



Kernel Methods for Transfer Learning

Clayton Scott

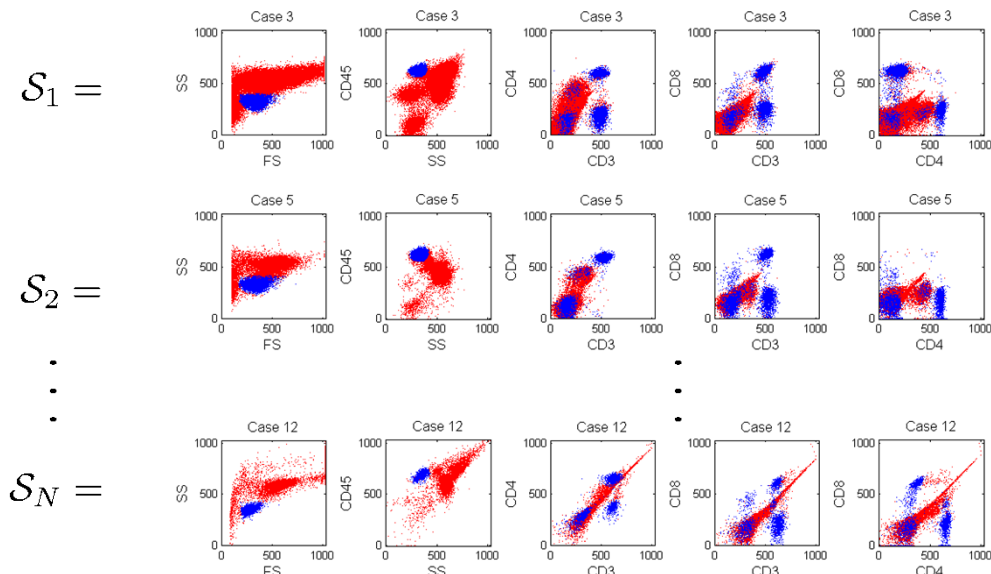
EECS and Statistics
University of Michigan

Collaborators:

Gilles Blanchard, Gyemin Lee, Lloyd Stoolman

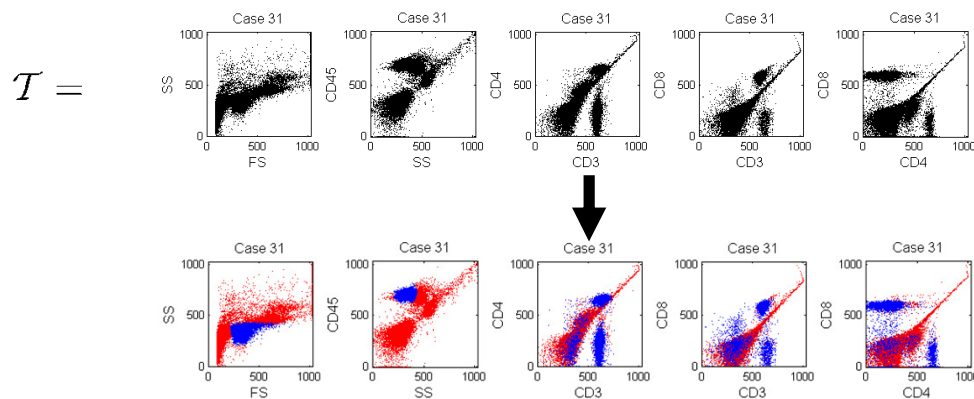
Transfer Learning in Flow Cytometry

Training data sets



Flow cytometry data

Testing data set



Classification

\mathcal{X} = set of patterns

\mathcal{Y} = set of class labels

Problem:

Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$

Example: Heart attack within next 5 years?

X = (age, weight) $\in \mathbb{R}^2$

Y = yes or no

Probabilistic Framework

Assume (X, Y) is a **random variable** with probability measure \mathbf{P}

Note that Y is not necessarily a deterministic function of X

Classifiers

Assume $\mathcal{Y} = \{-1, +1\}$. A real-valued function

$$f : \mathcal{X} \rightarrow \mathbb{R}$$

defines a **classifier** via

$$\hat{y} = \text{sign}(f(x)).$$

The **probability of error**, or **risk** of f is

$$\mathcal{E}(f) := \mathbf{E}[\ell(f(X), Y)]$$

where

$$\ell(t, y) = \mathbf{1}_{\{\text{sign}(t) \neq y\}}$$

Learning from Data

\mathbf{P} is unknown and potentially quite complex
(not amenable to parametric modeling)

A **training sample** is a collection

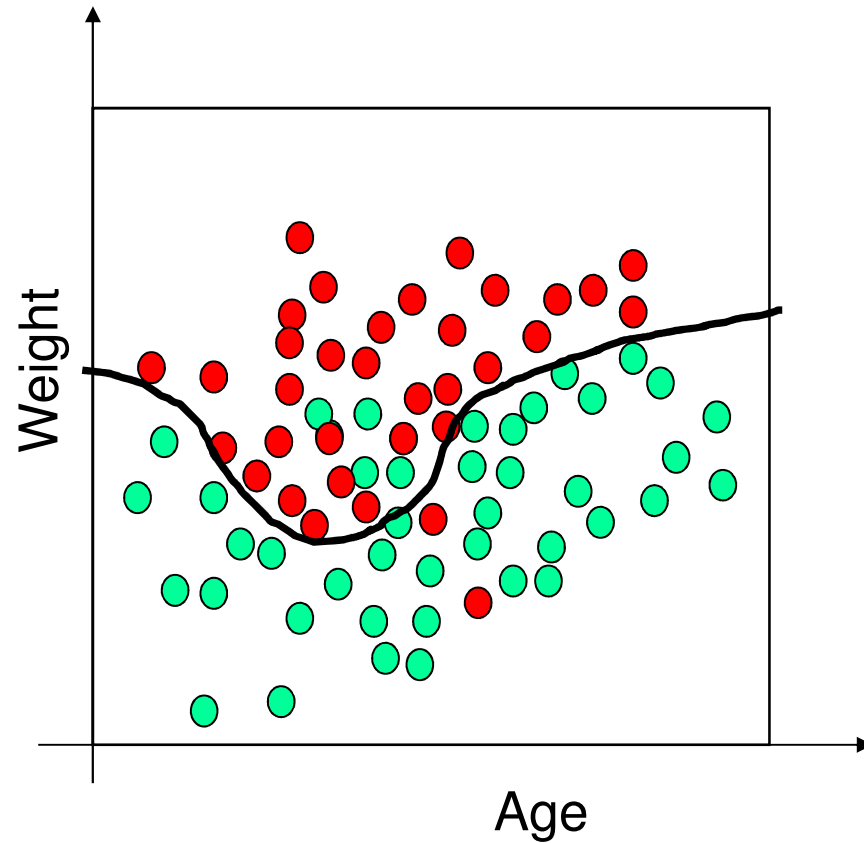
$$((X_1, Y_1), \dots, (X_n, Y_n))$$

of **IID** realizations of (X, Y) .

The **goal of classification** is to construct a classifier \hat{f} from the training sample such that $\mathcal{E}(\hat{f})$ is small, e.g., as $n \rightarrow \infty$

$$\mathcal{E}(\hat{f}) \xrightarrow{i.p.} \mathcal{E}^* = \inf_{\text{all } f} \mathcal{E}(f)$$

Learning from Data



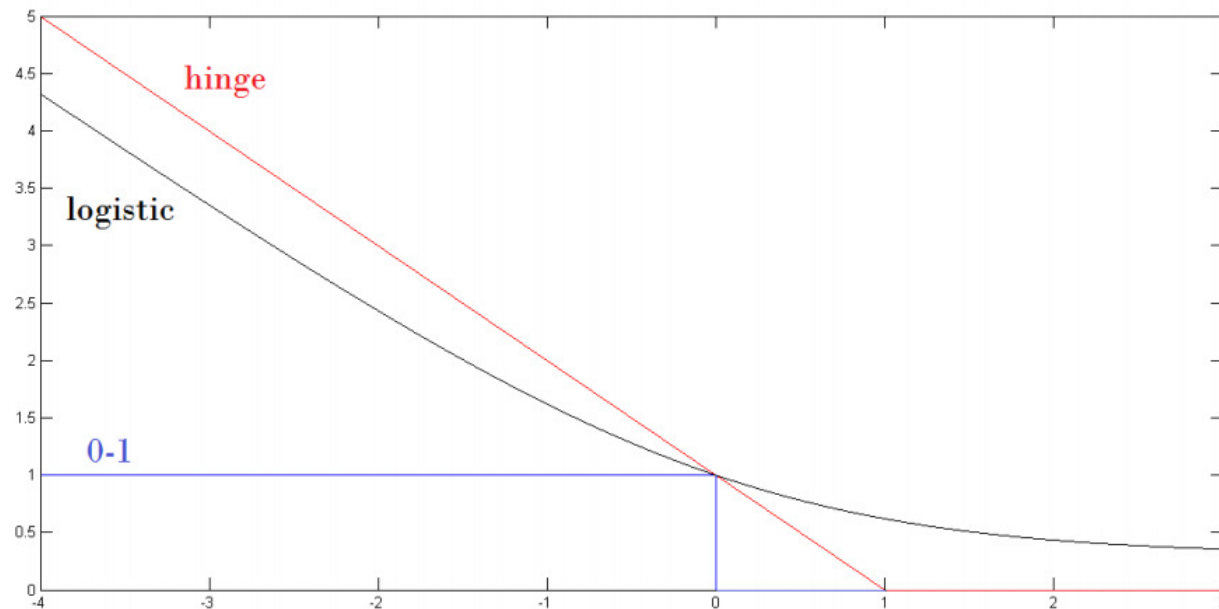
Task: construct a classifier using a training sample

Kernel Methods

$$\hat{f}_\lambda = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

where

- \mathcal{H} is a reproducing kernel Hilbert space
- $\ell(t, y) = \phi(yt)$ where ϕ is convex



Reproducing Kernel Hilbert Spaces

A **reproducing kernel Hilbert space** is a Hilbert space \mathcal{H} whose elements are functions $f : \mathcal{X} \rightarrow \mathbb{R}$, and such that there exists a function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ satisfying

- $k(\cdot, x) \in \mathcal{H}$ for every x
- for every $f \in \mathcal{H}$ and $x \in \mathcal{X}$,

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$$

(reproducing property)

Any k satisfying the above conditions is called a **reproducing kernel**

Reproducing Kernel Hilbert Spaces

Useful facts about reproducing kernel Hilbert spaces:

- k is a reproducing kernel iff k is positive definite
- Every RKHS has a unique kernel
- Every positive definite kernel is the reproducing kernel of a unique RKHS
- The set of functions

$$\mathcal{H}_0 = \left\{ f = \sum_{j=1}^m \alpha_j k(\cdot, z_j) \right\}$$

is dense in the RKHS associated with k

The Representer Theorem

Every solution of

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

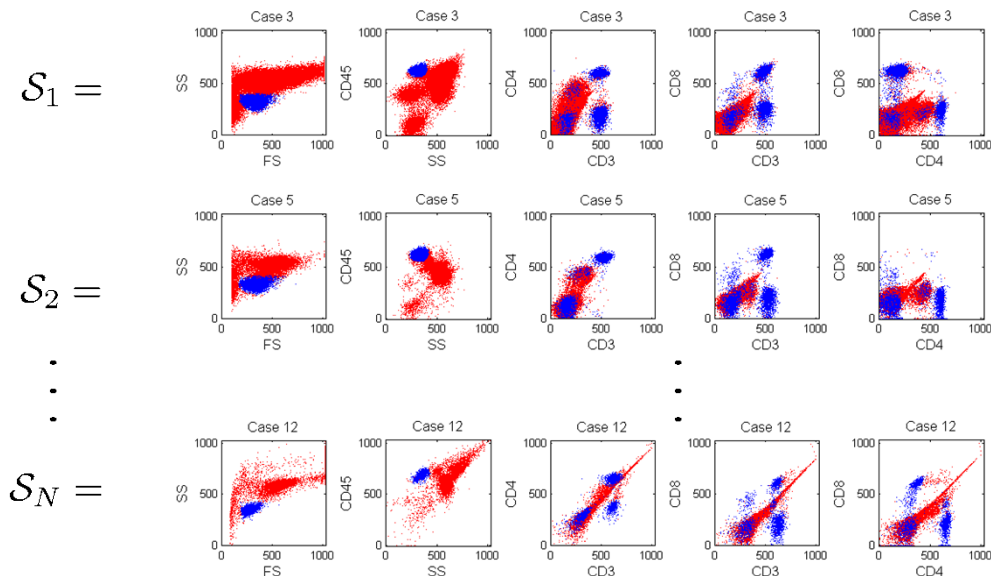
has the form

$$f = \sum_{i=1}^n \alpha_i k(\cdot, X_i)$$

where $\alpha_i \in \mathbb{R}$.

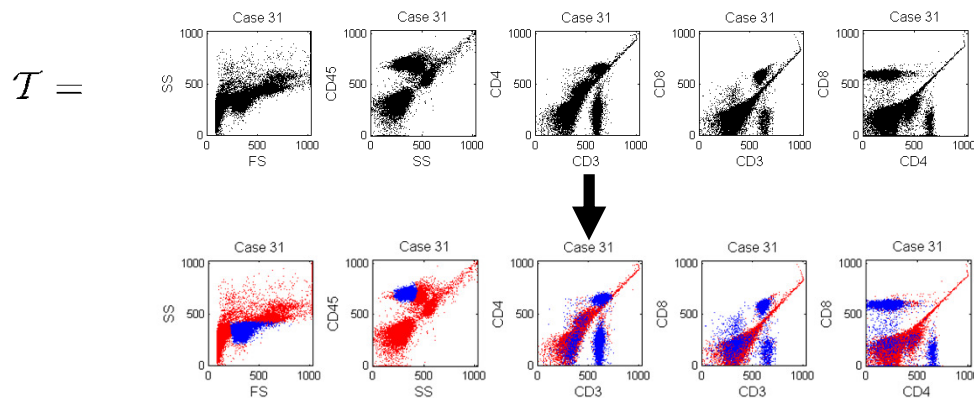
Transfer Learning in Flow Cytometry

Training data sets



Flow cytometry data

Testing data set



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

Training data

$$\begin{aligned} \mathcal{S}_i &= ((X_{ij}, Y_{ij}))_{1 \leq j \leq n_i} \\ (X_{ij}, Y_{ij}) &\stackrel{iid}{\sim} P_{XY}^{(i)} \quad \text{for each } i \\ P_{XY}^{(i)} &\stackrel{iid}{\sim} \mu \quad (\text{distribution on distributions}) \end{aligned}$$

Testing data

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

Prediction function

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

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

- Equivalent representation:

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

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

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

Measuring Performance

- loss

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

- empirical risk on the test sample

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

- generalization error

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

Kernel-based Algorithm

RHKS framework

- \bar{k} = kernel on $\mathfrak{B}_{\mathcal{X}} \times \mathcal{X}$
- $\mathcal{H}_{\bar{k}}$ = RKHS
- “extended data”

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

Minimize empirical risk plus complexity penalty

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

Implementation

Representer Theorem implies

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

Implementation

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

Kernels

Product kernel:

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

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

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

Analysis

Assumptions:

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

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

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

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

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

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	$7 \cdot 10^{-7}$
MT ($\tau = 0.5$)	1.72 %	26/35	$9 \cdot 10^{-4}$
Pooling ($\tau = 1$)	1.71 %	26/35	$2.5 \cdot 10^{-3}$
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