( 1 - \sum_{i=1}^M \nu_i \right) G + \sum_{i=1}^M \nu_i H_i \right\}.$$ 

- Arises in other multiclass settings, e.g., topic modelling, learning with partial labels

- A universally consistent estimator $\hat{\kappa}(\hat{F} \mid \hat{H}_1, \ldots, \hat{H}_M)$ exists (AISTATS 2014), but practical estimators are needed.
Classification with Unknown Class Skew

- Binary classification training data
  \[ X_1, \ldots, X_m \overset{iid}{\sim} P_0 \]
  \[ X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} P_1 \]

- Test data:
  \[ Z_1, \ldots, Z_k \overset{iid}{\sim} P_{\text{test}} = \pi P_0 + (1 - \pi) P_1 \]

- \( \pi \) unknown

- \( \pi \) needs to be known for several performance measures (probability of error, precision)

- MPE: If \( \kappa^*(P_1, P_0) = 0 \) then \( \pi = \kappa^*(P_{\text{test}}, P_0) \)
  \[ \hat{\pi} = \hat{\kappa}(\{X_i\}_{i=1}^m, \{Z_i\}_{i=1}^k) \]
Classification with Reject Option

- Binary classification training data
  \[ X_1, \ldots, X_m \stackrel{iid}{\sim} P_0 \]
  \[ X_{m+1}, \ldots, X_{m+n} \stackrel{iid}{\sim} P_1 \]

- Test data:
  \[ Z_1, \ldots, Z_k \stackrel{iid}{\sim} P_{test} = \pi_0 P_0 + \pi_1 P_1 + (1 - \pi_0 - \pi_1) P_2 \]

- \( P_2 \) = distribution of everything else (reject)

- \( \pi_0, \pi_1 \) unknown

- Use MPE (twice) to estimate \( \pi_0, \pi_1 \)
  \[ \implies \text{estimate probability of class 2 error} \]
  \[ \implies \text{design a classifier} \]
Collaborators

- Gilles Blanchard
- Gregory Handy, Tyler Sanderson
- Marek Flaska, Sara Pozzi
- Harish Ramaswamy, Ambuj Tewari
Conclusion

Mixture proportion estimation can be used to solve

- Learning with positive and unlabeled examples
- Classification with label noise
- Multiclass label noise
- Classification with unknown class skew
- Classification with reject option
- Co-training
- Classification with partial labels
- Mixed membership models
- ???