Active Learning via Transductive Experimental Design	ICML 2006
Non-greedy Active Learning for Text Categorization using Convex Transductive Experimental Design	SIGIR 2008
Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples	JMLR 2006
Beyond the Point Cloud: from Transductive to Semi-supervised Learning	ICML 2005
Manifold Regularized Experimental Design for Active Learning	TIP 2017
Semi-Supervised Learning with Graphs	PHD 2005

2018.3.29

Experimental design

Classic experiment design considers learning a linear function $f(x) = w^T x \quad w \in \mathbb{R}^d$, from measurements $y_i = w^T x_i + \epsilon_i, i = 1, \dots, m, \epsilon_i \sim N(0, \sigma^2)$ is measurement noise. $x_1, .., x_m$ are are experiments chosen from n candidates $v_1, \dots, v_n \in \mathbb{R}^d, n > m$.

The goal of experimental design is to **find a set of experiments** x_i that together are maximally informative.

$$X: [x_1, \dots, x_m]^T \in \mathbb{R}^{m \times d}, set\{x_i\} \ |X| = m \quad \text{V}: [v_1, \dots, v_n]^T \in \mathbb{R}^{n \times d}, set\{v_i\} \ |V| = n$$

The maximum-likelihood estimate of w:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ J(\mathbf{w}) = \sum_{i=1}^{m} \left(\mathbf{w}^\top \mathbf{x}_i - y_i \right)^2 \right\}$$

Estimation error $e = w - \hat{w}$ $mean: 0$ $covariance\ matrix$: $\sigma^2 C_w$, C_w is the inverted Hessian of $J(w), \sigma$ is a constant

$$\mathbf{C_w} = \left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1}$$

The matrix C_w characterizes the confidence of the estimation, or the informativeness of the selected data.

Let m_j denote the number of times for which v_j is chosen in **X,** $m_1 + \cdots m_n = m$

$$\min_{m_1,\ldots,m_n} \quad \mathrm{Tr}\left[\left(\sum_{j=1}^{n} m_j \mathbf{v}_j \mathbf{v}_j^\top \right)^{-1} \right]$$

$$\text{subject to} \quad m_j \geq 0, m_1 + \cdots + m_n = m, m_i \in \mathbb{Z}$$

$\tau_j = m_j / m$

$$\min_{\tau_1,\ldots,\tau_n} \quad \mathrm{Tr}\left[\left(\sum_{j=1}^{n} \tau_j \mathbf{v}_j \mathbf{v}_j^\top \right)^{-1} \right] \qquad \text{subject to} \quad \boldsymbol{\tau} \succeq 0, \mathbf{1}^\top \boldsymbol{\tau} = 1$$

Transductive Experimental Design

- The optimization criteria based on C_w does not directly characterize the quality of predictions on test data
- Standard experimental design only considers linear functions and is thus restrictive in applications.
- Very importantly, classic experimental design has to solve a SDP problem, which is often very slow when dealing with hundreds of data points

A general setting may consider a different set **T** of test data points besides candidates in **V**. Assume the two sets are the same.

$$\min_{\mathbf{w}} \left\{ J(\mathbf{w}) = \sum_{i=1}^{m} \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2 + \mu \|\mathbf{w}\|^2 \right\}$$

$$\mathbf{C}_\mathbf{w} = \left(\frac{\partial J(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \right)^{-1} = (\mathbf{X}^\top \mathbf{X} + \mu \mathbf{I})^{-1}$$

$\mathbf{f} = \left[f\left(v_1, \ldots, f\left(v_n\right)\right)\right]$ be the function values on all the available data **V,**
the predictive error $\mathbf{f} - \hat{\mathbf{f}}$ has the covariance matrix $\sigma^2 C_{\mathrm{f}}$

$$\begin{aligned}\mathbf{C_f} &= \mathbf{V}\mathbf{C_w}\mathbf{V}^\top = \mathbf{V}(\mathbf{X}^\top\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{V}^\top \\ &= \frac{1}{\mu}\left[\mathbf{V}\mathbf{V}^\top - \mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right]\end{aligned}$$

Woodbury matrix identity

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

The average predictive variance on **V** is given by $\frac{\sigma^2}{n}\mathrm{Tr}(C_{\mathrm{f}})$

Transductive experimental design

$$\begin{aligned}\max_{\mathbf{X}} \quad & Tr\left[\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right] \\ subject\ to \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m\end{aligned}$$

$$\mathrm{Tr}(\mathbf{C_f}) = \mathrm{Tr}(\mathbf{C_w}\mathbf{V}^\top\mathbf{V})$$

Transductive experimental design is equivalent to

$$\begin{aligned} \min_{\mathbf{X},\mathbf{A}} \quad & \sum_{i=1}^{n} \|\mathbf{v}_i - \mathbf{X}^\top \mathbf{a}_i\|^2 + \mu \|\mathbf{a}_i\|^2 \\ \textit{subject to} \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m, \\ & \mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\top \in \mathbb{R}^{n \times m} \end{aligned}$$

Kernel Transductive Experimental Design

$$k(\mathbf{x}, \mathbf{v}) = \langle \phi(\mathbf{x}), \phi(\mathbf{v}) \rangle, \ \mathbf{x}, \mathbf{v} \in \mathbb{R}^d$$

$$\begin{aligned} \max_{\mathbf{X}} \quad & \mathrm{Tr}\left[\mathbf{K_{vx}}(\mathbf{K_{xx}} + \mu\mathbf{I})^{-1}\mathbf{K_{xv}}\right] \\ \text{subject to} \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m \end{aligned}$$

$$\begin{aligned} \max_{\mathbf{X}} \quad & Tr\left[\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right] \\ \textit{subject to} \quad & \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m \end{aligned}$$

$$(\mathbf{K})_{ij} = k(\mathbf{v}_i, \mathbf{v}_j), \ (\mathbf{K_{vx}})_{ij} = k(\mathbf{v}_i, \mathbf{x}_j) \quad (\mathbf{K_{xx}})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

Given previously selected data X_1, a sequential transductive design seeks *m* new data $X_2 \subset V$ in the following way.

$$\begin{aligned} \max_{\mathbf{X}_2} \quad & \mathrm{Tr}\left[\mathbf{K}_{\mathbf{vx}}(\mathbf{K}_{\mathbf{xx}}+\mu\mathbf{I})^{-1}\mathbf{K}_{\mathbf{xv}}\right] \\ \text{subject to} \quad & \mathbf{X}=\mathbf{X}_1\cup\mathbf{X}_2, \mathbf{X}_2\subset\mathbf{V}, |\mathbf{X}_2|=m \end{aligned}$$

$$\begin{aligned} \max_{\mathbf{X}_2} \quad & \mathrm{Tr}\left[\tilde{\mathbf{K}}_{\mathbf{vx}_2}(\tilde{\mathbf{K}}_{\mathbf{x}_2\mathbf{x}_2}+\mu\mathbf{I})^{-1}\tilde{\mathbf{K}}_{\mathbf{x}_2\mathbf{v}}\right] \\ \text{subject to} \quad & \mathbf{X}_2\subset\mathbf{V}, |\mathbf{X}_2|=m \end{aligned}$$

kernel matrix $\tilde{K}$ is obtained by deflating the original kernel matrix K by X_1:

$$\tilde{\mathbf{K}}=\mathbf{K}-\mathbf{K}_{\mathbf{vx}_1}(\mathbf{K}_{\mathbf{x}_1\mathbf{x}_1}+\mu\mathbf{I})^{-1}\mathbf{K}_{\mathbf{x}_1\mathbf{v}}$$

procedure:

After approximating V by X_1, the approximation residuals $\tilde{V}$ form a new kernel matrix $\tilde{K}=\tilde{V}\tilde{V}^T$, and a set of *m* vectors from $\tilde{V}$ are selected to further approximate $\tilde{V}$.

kernel version

Algorithm 1: Sequential Design

- Select $\mathbf{x} \in \mathbf{V}$ with the highest $\|\mathbf{K}_\mathbf{x}\|^2/(k(\mathbf{x},\mathbf{x}) + \mu)$, and add $\mathbf{x}$ into $\mathbf{X}$, where $\mathbf{K}_\mathbf{x}$ and $k(\mathbf{x},\mathbf{x})$ are $\mathbf{x}$'s corresponding column and diagonal entry in current $\mathbf{K}$;
- Update $\mathbf{K} \leftarrow \mathbf{K} - \frac{\mathbf{K}_\mathbf{x}\mathbf{K}_\mathbf{x}^\top}{(k(\mathbf{x},\mathbf{x})+\mu)}$

Alternating Optimization

Let $Q = [q_1 \ldots, q_n]^T$ and $\pi_1 \geq \cdots \geq \pi_n$ be the eigenvectors and eigenvalues of $K = VV^T$. Then transductive experimental design is equivalent to

$$\min_{\mathbf{X},\mathbf{C}} \quad \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \mathbf{K}_{\mathbf{vx}}\mathbf{c}_i\|^2 + \mu\pi_i\|\mathbf{c}_i\|^2$$

$$\textit{subject to} \quad \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m, \quad \mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_n]^\top \in \mathbb{R}^{n\times m}$$

$$\min_{\mathbf{B},\boldsymbol{\alpha}_i} \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \mathbf{K}\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \mu\pi_i\|\mathbf{B}\boldsymbol{\alpha}_i\|^2$$

$$\text{subject to } \mathbf{B} = diag(\boldsymbol{\beta}), \quad \text{Card}(\boldsymbol{\beta}) = m, \quad \beta_j \in \{0,1\}, \quad j = 1, \cdots, n.$$

If $\beta_j = 1$, v_j is included in X. Then $\mathbf{K}_{\mathbf{vx}}\mathbf{c}_i = \mathbf{K}\mathbf{B}\boldsymbol{\alpha}_i$

$$\min_{\mathbf{B},\alpha_i} \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \mathbf{K}\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \mu\pi_i\|\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \gamma\|\boldsymbol{\beta}\|_1$$

$$\text{subject to} \quad \mathbf{B} = diag(\boldsymbol{\beta}), \quad \beta_j \geq 0, \quad j = 1, \cdots, n.$$

$$\|\boldsymbol{\beta}\|_1 = \sum \beta_j$$

Algorithm 2: Alternating Design

- Fix $\mathbf{B}$ to the current solution (initially to the identity matrix $\mathbf{I}$), convert $\tilde{\mathbf{K}} \leftarrow \mathbf{K}\mathbf{B}$, solve the following problem for optimal $\boldsymbol{\alpha}_i$,

$$\min_{\boldsymbol{\alpha}_i} \quad \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \tilde{\mathbf{K}}\boldsymbol{\alpha}_i\|^2 + \mu\pi_i\|\mathbf{B}\boldsymbol{\alpha}_i\|^2$$

- Fix $\boldsymbol{\alpha}_i$ to the solution obtained at the above step, convert $\mathbf{K}_i \leftarrow \mathbf{K} \cdot diag(\boldsymbol{\alpha}_i)$, solve the following problem for optimal $\hat{\boldsymbol{\beta}}$,

$$\min_{\boldsymbol{\beta}\geq 0} \quad \sum_{i=1}^{n} \|\sqrt{\pi_i}\mathbf{q}_i - \mathbf{K}_i\boldsymbol{\beta}\|^2 + \mu\pi_i\|\boldsymbol{\beta} \otimes \boldsymbol{\alpha}_i\|^2 + \gamma\|\boldsymbol{\beta}\|_1$$

- $\mathbf{B} \leftarrow \mathbf{B} \otimes diag(\hat{\boldsymbol{\beta}})$ $\otimes$ denotes the component-wise multiplication between two matrices

Non-greedy Active Learning for Text Categorization using Convex Transductive Experimental Design

SIGIR 2008

$$\min_{\mathcal{A}\subset\mathcal{C},|\mathcal{A}|=K} \frac{1}{M}\mathrm{trace}\left(\mathbf{X}_{\mathcal{P}}\mathbf{H}^{-1}\mathbf{X}_{\mathcal{P}}^{\top}\right) \qquad \mathbf{H} = \frac{\partial J(\mathbf{w};\mathbf{X}_{\mathcal{A}})}{\partial \mathbf{w}\partial \mathbf{w}^{\top}} = \mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{\top} + \mu\mathbf{I}$$

$$\begin{aligned} \min_{\mathcal{A},\boldsymbol{\alpha}_i\in\mathbb{R}^K} \quad & \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{A}}^{\top}\boldsymbol{\alpha}_i\|^2 + \mu\|\boldsymbol{\alpha}_i\|^2 \\ \text{subject to} \quad & |\mathcal{A}| = K, \quad \mathcal{A}\subset\mathcal{C}, \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}} \end{aligned}$$

TED has a geometric interpretation that it tends to find representative data X_A spanning a linear subspace to retain most of the information of the whole set of test data X_P

Therefore, given a sufficiently X_P large set , TED actually explores the information about the distribution of unlabeled data.

$$\min_{\boldsymbol{\beta},\boldsymbol{\alpha}_i\in\mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{C}}^{\top}\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \mu\|\mathbf{B}\boldsymbol{\alpha}_i\|^2 + \gamma\|\boldsymbol{\beta}\|_1$$
$$\text{subject to} \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}}, \quad \mathbf{B} = \text{diag}(\boldsymbol{\beta}), \quad \mathbf{B} \succeq 0.$$

The optimization is done by alternatively optimizing β_j or $\boldsymbol{\alpha}_j$ while fixing the other.

A Convex Formulation

$$\min_{\boldsymbol{\beta},\boldsymbol{\alpha}_i\in\mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{C}}^{\top}\boldsymbol{\alpha}_i\|^2 + \sum_{j=1}^{N} \frac{\alpha_{i,j}^2}{\beta_j} + \gamma\|\boldsymbol{\beta}\|_1$$
$$\text{subject to} \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}}, \quad \beta_j \geq 0, \quad j = 1, \cdots, N,$$

$$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_N]$$
$$\boldsymbol{\alpha}_i = [\alpha_{i,1}, \ldots, \alpha_{i,N}]^{\top}$$

$$\frac{\sum_{i=1}^{M} \alpha_{i,j}^2}{|\beta_j|} + \gamma|\beta_j| \geq 2\sqrt{\gamma \sum_{i=1}^{M} \alpha_{i,j}^2},$$

$$\beta_j^2 = \frac{1}{\gamma}\sum_{i=1}^{M} \alpha_{i,j}^2$$

$$\min_{\boldsymbol{\alpha}_i\in\mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{C}}^{\top}\boldsymbol{\alpha}_i\|^2 + 2\sum_{j=1}^{N} \sqrt{\gamma \sum_{i=1}^{M} \alpha_{i,j}^2}$$

$$\boldsymbol{\alpha}_i = (\mathrm{diag}(\boldsymbol{\beta})^{-1} + \mathbf{X}_{\mathcal{C}}\mathbf{X}_{\mathcal{C}}^{\top})^{-1}\mathbf{X}_{\mathcal{C}}\mathbf{x}_i, \quad i = 1, \ldots, M$$

$$\beta_j = \sqrt{\frac{1}{\gamma}\sum_{i=1}^{M}\alpha_{i,j}^2}, \quad j = 1, \ldots, N.$$

Algorithm 2 Convex TED

Require: candidates $\mathbf{X}_{\mathcal{C}}$, unlabeled data $\mathbf{X}_{\mathcal{P}}$, $\gamma > 0$;

1: initialize $(\alpha_{i,j})$;
2: **repeat**
3: $\quad \beta_j \leftarrow \sqrt{\frac{1}{\gamma}\sum_{i=1}^{M}\alpha_{i,j}^2}$ for $j = 1, \ldots, N$;
4: $\quad \boldsymbol{\alpha}_i \leftarrow (\mathrm{diag}(\boldsymbol{\beta})^{-1} + \mathbf{X}_{\mathcal{C}}\mathbf{X}_{\mathcal{C}}^{\top})^{-1}\mathbf{X}_{\mathcal{C}}\mathbf{x}_i$, for $i = 1, \ldots, M$;
5: **until** converge;
6: $\mathbf{X}_{\mathcal{A}} \leftarrow \{\mathbf{x}_j | \mathbf{x}_j \in \mathbf{X}_{\mathcal{C}}, \beta_j \neq 0\}$;
7: **return** $\mathbf{X}_{\mathcal{A}}$

Beyond the Point Cloud: from Transductive to Semi-supervised Learning

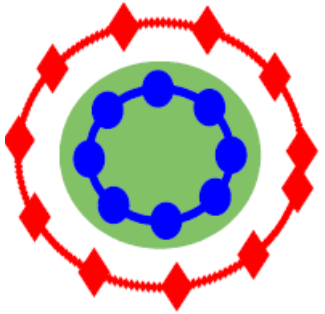
ICML 2005

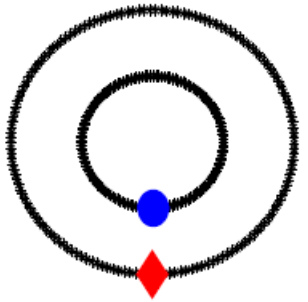
(x_i, y_i) $x_i \in \mathbb{R}^2$ $y_i \in \{-1, +1\}$

$$f = \arg\min_{h \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} V(h, x_i, y_i) + \gamma \|h\|_{\mathcal{H}}^2$$

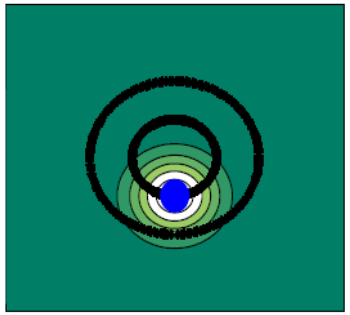
The solution can be expressed as

$$f(x) = \sum_{i=1}^{l} \alpha_i k(x, x_i)$$

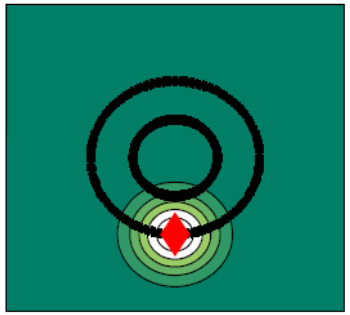




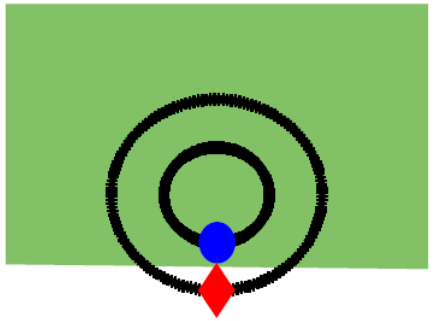
(a) gaussian kernel centered on labeled point 1



(b) gaussian kernel centered on labeled point 2



(c) classifier learnt in the RKHS



Can we define a kernel $\tilde{k}$ that is adapted to the geometry of the data distribution?

Such a kernel $\tilde{k}$ must have the property:

It is a valid Mercer kernel $\tilde{k}: X \times X \to \mathbb{R}$ and therefore denes a new RKHS $\tilde{H}$.

It implements our intuitions about the geometry of the data

$$g = \arg\min_{h \in \tilde{\mathcal{H}}} \frac{1}{2} \sum_{i=1}^{2} V(h, x_i, y_i) + \|h\|^2_{\tilde{\mathcal{H}}} \qquad g(x) = \sum_{i=1}^{2} \alpha_i \tilde{k}(x, x_i)$$

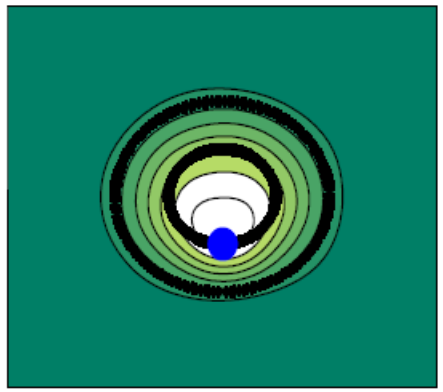
this solution must produce an intuitive decision surface that separates the two circles

$$\tilde{k}(x, z) = k(x, z) - \mathbf{k}_x^t (I + MK)^{-1} M \mathbf{k}_z$$

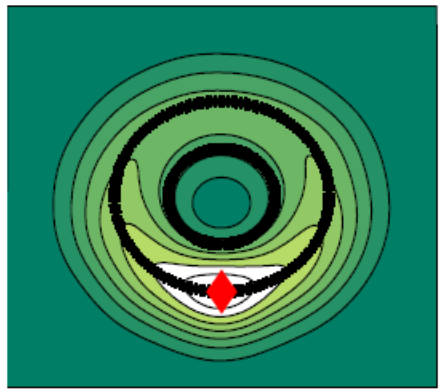
$$M = L^p$$

$$L = D - W \qquad W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \qquad D_{ii} = \sum_i W_{ij}$$

(a) deformed kernel centered on labeled point 1



(b) deformed kernel centered on labeled point 2



(c) classifier learnt in the deformed RKHS

Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples

JMLR 2006

If two points $x_1, x_2 \in X$ are close in the intrinsic geometry of P_X , then the conditional distributions $P(y|x_1)$ and $P(y|x_2)$ are similar.

the conditional probability distribution $P(y|x)$ varies smoothly along the geodesics in the intrinsic geometry of P_X.

For a Mercer kernel $K: X \times X \to R$, there is an associated RKHS $\mathcal{H}_K$ of functions $X \to \mathbb{R}$ with the corresponding norm $||*||_K$. Given a set of labeled examples $(x_i, y_i), i = 1, \ldots, l$

$$f^* = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma ||f||_K^2$$

$$f^*(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x)$$

Marginal P_X is Known

Our goal is to incorporate additional information about the geometric structure of the marginal P_X

$$f^* = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A ||f||_K^2 + \gamma_I ||f||_I^2$$

If the probability distribution is supported on a low-dimensional manifold, $||f||_I^2$ may penalize f along that manifold. γ_A controls the **complexity of the function in the ambient space** while γ_I controls the **complexity of the function in the intrinsic geometry** of P_X.

Theorem 1

Assume that the penalty term $||f||_I$ respect to the RKHS norm $||f||_K$

$$f^*(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) + \int \alpha(z) K(x, z) dP_x(z)$$

Marginal P_X Unknown

we must attempt to get empirical estimates of P_X and $||*||_I$.

when the support of P_X is a compact submanifold $\mathcal{M} \subset \mathbb{R}^n$. $\quad ||f||_I \equiv \int \left||\nabla_{\mathcal{M}} f|\right|^2 dP_X(x)$

$\nabla_{\mathcal{M}}$ is the gradient of f along the manifold $\mathcal{M}$ and the integral is taken over the marginal distribution.

$$f^* = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A ||f||_K^2 + \gamma_I \int \left||\nabla_{\mathcal{M}} f|\right|^2 dP_X(x)$$

$\int \left||\nabla_{\mathcal{M}} f|\right|^2 dP_X(x)$ may be approximated on the basis of labeled and unlabeled data using the graph Laplacian associated to the data

Under certain conditions choosing exponential weights for the adjacency graph leads to convergence of the graph Laplacian to the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ (or its weighted version) on the manifold

$$\{(x_i, y_i)\}_{i=1}^{l} \qquad \{(x_j)\}_{j=l+1}^{j=l+u}$$

$$f^* = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A ||f||_K^2 + \frac{\gamma_I}{(u+l)^2} \sum_{i,j=1}^{l+u} (f(x_i) - f(x_j))^2 W_{ij}$$

$$= \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A ||f||_K^2 + \frac{\gamma_I}{(u+l)^2} \mathbf{f}^T L \mathbf{f}$$

$$\mathbf{f} = [f(x_1), \ldots, f(x_{l+u})]^T \qquad L = D - \mathrm{W} \qquad D_{ii} = \sum_{j=1}^{l+u} W_{ij}$$

Manifold Regularized Experimental Design for Active Learning TIP 2017

MRED for Active Learning

$$w^* = \arg\min_{w} \left\{ J(w) = \sum_{i=1}^{l} (w^T z_i - y_i)^2 + \gamma_1 ||w||^2 + \gamma_2 ||w||_I^2 \right\}$$

we employ a data-dependent deformed kernel function to incorporate the manifold structure of abundant unlabeled samples, We use H_K and $\widetilde{H}_{\widetilde{K}}$ to denote the original RKHS and the new kernel space.

$$\tilde{k}(x_i, x_j) = k(x_i, x_j) - \gamma k_{x_i}^T (I + ML)^{-1} M k_{x_j}$$

kernel Gram matrix $K = [k(x_i, x_j)]_{n \times n}$ $k_{x_i} = [k(x_i, x_1), \ldots, k(x_i, x_n)]^T$

Adopt the graph Laplacian to capture the intrinsic manifold of unlabeled samples.

$$W_{ij} = \begin{cases} 1, & if\ x_i \in (x_j)\ or\ x_j \in (x_i) \\ 0, & otherwise \end{cases}$$

$$\hat{w}^* = \arg\min_{\hat{w}\in\tilde{H}_{\tilde{K}}} \left\{ J(\hat{w}) = \sum_{i=1}^{l} (\hat{w}^T \tilde{\phi}(z_i) - y_i)^2 + \gamma_1 ||\hat{w}||^2 \right\}$$

$\tilde{\phi}(z_i)$ indicates the data sample z_i in the high dimensional kernel space $\tilde{H}_K$, which shows the intrinsic manifold structure of a large number of the unlabeled samples in the database.

we notice that $\hat{w}^*$ is defined as a linear combination of $\tilde{\phi}(z_i),\ i = 1, \dots, l$:

$$\hat{w} = \sum_{i=1}^{l} v_i \tilde{\phi}(z_i) = \tilde{\phi}(Z)v, \qquad v = [v_1, \dots, v_l]^T \in R^l \qquad \tilde{\phi}(Z) = [\tilde{\phi}(z_1), \dots, \tilde{\phi}(z_l)]$$

$$\hat{w}^* = \arg\min_{\hat{w}\in\tilde{H}_{\tilde{K}}} \left\{ J(v) = \left\| \tilde{K}_z v - y \right\|^2 + \gamma_1 v^T \tilde{K}_z v \right\}, \qquad y = [y_1, \dots, y_l]^T \quad \tilde{K}_z \in R^{l\times l}$$

$$\frac{\partial J(v)}{\partial v} = 0 \qquad v^* = (\tilde{K}_z + \gamma_1 I)^{-1} y \qquad f(x) = \sum_{i=1}^{l} \tilde{k}(x, z_i) v^*$$

MRED Solution

Find the informative samples by minimizing the expected prediction variance on the test data.

$$\min_{\alpha_i \in R^l} \sum_{i=1}^{l} \left\| \tilde{\phi}(x_i) - \tilde{\phi}(Z)\alpha_i \right\|^2 + \gamma_1 \|\alpha_i\|^2$$

$$\tilde{\phi}(Z) = [\tilde{\phi}(z_1), \dots, \tilde{\phi}(z_l)]$$

$$\min_{\alpha_i, \beta \in R^l} \sum_{i=1}^{l} \left(\left\| \tilde{\phi}(x_i) - \tilde{\phi}(X)\alpha_i \right\|^2 + \sum_{j=1}^{n} \frac{\alpha_{i,j}^2}{\beta_j} \right) + \lambda \|\beta\|_1$$

$$s.t. \beta_j \geq 0, \qquad j = 1, \dots, n$$

$$\alpha_i = (\alpha_{i,1}, \dots, \alpha_{i,n})^T$$

分

割

线

$$\min_{\mathcal{A}, \boldsymbol{\alpha}_i \in \mathbb{R}^K} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{A}}^{\top} \boldsymbol{\alpha}_i\|^2 + \mu \|\boldsymbol{\alpha}_i\|^2$$

$$\text{subject to} \quad |\mathcal{A}| = K, \quad \mathcal{A} \subset \mathcal{C}, \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}}$$

$$\min_{\boldsymbol{\beta}, \boldsymbol{\alpha}_i \in \mathbb{R}^N} \quad \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_{\mathcal{C}}^{\top} \boldsymbol{\alpha}_i\|^2 + \sum_{j=1}^{N} \frac{\alpha_{i,j}^2}{\beta_j} + \gamma \|\boldsymbol{\beta}\|_1$$

$$\text{subject to} \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}}, \quad \beta_j \geq 0, \quad j = 1, \cdots, N,$$

$$\boldsymbol{\beta} = [\beta_1, \dots, \beta_N]$$

$$\boldsymbol{\alpha}_i = [\alpha_{i,1}, \dots, \alpha_{i,N}]^{\top}$$

Algorithm 2 Convex TED

Require: candidates $\mathbf{X}_{\mathcal{C}}$, unlabeled data $\mathbf{X}_{\mathcal{P}}$, $\gamma > 0$;

initialize $(\alpha_{i,j})$;

repeat

  $\beta_j \leftarrow \sqrt{\frac{1}{\gamma}\sum_{i=1}^{M}\alpha_{i,j}^2}$ for $j = 1, \ldots, N$;

  $\boldsymbol{\alpha}_i \leftarrow (\mathrm{diag}(\boldsymbol{\beta})^{-1} + \mathbf{X}_{\mathcal{C}}\mathbf{X}_{\mathcal{C}}^{\top})^{-1}\mathbf{X}_{\mathcal{C}}\mathbf{x}_i$, for $i = 1, \ldots, M$;

until converge;

$\mathbf{X}_{\mathcal{A}} \leftarrow \{\mathbf{x}_j | \mathbf{x}_j \in \mathbf{X}_{\mathcal{C}}, \beta_j \neq 0\}$;

return $\mathbf{X}_{\mathcal{A}}$

Algorithm 1 MRED for Active Learning

Input: The n unlabeled data samples X, the number of the selected most information data samples l, the number of the nearest neighbor data samples k

Step 1: Construct a nearest neighbor Laplacian graph with the weight matrix W as calculated in Eq. (15) on the unlabeled samples X and calculate

Step 2: Construct the kernel Gram matrix K with an selected input kernel type and let $M = L$.

Step 3: Construct the data-dependent deformed kernel Gram matrix $\tilde{K}$ according to Eq. (14).

Step 4: Let u_i be the ith column vector of K and initialize $\alpha_{i,j} = 1$.

  Step 4.1: Repeat

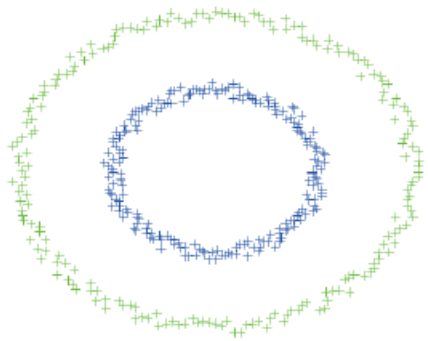
    Step 4.2: Compute β_j according to Eq. (29), i.e., $\beta_j = \sqrt{\sum_{i=1}^{n}\alpha_{i,j}^2/\lambda}$.

    Step 4.3: Compute α_{i} according to Eq. (27), i.e., $\alpha_{\mathrm{i}} = (D_{\beta}^{-1} + \tilde{K})^{-1}\tilde{K}_i$.

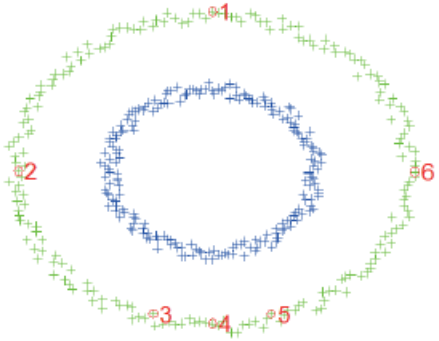
  Step 4.4: Until Convergence

Step 5: Rank the samples in X by following $\beta_j (j = 1, \ldots, n)$ in a descending order and then return the top l samples as the selected most informative ones Z.

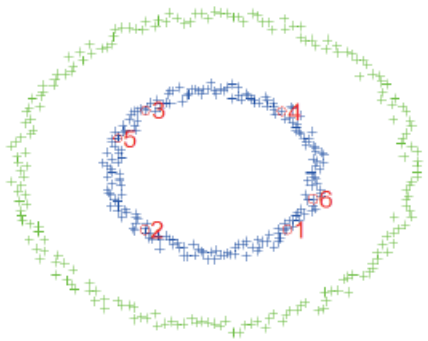
Output The l selected most informative samples can be labeled as the training samples.



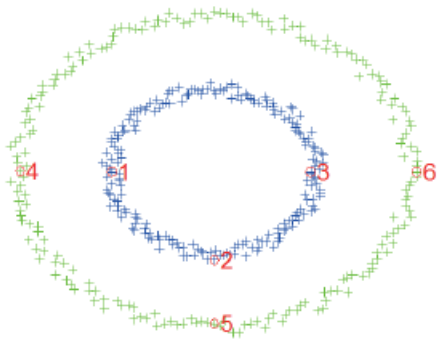
(a)


1
2
3
4
5
6

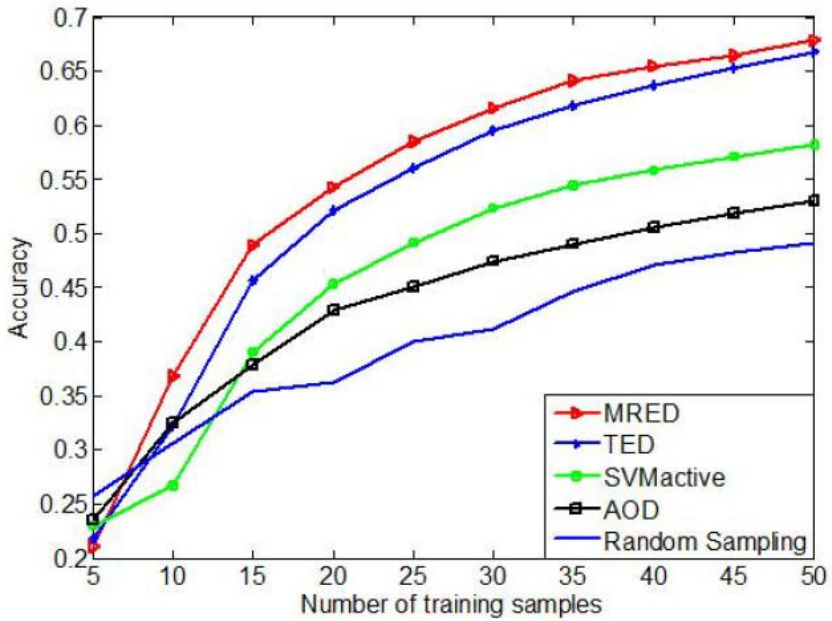
(b)


3
1
5
6
2
1

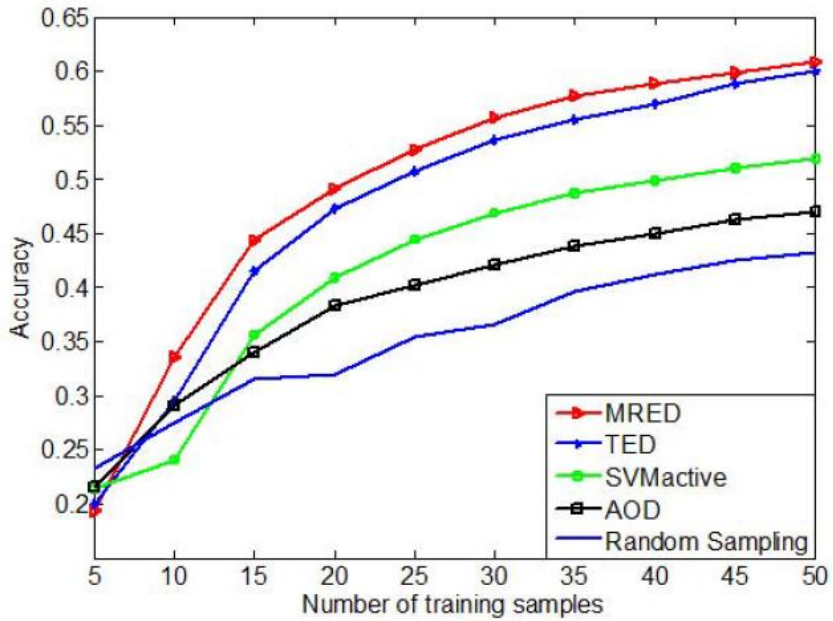
(c)


4
1
3
6
2
5

(d)


Accuracy
Number of training samples
MRED
TED
SVMactive
AOD
Random Sampling

(a)


Accuracy
Number of training samples
MRED
TED
SVMactive
AOD
Random Sampling

(b)

Semi-Supervised Learning with Graphs

PHD 2005

Xiaojin Zhu

Problem Setup

labeled data:$\{(x_1, y_1) \dots (x_l, y_l)\}$, $y \in \{1 \dots C\}$ unlabeled data: $\{(x_1, y_1) \dots (x_{l+u}, y_{l+u})\}$ $n = l + u$

Assume the number of classes C is known, and all classes are present in the labeled data. In most of the thesis we study the ***transductive*** problem of finding the labels for U.

Assume the graph is fully connected with the following weights

$$w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\alpha^2}\right)$$

The Algorithm

Propagate the labels through the edges. Larger edge weights allow labels to travel through more easily.

probabilistic transition matrix $P_{ij} = P(i \to j) = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}}$

Also define a $l \times C$ label matrix Y_L, whose i_{th} row is an indicator vector for $y_i, i \in L: Y_{ic} = \delta(y_i, c)$.We will compute soft labels f for the nodes. f is a $n \times C$ matrix, the rows can be interpreted as the probability distributions over labels.

1. Propagate $f \leftarrow Pf$

2. Clamp the labeled data $f_L = Y_L$

3. Repeat from step 1 until f converges.

Convergence

$$f = \begin{pmatrix} f_L \\ f_U \end{pmatrix} \qquad P = \begin{bmatrix} P_{LL} & P_{LU} \\ P_{UL} & P_{UU} \end{bmatrix} \qquad P_{ij} = P(i \to j) = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}}$$

$$f_U \leftarrow P_{UU} f_U + P_{UL} Y_L$$

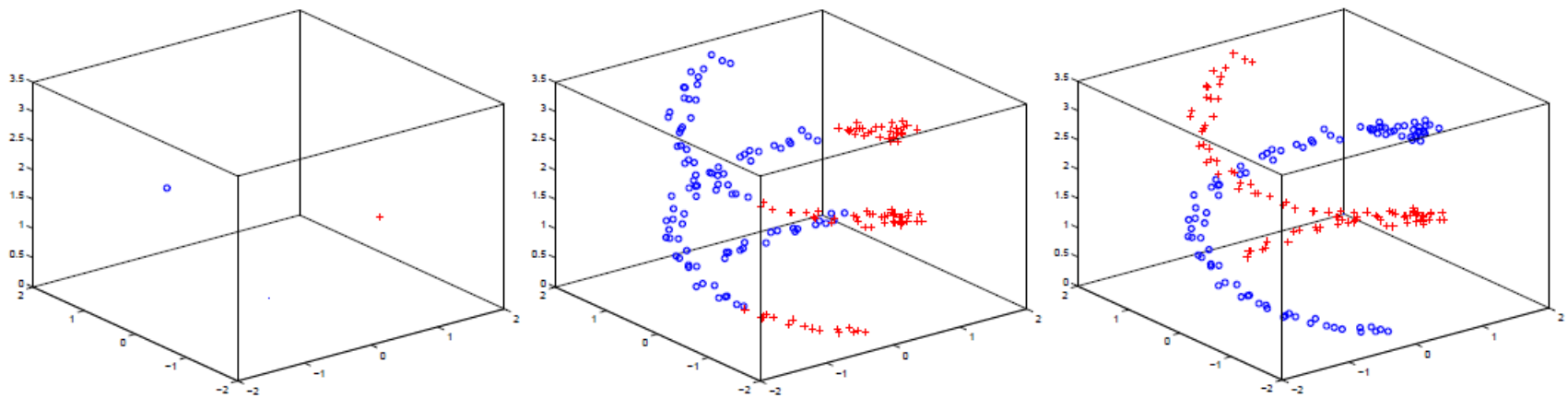
$$f_U = \lim_{n \to \infty} (P_{UU})^n f_U^0 + \left(\sum_{i=1}^{n} (P_{UU})^{(i-1)} \right) P_{UL} Y_L$$

Since P is row normalized, and P_{UU} is a sub-matrix of P, it follows

$$\exists \gamma < 1, \sum_{j=1}^{u} (P_{UU})_{ij} \leq \gamma, \forall i = 1 \ldots u$$

$$\begin{aligned}\sum_j (P_{UU})^n{}_{ij} &= \sum_j \sum_k (P_{UU})^{(n-1)}{}_{ik} (P_{UU})_{kj} \\ &= \sum_k (P_{UU})^{(n-1)}{}_{ik} \sum_j (P_{UU})_{kj} \qquad \exists \gamma < 1, \sum_{j=1}^{u} (P_{UU})_{ij} \leq \gamma, \forall i = 1 \ldots u \\ &\leq \sum_k (P_{UU})^{(n-1)}{}_{ik} \gamma \\ &\leq \gamma^n\end{aligned}$$

$$(P_{UU})^n f_U^0 \rightarrow 0 \qquad f_U = (I - P_{UU})^{-1} P_{UL} Y_L$$



(a) The data (b) 1NN (c) Label Propagation

Formalize label propagation with **a probabilistic framework**, assume binary classification $y \in \{0,1\}$. W has to be **symmetric with non-negative entries**, but otherwise need not to be positive semidefinite. Intuitively W specifies the 'local similarity' between points. The task is to assign labels to unlabeled nodes

Gaussian Random Fields

The strategy is to define a continuous random field on the graph. Intuitively, we want unlabeled points that are similar (as determined by edge weights) to have similar labels.

$$E(f) = \frac{1}{2}\sum_{i,j} w_{ij}\,(f(i) - f(j))^2 \qquad f(i) = y_i, i \in L$$

Assign a probability distribution to functions f by a Gaussian random field

$$p(f) = \frac{1}{Z}e^{-\beta E(f)} \qquad Z = \int_{f_L = Y_L} \exp\left(-\beta E(f)\right)\, df$$

Because of the quadratic energy, $p(f)$ $p(f_U|Y_L)$ are both multivariate Gaussian distributions

$$p(f) = \frac{1}{Z} e^{-\beta E(f)} \qquad Z = \int_{f_L = Y_L} \exp\left(-\beta E(f)\right) \, df$$

This is why p is called a Gaussian random field. The marginals $p(f_i|Y_L)$ are univariate Gaussian too, and have closed form solutions

The Graph Laplacian

$$D_{ii} = \sum_j W_{ij} \qquad \Delta \equiv D - W$$

$$E(f) = \frac{1}{2} \sum_{i,j} w_{ij} \left(f(i) - f(j)\right)^2 = f^\top \Delta f$$

The Gaussian random field can be written as

$$p(f) = \frac{1}{Z} e^{-\beta f^\top \Delta f}$$

Δ plays the role of the precision (inverse covariance) matrix in a multivariate Gaussian distribution. It is always positive semi-definite if W is symmetric and non-negative

$$p(f) = \frac{1}{Z} e^{-\beta f^\top \Delta f}$$

Harmonic Functions

$$\nabla^2 f = 0, \Delta f = 0 \qquad \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_1^2} = 0$$

$$f = \arg\min_{f_L = Y_L} E(f) \xrightarrow{\text{harmonic}} \begin{matrix} Y_L \\ \Delta f = 0 \end{matrix}$$

Use h to represent this harmonic function. The harmonic property means that the value of $h(i)$ at each unlabeled data point i is the average of its neighbors in the graph

$$h(i) = \frac{1}{D_{ii}} \sum_{j \sim i} w_{ij} h(j), \text{ for } i \in U$$

$$W = \begin{bmatrix} W_{LL} & W_{LU} \\ W_{UL} & W_{UU} \end{bmatrix}$$

$$\Delta h = 0$$

$$h_L = Y_L$$

$$\begin{aligned} h_U &= (D_{UU} - W_{UU})^{-1} W_{UL} Y_L \\ &= -(\Delta_{UU})^{-1} \Delta_{UL} Y_L \\ &= (I - P_{UU})^{-1} P_{UL} Y_L \end{aligned}$$

$$P = D^{-1} W$$

$$f_U = (I - P_{UU})^{-1} P_{UL} Y_L$$

Chapter 5 Active Learning(略)

Active learning with semi-supervised learning for open-set recognition

Probability Models for Open Set Recognition PAMI 2014

$$y = \mathrm{sgn}\left(\frac{1}{n_+} \sum_{\{i:y_i=+1\}} \underbrace{\langle \Phi(x), \Phi(x_i) \rangle}_{k(x,x_i)} - \frac{1}{n_-} \sum_{\{i:y_i=-1\}} \underbrace{\langle \Phi(x), \Phi(x_i) \rangle}_{k(x,x_i)} + b \right)$$

$$b = \frac{1}{2}\left(\frac{1}{n_-^2} \sum_{\{(i,j):y_i=y_j=-1\}} k(x_i, x_j) - \frac{1}{n_+^2} \sum_{\{(i,j):y_i=y_j=+1\}} k(x_i, x_j) \right)$$

Weibull $P^+(y|x) = 1 - P^-(\mathcal{Y} \backslash \mathrm{y}|x)$ reverse Weibull

$$P_\eta(y|f(x)) = 1 - e^{-\left(\frac{-f(x)-v_\eta}{\lambda_\eta}\right)^{\kappa_\eta}}$$

$$P_\psi(y|f(x)) = 1 - e^{-\left(\frac{-f(x)-v_\psi}{\lambda_\psi}\right)^{\kappa_\psi}}$$

$$y^* = \arg\max_{y \in \mathcal{Y}} P_{\eta,y}(x) \times P_{\psi,y}(x) \times l_y \qquad \text{subject to} \qquad P_{\eta,y^*}(x) \times P_{\psi,y^*}(x) \geq \delta_R$$

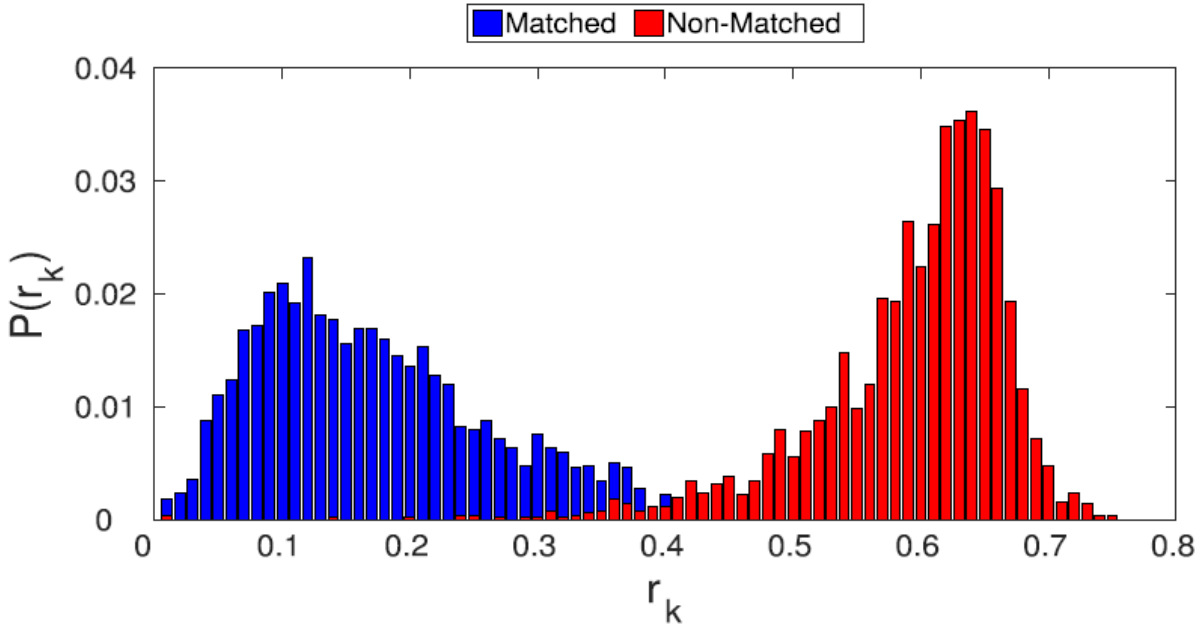
$$l_y = \begin{cases} 1 & \text{if } P_O(y|x) > \delta_\tau \\ 0 & \text{otherwise} \end{cases}$$

Sparse Representation-Based Open Set Recognition PAMI 2017

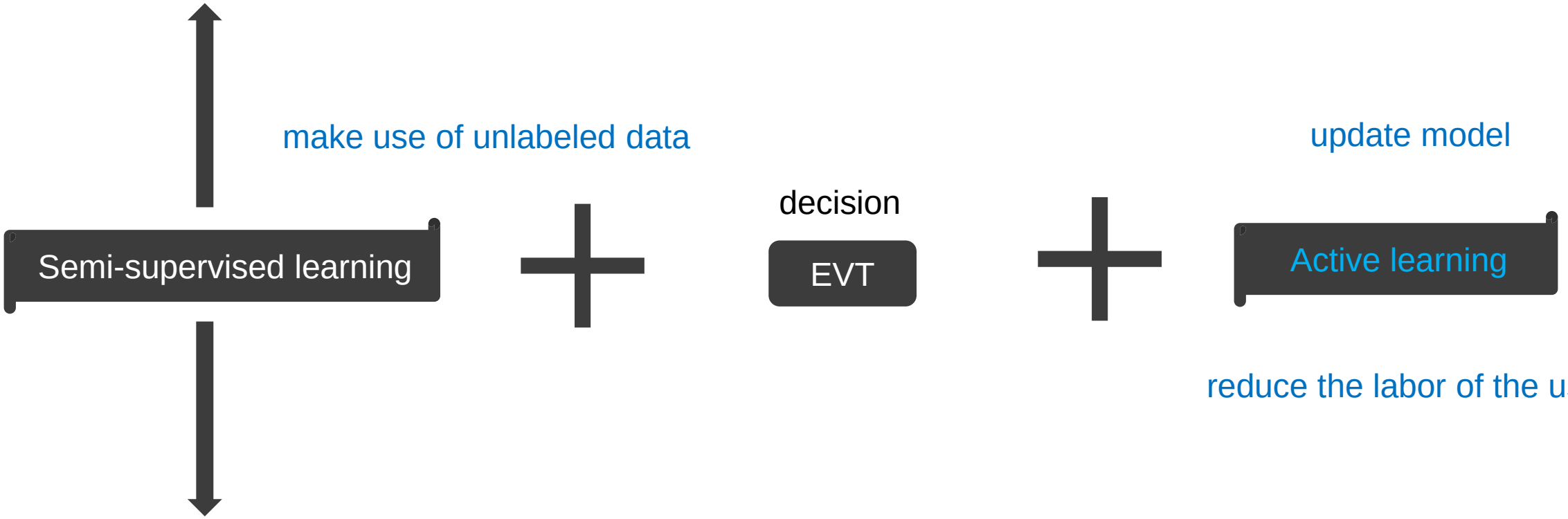
$$y_t = Yx \qquad x \in R^N$$

Use the GPD to model the tail of the matched distribution

$$\mathcal{H}_0: \ G(r_k) \leq \delta_g$$
$$\mathcal{H}_1: \ G(r_k) > \delta_g$$



Label Propagation $f_U = (I - P_{UU})^{-1} P_{UL} Y_L$



Manifold Regularization $f^* = \arg\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A ||f||_K^2 + \frac{\gamma_I}{(u+l)^2} \sum_{i,j=1}^{l+u} (f(x_i) - f(x_j))^2 W_{ij}$