Marginal Transfer Learning

Clayton Scott

EECS and Statistics
University of Michigan

Collaborators:
Gilles Blanchard, Ürün Dogan, U. Potsdam
Gyemin Lee, Lloyd Stoolman, U. Michigan

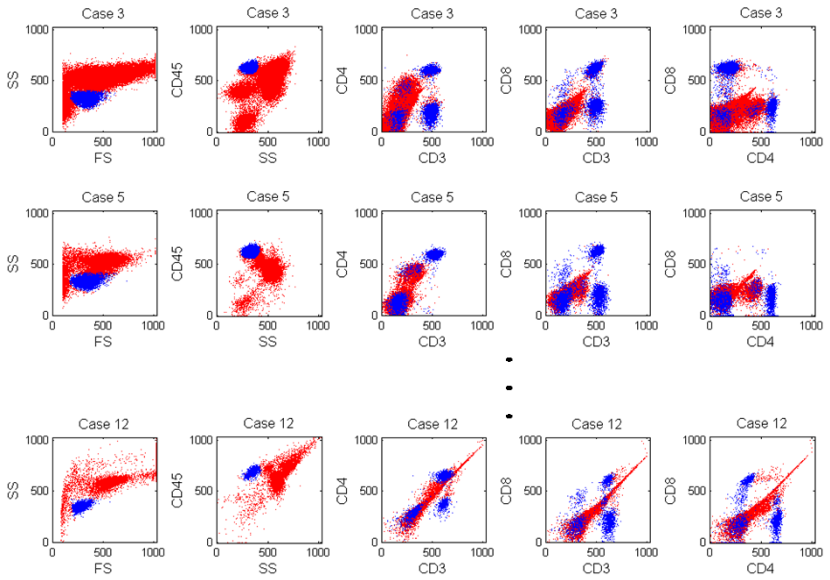
Marginal Transfer

Training data sets

$\mathcal{S}_1 =$

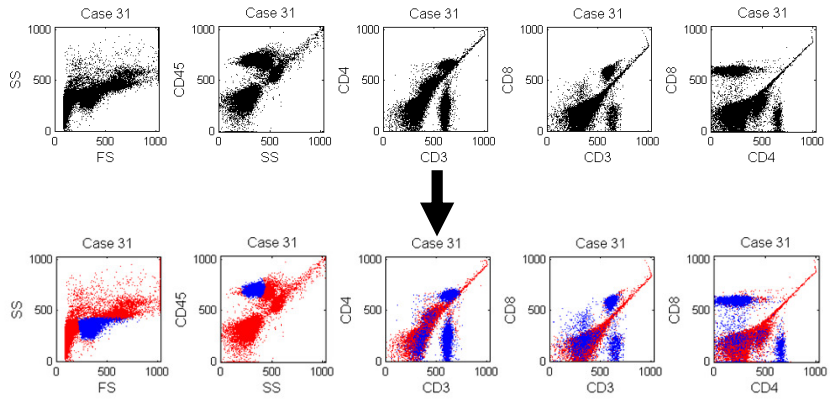
$\mathcal{S}_2 =$

$\mathcal{S}_N =$



Testing data set

$\mathcal{T} =$



Flow cytometry data

Related work:

- Multi-task learning
- Transfer learning

Applications

Application	Feature	Label
Flow cytometry	Cell	Cell type
ECG	Heartbeat	Abnormal heartbeat?
EEG	EEG window	Seizure imminent?
Microchip inspection	Chip	Defect?
…		

Formal Setup

$$\mathcal{X} = \text{feature space (compact)}, \quad \mathcal{Y} = \{-1, 1\}$$

Training data

$$\begin{aligned}
\mathcal{S}_i &= ((X_{ij}, Y_{ij}))_{1 \le j \le n_i} \\
(X_{ij}, Y_{ij}) &\overset{iid}{\sim} P_{XY}^{(i)} \quad \text{for each } i \\
P_{XY}^{(i)} &\overset{iid}{\sim} \mu \\
\mu &\in \mathfrak{B}_{\mathcal{X} \times \mathcal{Y}} \quad (\text{distributions on } \mathcal{X} \times \mathcal{Y})
\end{aligned}$$

Testing data

$$\begin{aligned}
\mathcal{T} &= ((X_j^T, Y_j^T))_{1 \le j \le n_T} \\
(X_j^T, Y_j^T) &\overset{iid}{\sim} P_{XY}^T, \quad Y_j^T \textbf{ not observed} \\
P_{XY}^T &\sim \mu
\end{aligned}$$

Prediction function

- $\mathfrak{B}_{\mathcal{X}}$ = distributions on $\mathcal{X}$
- Map marginal distributions to classifiers

$$g : \mathfrak{B}_{\mathcal{X}} \to (\mathcal{X} \to \mathbb{R})$$

- Equivalent represention:

$$f : \mathfrak{B}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$$

$$f(P_X, x) := g(P_X)(x)$$

- Classifier on "extended feature space" $\mathfrak{B}_{\mathcal{X}} \times \mathcal{X}$

Measuring Performance

- loss

$$\ell(\widehat{y}, y) \quad = \quad \text{loss of prediction value } \widehat{y} \text{ when true label is } y$$

- empirical risk on the test sample

$$\widehat{\mathcal{E}}(f, T) := \frac{1}{n_T} \sum_{i=1}^{n_T} \ell(f(\widehat{P}_X^T, X_i^T), Y_i^T) \ ,$$

- generalization error

$$\mathcal{E}(f) := \mathbb{E}_{P_{XY}^T \sim \mu} \mathbb{E}_{(X^T, Y^T) \sim P_{XY}^T} \left[\ell(f(P_X^T, X^T), Y^T) \right]$$

Kernel-based Algorithm

RHKS framework

- $\overline{k} = \text{kernel on } \mathfrak{B}_{\mathcal{X}} \times \mathcal{X}$
- $\mathcal{H}_{\overline{k}} = \text{RKHS}$
- "extended data"

$$\widetilde{X}_{ij} = (\widehat{P}_X^{(i)}, X_{ij}) \in \mathfrak{B}_{\mathcal{X}} \times \mathcal{X}$$

Minimize empirical risk plus complexity penalty

$$\begin{aligned} \widehat{f}_\lambda &= \arg\min_{f \in \mathcal{H}_{\overline{k}}} \frac{1}{N} \sum_{i=1}^{N} \widehat{\mathcal{E}}(f, \mathcal{S}_i) + \lambda \|f\|^2_{\mathcal{H}_{\overline{k}}} \\ &= \arg\min_{f \in \mathcal{H}_{\overline{k}}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(f(\widetilde{X}_{ij}), Y_{ij}) + \lambda \|f\|^2_{\mathcal{H}_{\overline{k}}} \end{aligned}$$

Implementation

Representer Theorem implies

$$\widehat{f}_\lambda(P_X, x) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \alpha_{ij} \overline{k}((\widehat{P}_X^{(i)}, X_{ij}), (P_X, x))$$

Implementation

- hinge loss $\Longrightarrow$ SVM packages
- logistic loss $\Longrightarrow$ kernel logistic regression algorithms
- etc.

Kernels

Product kernel:

$$\overline{k}((P_1, x_1), (P_2, x_2)) = k_P(P_1, P_2) k_X(x_1, x_2)$$

Kernels on distributions:

- Universal kernels developed by Steinwart and Christmann (NIPS 2010)
- Embedding of distributions: Fix another kernel k'_X on $\mathcal{X}$ and set

$$\Psi(P) := \int k'_X(\cdot, x) dP(x) \in \mathcal{H}_{k'_X}$$

  (related work: Sriperumbudur, Gretton, Fukumizu, Schölkopf, Lanckriet, JMLR 2011)

- Gaussian-like kernel

$$k_P(P_1, P_2) := \exp\left\{ -\frac{1}{2\sigma_P^2} \|\Psi(P_1) - \Psi(P_2)\|^2_{\mathcal{H}_{k'_X}} \right\}$$

Analysis

Assumptions:

- kernels k_X, k_P are universal and bounded
- the loss $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is Lipschitz in its first variable
- all samples $\mathcal{S}_i$ have the same size n

Theorem: With probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{B}_{\overline{k}}(R)} \left| \frac{1}{N} \sum_{i=1}^{N} \widehat{\mathcal{E}}(f, S_i) - \mathcal{E}(f) \right| \leq CR \left(\sqrt{\frac{\log N + \log \delta^{-1}}{n}} + \sqrt{\frac{\log \delta^{-1}}{N}} \right)$$

Corollary: (Universal consistency) If $N, n \to \infty$ with $N = \mathcal{O}(n^\gamma)$ for some $\gamma > 0$, and $\lambda = \lambda(N, n) \to 0$ (but not too fast), then

$$\mathcal{E}(\widehat{f}_{\lambda(N,n)}) \to \inf_{f : \mathfrak{B}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}} \mathcal{E}(f)$$

in probability.

Results

- Flow cytometry data with $N = 35$ patients
- Subsampled $n = 5000$ examples per patient
- leave-one-patient-out test error
- Comparison to a "vanilla" multi-task (MT) learning kernel

$$k_P^\tau(P_1, P_2) := \begin{cases} 1 & \text{if } P_1 = P_2 \\ \tau & \text{if } P_1 \neq P_2 \end{cases}$$

Kernel k_P	Test error	Losses vs. proposed	Wilcoxon signed rank p
MT ($\tau = 0.01$)	1.92 %	29/35	$7\ 10^{-7}$
MT ($\tau = 0.5$)	1.72 %	26/35	$9\ 10^{-4}$
Pooling ($\tau = 1$)	1.71 %	26/35	$2.5\ 10^{-3}$
Proposed	1.67 %	-	-

Reference

"Generalizing from Several Related Classification Tasks to a New Unlabeled Sample"

NIPS 2011

Available at: http://www.eecs.umich.edu/~cscott

Supported by NSF