Distribution Output Learning

The Distribution-valued RKHS Approach

Krikamol Muandet
Empirical Inference Department
Max Planck Institute for Intelligent Systems

January 29, 2013

Distribution Output Learning

To goal of inference is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ from a sample

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}.$$

from some unknown distribution $P(X, Y)$.

Distribution Output Learning

To goal of inference is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ from a sample

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}.$$

from some unknown distribution $P(X, Y)$.

Standard Supervised Learning

$f : \mathcal{X} \to \{-1, +1\}$ or $\mathbb{R}$

Distribution Output Learning

To goal of inference is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ from a sample

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}.$$

from some unknown distribution $P(X, Y)$.

Standard Supervised Learning

$f : \mathcal{X} \to \{-1, +1\}$ or $\mathbb{R}$

Vector-valued Learning

$f : \mathcal{X} \to \mathbb{R}^d$

Distribution Output Learning

To goal of inference is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ from a sample

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}.$$

from some unknown distribution $P(X, Y)$.

Standard Supervised Learning

$$f : \mathcal{X} \to \{-1, +1\} \text{ or } \mathbb{R}$$

Structured Output Learning

$$f : \mathcal{X} \to \mathcal{Y}, \quad \mathcal{Y} \neq \mathbb{R}^d$$

Vector-valued Learning

$$f : \mathcal{X} \to \mathbb{R}^d$$

Distribution Output Learning

To goal of inference is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ from a sample

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}.$$

from some unknown distribution $P(X, Y)$.

Standard Supervised Learning

$f : \mathcal{X} \to \{-1, +1\}$ or $\mathbb{R}$

Structured Output Learning

$f : \mathcal{X} \to \mathcal{Y}, \quad \mathcal{Y} \neq \mathbb{R}^d$

Vector-valued Learning

$f : \mathcal{X} \to \mathbb{R}^d$

Distribution Output Learning

$f : \mathcal{X} \to \mathcal{P}(\mathcal{Y}), \quad P(Y) \in \mathcal{P}(\mathcal{Y})$

Collaborative Filtering

Preference

- In psychology, preferences could be conceived of as an individual's attitude towards a set of objects.
- Evaluative judgment in the sense of liking or disliking an object.
- Preference is not necessarily stable over time.

Assumption

The preference distribution $\mathbb{P}$ is defined as the distribution over the set of objects $\mathcal{Y}$. The observations $y_1, y_2, \ldots, y_n$ are assumed to be i.i.d. realizations from the preference distribution.

Preference Prediction

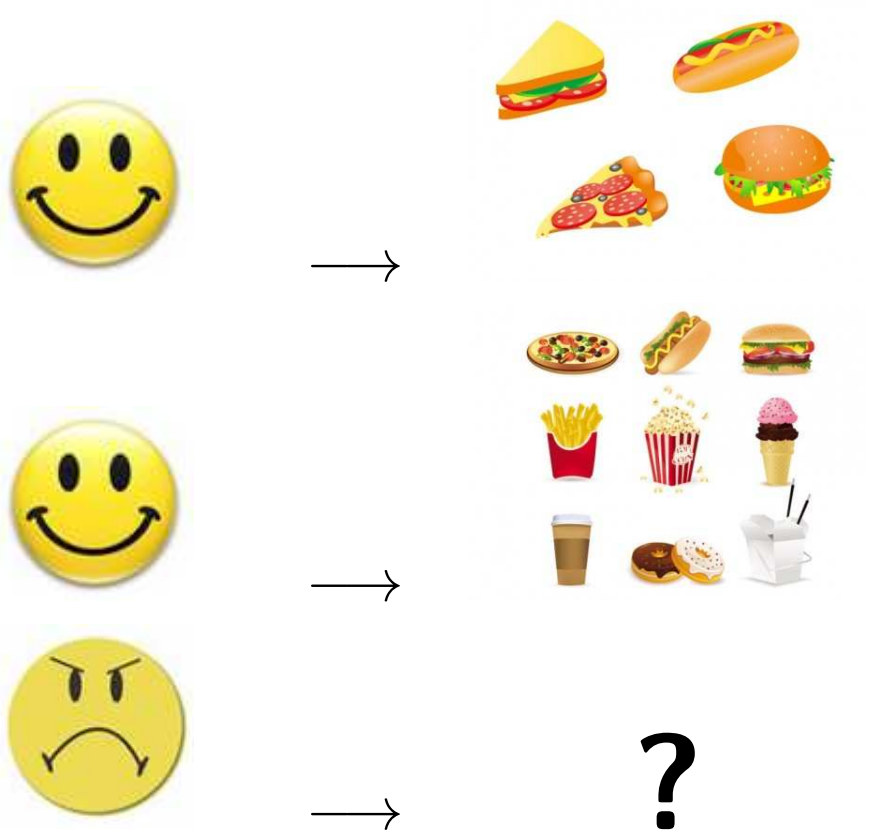
Customers **Products**

Preference Prediction

Customers **Products**

?

Preference Prediction



Cold-start Problem

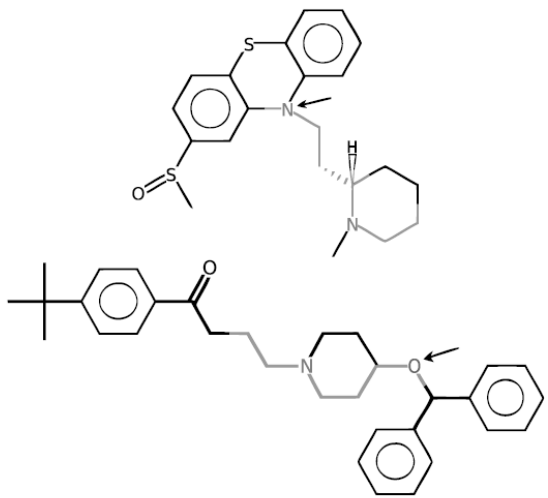
It is very difficult to make accurate predictions in the absence of rating data.

Structured Multi-Class Classification

In multi-class classification problem, one would like to predict the class to which the object belongs. That is, we want to learn a mapping

$$\mathcal{X} \longrightarrow \mathcal{P}(\mathcal{Y}), \qquad X \longmapsto P(Y|X = \cdot)$$

The conditional distribution $P(Y|X = \cdot)$ may be difficult to estimate for highly structured output.



Domain Prediction

The goal of domain adaptation is to learn a function f that perform well in a test domain. One approach is to **reweight** the training data so that training domain $\mathbb{P}_{train}(X)$ and test domain $\mathbb{P}_{test}(X)$ become similar:

$$\arg\min_{\beta} \quad \mathbb{D}\left(\mathbb{P}_{train}(X), \mathbb{P}_{test}(X)\right)$$

where $\mathbb{D}$ is a divergence between distributions. For domain prediction problem, we want to solve

$$\arg\min_{\beta} \quad \mathbb{D}\left(\mathbb{P}_{train}(X), f(u)\right)$$

where $f : \mathcal{U} \to \mathcal{P}(\mathcal{X})$.

Notations

$\mathcal{X}$	a structured input space, e.g., $\mathbb{R}^d$
$\mathcal{Y}$	s structured output space, e.g., non-Euclidean space
$\mathbb{P}_Y$	a probability distribution over the output space $(\mathcal{Y}, \mathcal{A})$
$\mathfrak{P}_\mathcal{Y}$	a set of all probability distribution $\mathbb{P}_Y$
$\mathcal{H}_X$	a reproducing kernel Hilbert space of function from $\mathcal{X}$ to $\mathbb{R}$
$\mathcal{H}_Y$	a reproducing kernel Hilbert space of function from $\mathcal{Y}$ to $\mathbb{R}$
k_x	the reproducing kernel of $\mathcal{H}_X$
k_y	the reproducing kernel of $\mathcal{H}_Y$

Distribution-valued Functions

- Learning a function $f : \mathcal{X} \to \mathfrak{P}_{\mathcal{Y}}$ from a training sample $\mathcal{S} = \{(x_1, \mathbb{P}_1), \ldots, (x_n, \mathbb{P}_n)\} \in \mathcal{X} \times \mathfrak{P}_{\mathcal{Y}}$.
- Consider the following mean embedding

$$\mu : \mathfrak{P}_{\mathcal{Y}} \longrightarrow \mathcal{H}_Y, \ \mathbb{P}_Y \longmapsto \int_{\mathcal{Y}} k_y(y, \cdot) \, \mathrm{d}\mathbb{P}_Y(y)$$

- Thus, we can consider a function $g : \mathcal{X} \to \mathcal{H}_Y$.
- Let $\mathcal{L}(\mathcal{H}_Y)$ be the set of all bounded operators from $\mathcal{H}_Y$ to $\mathcal{H}_Y$.

Distribution-valued Functions

Definition (Non-negative $\mathcal{L}(\mathcal{H}_Y)$-valued kernel)

A non-negative $\mathcal{L}(\mathcal{H}_Y)$-valued kernel K is an operator-valued function on $\mathcal{X} \times \mathcal{X}$, i.e., $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H}_Y)$, such that for all $x_i, x_j \in \mathcal{X}, \mu_i, \mu_j \in \mathcal{H}_Y$, and $m \in \mathbb{N}_+^*$

i) $K(x_i, x_j) = K(x_j, x_i)^*$ where $*$ denotes the adjoint,

ii) $\sum_{i,j=1}^{m} \langle K(x_i, x_j)\mu_i, \mu_j \rangle_{\mathcal{H}_Y} \geq 0$.

Distribution-valued RKHS

Given a non-negative $\mathcal{L}(\mathcal{H}_Y)$-valued kernel K on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS of $\mathcal{H}_Y$-valued functions defined over $\mathcal{X} \times \mathcal{X}$ whose reproducing kernel is K.

Definition ($\mathcal{H}_Y$-valued RKHS)

A RKHS $\mathcal{F}$ of $\mathcal{H}_Y$-valued functions $g : \mathcal{X} \to \mathcal{H}_Y$ is a Hilbert space such that there is a non-negative $\mathcal{L}(\mathcal{H}_Y)$-valued kernel K with the following properties:

i) $\forall x \in \mathcal{X}, \forall \mu \in \mathcal{H}_Y, \quad K(x, \cdot)\mu \in \mathcal{F},$

ii) $\forall g \in \mathcal{F}, \forall x \in \mathcal{X}, \forall \mu \in \mathcal{H}_Y$

$$\langle g, K(x, \cdot)\mu \rangle_{\mathcal{F}} = \langle g(x), \mu \rangle_{\mathcal{H}_Y}$$

Regularization on Distributions

Given a training sample $(X_1, \mathbb{P}_1), \ldots, (X_n, \mathbb{P}_n)$, we minimize the regularized loss functional

$$\ell(X_i, \mathbb{E}_{\mathbb{P}_i}[k_y(y, \cdot)], f(X_i)\}_{i=1}^n) + \lambda\Omega(\|f\|_{\mathcal{F}})$$

where $f \in \mathcal{F}$.

Kernel Ridge Regression:

$$\frac{1}{n}\sum_{i=1}^{n} \|f(X_i) - \mu[\mathbb{P}_i]\|_{\mathcal{H}_Y}^2 + \lambda\|f\|_{\mathcal{F}}^2$$

Distribution Estimation

We solve the following kernel ridge regression problem:

$$\arg\min_{g \in \mathcal{F}} \sum_{i=1}^{n} \|g(X_i) - \mu[\mathbb{P}_i]\|^2_{\mathcal{H}_Y} + \lambda \|g\|^2$$

where $\lambda > 0$ is a regularization parameter. Using the representer theorem, the solution can be written as

$$g(\cdot) = \sum_{i=1}^{n} K(\cdot, X_i)\Psi_i$$

where $\Psi \in \mathcal{H}$. Substituting back, we obtain an analytic solution

$$\mathbf{\Psi} = (\mathbf{K} + \lambda I)^{-1}\mathbf{\Phi}_I$$

where $\mathbf{\Phi}_I$ is the column vector of $[\mu[\mathbb{P}_i] \in \mathcal{H}_\mathcal{Y}]_{i=1}^n$.

Distribution Prediction

Given a set of objects $Y_1, Y_2, \ldots, Y_m$, we want to know the relative weights $\beta_1, \beta_2, \ldots, \beta_m$ that encodes the probability over set. That is, the deviation between the predicted distribution $g(X)$ and the empirical distribution $\mu[\widehat{\mathbb{P}}] = \frac{1}{m}\sum_{i=1}^{m} \beta_i k_y(Y_i, \cdot)$ is small.

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^m} \left\| \frac{1}{m} \sum_{i=1}^{m} \beta_i k_y(Y_i, \cdot) - g(X) \right\|_{\mathcal{H}_\mathcal{Y}}^2 \qquad \text{s.t. } \boldsymbol{\beta}^\top \mathbf{1} = 1, \beta_i \geq 0$$

Distribution Prediction

To prevent overfitting, we introduce a regularizer $\Omega(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2$ with a regularization parameter $\varepsilon > 0$. Substituting

$$g(X) = \sum_{i=1}^{n} K(X, X_i)\Psi_i$$

back yields a quadratic programming (QP) for $\boldsymbol{\beta}$

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \frac{1}{2}\boldsymbol{\beta}^\top(\mathbf{K} + \varepsilon\mathbf{I})\boldsymbol{\beta} - \mathbf{Q}^\top\boldsymbol{\beta} \quad \text{s.t. } \boldsymbol{\beta}^\top\mathbf{1} = 1, \beta_i \geq 0$$

where $\mathbf{I}$ is the identity matric, $\mathbf{K} \in \mathbb{R}^{m\times m}$ and $\mathbf{Q} \in \mathbb{R}^m$ are given by

$$\mathbf{K}_{ij} = k_y(Y_i, Y_j), \qquad \mathbf{Q}_j = \mathbf{K}_x(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{L}_j$$

Summary

- a learning framework which can predict a distribution.
- a distribution-valued RKHS approach
- an operator-valued kernel for distributions.
- it does not suffer from the *pre-image* problem compared to standard structured output learning.

"What does a fish say when it hits a wall?"