Decontamination of Mutually Contaminated Models

Gilles Blanchard and Clay Scott

Universität Potsdam

MICHIGAN

Standard Multiclass Classification

- P_i = (class-conditional) distribution of X given $Y = i$, on space $\mathcal{X}$
- Training data

$$X_1^i, \ldots, X_{n_i}^i \overset{i.i.d.}{\sim} P_i, \qquad i = 1, \ldots, L$$

- **Goal:** Estimate a good classifier $f : \mathcal{X} \to \{1, \ldots, L\}$
- Various performance measures: Classification accuracy, cost-sensitive risk, minmax, Neyman-Pearson, etc.

Label Noise: Mutual Contamination Model

- Contaminated training data

$$X_1^i, \ldots, X_{n_i}^i \overset{i.i.d.}{\sim} \tilde{P}_i = \sum_{j=1}^{L} \pi_{ij} P_i, \qquad i = 1, \ldots, L$$

- **Goals:**
  - Estimate a good classifier $f : \mathcal{X} \to \{1, \ldots, L\}$
  - Estimate π_{ij}
  - Estimate P_i
- **Assumptions:**
  - π_{ij} unknown, possibly asymmetric
  - P_i unknown, possibly overlapping

Related Label Noise Model

- Assume (X, Y) jointly distributed
- First generate "clean" (but unobserved) training data drawn i.i.d from joint distribution
- Replace each true label Y with corrupted label $\tilde{Y}$ according to

$$\theta_{ij} = \Pr(\tilde{Y} = j \,|\, Y = i, X).$$

- π_{ij} and θ_{ij} related by Bayes rule
- **Note:** Label noise independent of X – does not encompass instance-dependent or adversarial label noise

Motivation

- Nuclear particle classification (pure training data unavailable)
- Crowdsourcing
- Topic modeling (# topics = # documents)
- Learning from partial labels

Related Work

- Previous work on related topics include:
  - Learning from positive and unlabeled data (LPUE) (Denis et al. 05, Liu et al. 03)
  - Co-training (Blum and Mitchell 98)
  - Label noise models and noise-tolerant PAC learning (Angluin and Laird 88, Kearns 93, Aslam and Deactur 96, Cesa-Bianchi et al. 97, Bshouty et al. 98, Kalai and Servedio 03, Stempfel and Ralaivola 09, Jabbari 10)
  - Some negative results (Long and Servido 10, Manwani and Sastry 11)
  - Surrogate losses and label noise (Stempfel and Ralaivola 09, Natarajan et al. 13)

Related Work (2/2)

- Previous theoretical work assumes $L = 2$
- Generally one or more of the following is assumed:
  - P_1, P_2 have non-overlapping support ($\leftrightarrow$ deterministic target concept)
  - symmetric label noise
  - known noise proportions π_{ij}/θ_{ij}
- We do not assume the above here

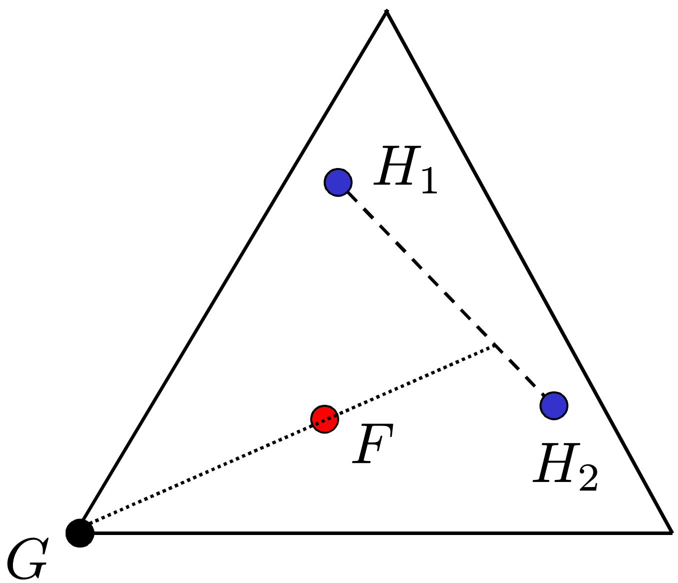
Maximum Mixture Proportions

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

$$\kappa^*(F \mid H_1, \ldots, H_M) = \max \left\{ \sum_{i=1}^{M} \nu_i \;\middle|\; \nu_i \geq 0, \sum_{i=1}^{M} \nu_i \leq 1, \text{ and } \right.$$

$$\exists \text{ a distribution } G \text{ s.t.}$$

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



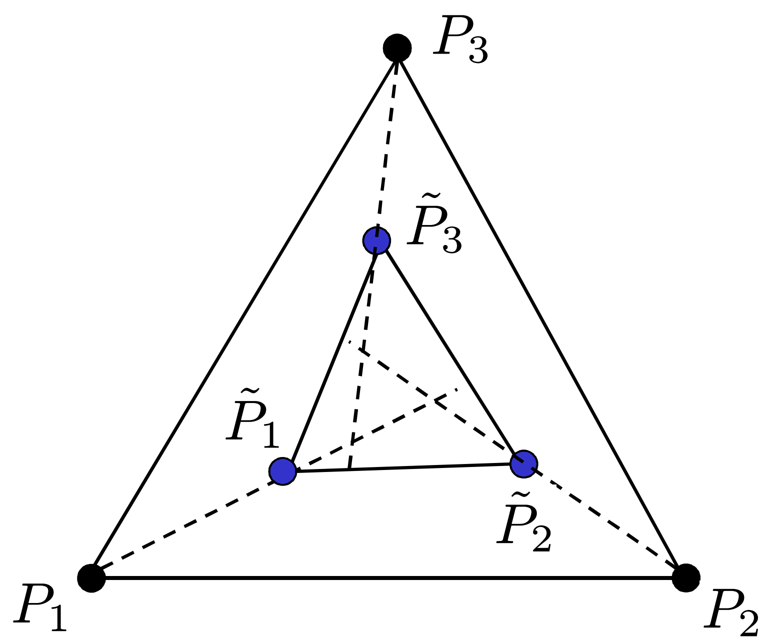
- If G achieving κ^* in unique, it is called the **residue** of F with respect to $H_1, \ldots, H_M$.

- We establish a **universally consistent** estimator $\widehat{\kappa}(\widehat{F} \mid \widehat{H}_1, \ldots, \widehat{H}_M)$. If $\nu_1, \ldots, \nu_M$ achieving κ^* are unique, these are also consistently estimated.

Identifiability

- Write $\widetilde{P} = \Pi P$ where $\Pi = [\pi_{ij}]$
- **Theorem:** If $P_1, \ldots, P_L$ are **jointly irreducible** and Π is **recoverable**, then for each ℓ, P_ℓ is the residue of $\widetilde{P}_\ell$ w.r.t. $\{\widetilde{P}_j, j \neq \ell\}$. Therefore
$$\widetilde{P}_\ell = (1-\kappa_\ell)P_\ell + \sum_{j \neq \ell} \nu_{\ell j} \widetilde{P}_j,$$
where $\kappa_\ell = \kappa^*(\widetilde{P}_\ell \,|\, \{\widetilde{P}_j, j \neq \ell\})$. In addition, $\kappa_\ell < 1$, the $\nu_{\ell j}$ are unique, and
$$(\Pi^{-1})_{\ell k} = -\frac{\nu_{\ell k}}{1-\kappa_\ell}; \qquad (\Pi^{-1})_{\ell \ell} = \frac{1}{1-\kappa_\ell}.$$
- **Conclusion:** κ_ℓ can be estimated consistently by $\widehat{\kappa}(\widehat{\widetilde{P}}_\ell | \{\widehat{\widetilde{P}}_j, j \neq \ell\})$; so can $\nu_{\ell j}$, Π^{-1} and finally Π.

Identifiability – Intuition



Consistent Discrimination

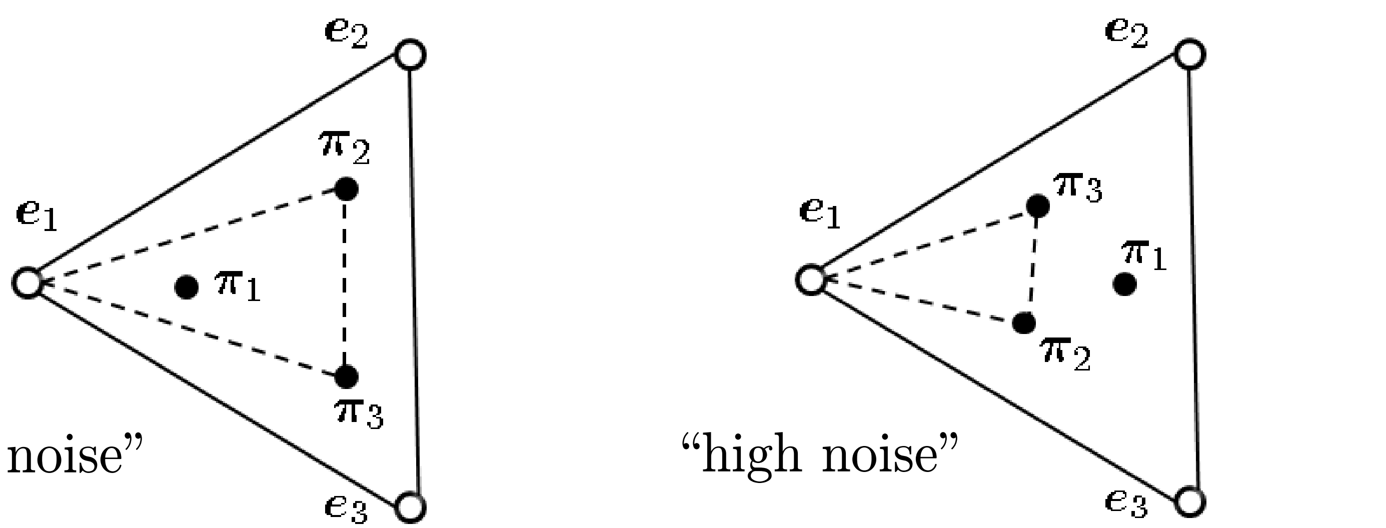
- Consistent estimation of Π/Π^{-1} enables construction of a consistent discrimination rules using standard techniques (e.g., empirical risk minimization over a growing sequence of VC classes)
- See poster/paper for details

Joint Irreducibility

- $P_1, \ldots, P_L$ are **jointly irreducible** if the following equivalent conditions hold:
  - If $\sum_{i=1}^{L} \gamma_i P_i$ is a distribution, then $\gamma_i \geq 0$ for all i.
  - For any $I \subset \{1, \ldots, L\}$; $1 \leq |I| \leq (L-1)$,
    for any distribution $Q \in \mathrm{ConvHull}\{P_i, i \in I\}$,
    for any distribution $Q' \in \mathrm{ConvHull}\{P_i, i \in I^c\}$,
    $\kappa^*(Q \,|\, Q') = 0$.
- If $P_1, \ldots, P_L$ are **discrete** on a finite domain, joint irreducibility is equivalent to the **separability** assumption from topic modeling

Recoverability

- Let $\boldsymbol{\pi}_\ell$ denote the ℓ-th row of $\Pi = [\pi_{ij}]$.
- Denote $\boldsymbol{e}_\ell = (0, \ldots, 0, 1, 0, \ldots, 0)$
- Π is **recoverable** if the following equivalent conditions hold:
  (a) For every ℓ there exists a decomposition $\boldsymbol{\pi}_\ell = \kappa_\ell \boldsymbol{e}_\ell + (1 - \kappa_\ell)\boldsymbol{\pi}'_\ell$ where $\kappa_\ell > 0$ and $\boldsymbol{\pi}'_\ell$ is a convex combination of $\boldsymbol{\pi}_j$ for $j \neq \ell$.
  (b) Π is invertible and Π^{-1} is a matrix with strictly positive diagonal entries and nonpositive off-diagonal entries.
  (c) For each ℓ, the residue of $\boldsymbol{\pi}_\ell$ with respect to $\{\boldsymbol{\pi}_j, j \neq \ell\}$ is $\boldsymbol{e}_\ell$.



"low noise" "high noise"

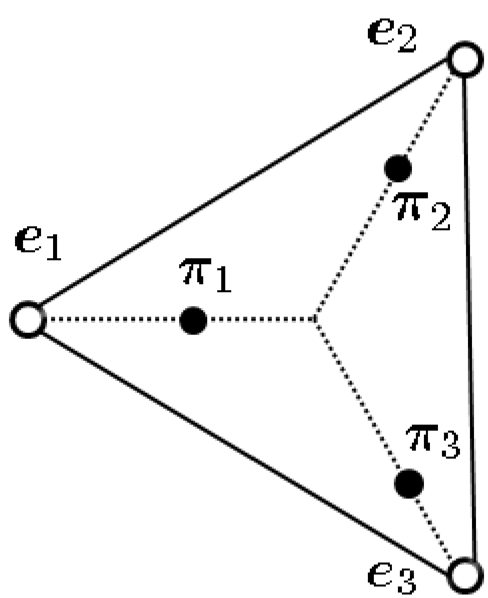
Recoverability (2/2)

- **Binary case:**

$$\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

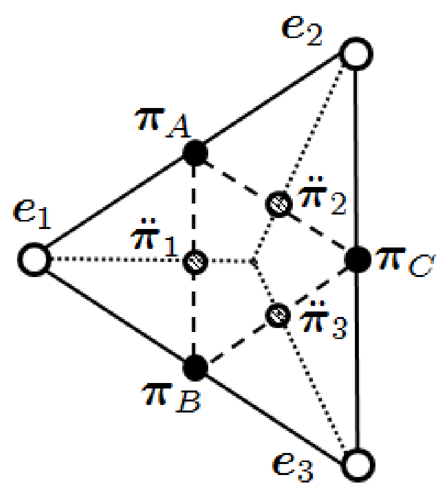
  is recoverable iff $\pi_{12} + \pi_{21} < 1$.

- Recoverability guarenteed by **common noise background model:**



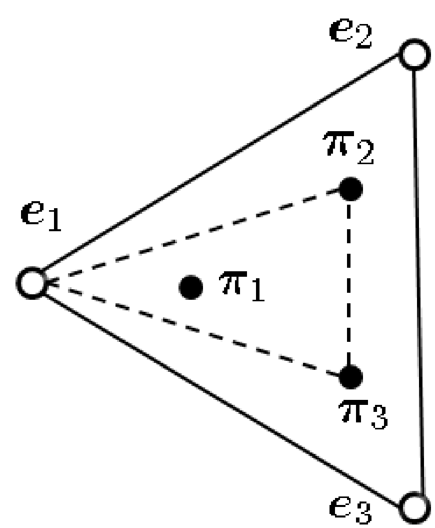
Connections to Other Problems

- Topic modeling: consistent estimation of topics with number of document = number of topics
- Learning with partial labels, $L = 3$. Labels are $A = \{1,2\}$, $B = \{1,3\}$ and $C = \{2,3\}$. Via resampling, can satisfy recoverability assumption.



Contributions

- Universally consistent estimator for maximum mixing proportions
- Sufficient conditions (joint irreducibility, recoverability) for decontamination of mutually contaminated models
- Consistent estimation of π_{ij}
- Consistent discrimination



Consistent Discrimination

- Denote $R_\ell(f) = P_\ell(f(X) \neq \ell)$ and $\widetilde{R}_{\ell j}(f) = \widetilde{P}_j(f(X) \neq \ell)$
- Can estimate

$$R_\ell(f) = \frac{\widetilde{R}_{\ell\ell}(f) - \sum_{j \neq \ell} \nu_{\ell j} \widetilde{R}_{j\ell}(f)}{1 - \kappa_\ell}.$$

via

$$\widehat{R}_\ell(f) := \frac{\widehat{\widetilde{R}}_{\ell\ell}(f) - \sum_{j \neq \ell} \widehat{\nu}_{\ell j} \widehat{\widetilde{R}}_{j\ell}(f)}{1 - \widehat{\kappa}_\ell}.$$

- Consistent discrimination rules can be constructed by establishing probabilistic control of

$$\sup_{f \in \mathcal{F}} \left| R_\ell(f) - \widehat{R}_\ell(f) \right|$$

and following standard arguments (e.g., structural risk minimization over a growing family of VC classes).