Kernel Distribution Propagation

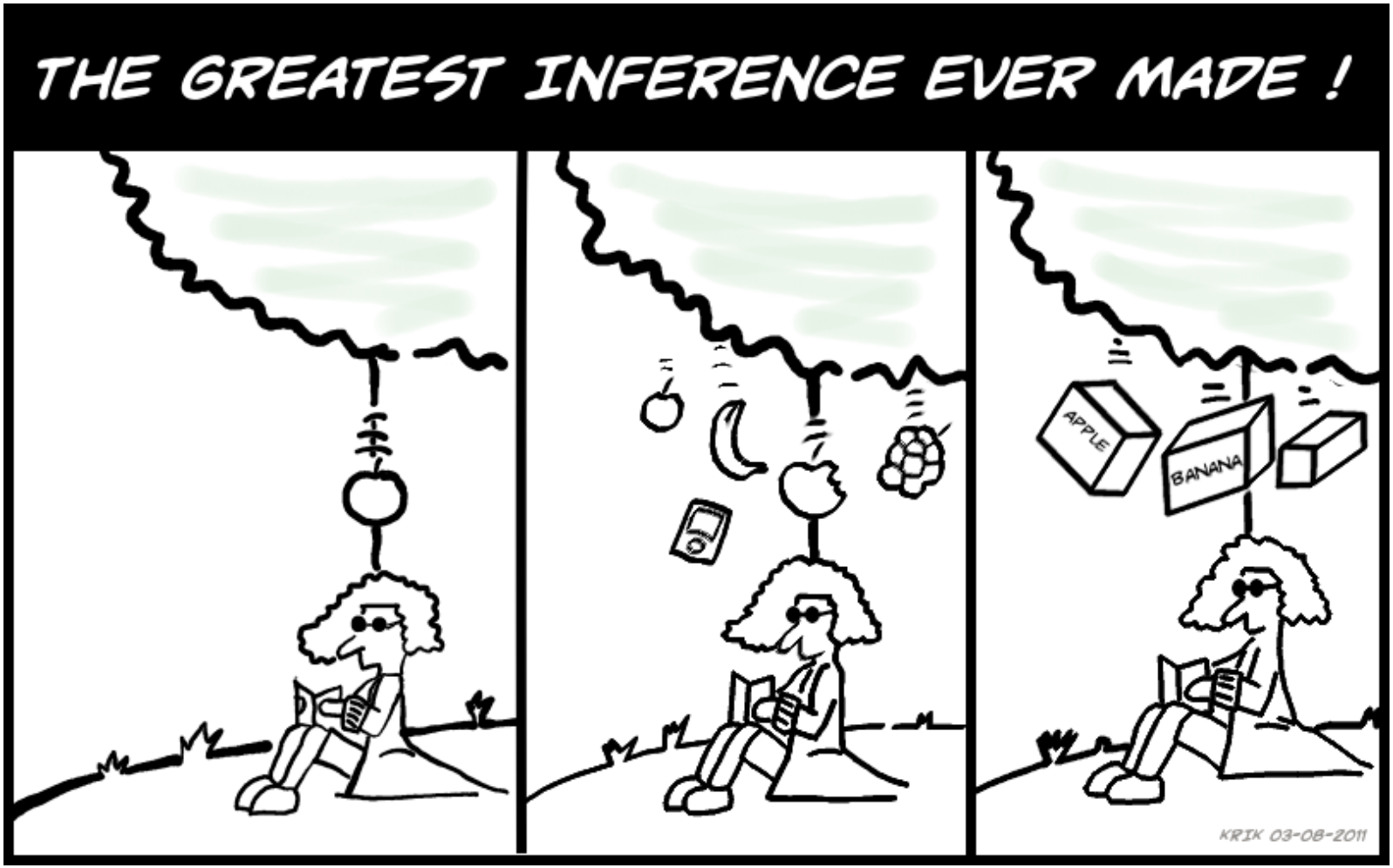
Krikamol Muandet

Max Planck Institute for Intelligent Systems

Tübingen, Germany

November 17, 2011

Recap


THE GREATEST INFERENCE EVER MADE !
APPLE
BANANA
KRIK 03-08-2011

Recap

Hilbert Space Embedding

A mean embedding of the distribution $\mathbb{P} \in \mathcal{Q}$ in the reproducing kernel Hilbert space $\mathcal{H}$ with kernel k is defined as

$$\begin{aligned} \mu &: \mathcal{Q} \longrightarrow \mathcal{H} \\ &: \mathbb{P} \longmapsto \int_{\mathcal{X}} k(x, \cdot)\, d\mathbb{P}(x) \triangleq \mu_{\mathbb{P}} \end{aligned}$$

For an appropriate choice of positive semidefinite kernel k, the map μ is injective. Therefore, all information about the distribution $\mathbb{P}$ are preserved.

For any $f \in \mathcal{H}$, we have

$$\langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} = \mathbb{E}_{x \sim \mathbb{P}}[f(x)]$$

Recap

Regularization on Distributions

Given training samples $(\mathbb{P}_i, y_i) \in \mathcal{Q} \times \mathbb{R}$, $i = 1, \ldots, m$, a strictly monotonically increasing function $\Omega : [0, +\infty) \to \mathbb{R}$, a loss function $\ell : (\mathcal{Q} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{+\infty\}$, then any $f \in \mathcal{H}$ minimizing the regularized risk functional

$$\ell\left(\mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{E}_{\mathbb{P}_m}[f]\right) + \Omega\left(\|f\|_{\mathcal{H}}\right) \tag{1}$$

admits a representation of the form

$$f = \sum_{i=1}^{m} \alpha_i \mu_{\mathbb{P}_i} \tag{2}$$

for some $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, m$.

Motivations

Distribution Propagation

Conditional Mean Embedding

Conditioning Maximum Mean Discrepancy

Kernel Distribution Propagation

Conclusions

Motivations

Distribution Propagation

Conditional Mean Embedding

Conditioning Maximum Mean Discrepancy

Kernel Distribution Propagation

Conclusions

Motivations

Objective: Distributions on output space $\mathcal{Y}$.

Motivations

Objective: Distributions on output space $\mathcal{Y}$.

Supervised learning with noisy and complex structures:

- The output space is highly structured and complex.
- It may be infeasible or expensive to obtain objective and reliable labels. Instead, we can collect subjective (possibly noisy) labels from multiple experts.
- Overfitting.

Motivations

Objective: Distributions on output space $\mathcal{Y}$.

Supervised learning with noisy and complex structures:

- The output space is highly structured and complex.
- It may be infeasible or expensive to obtain objective and reliable labels. Instead, we can collect subjective (possibly noisy) labels from multiple experts.
- Overfitting.

Applications

- semi-supervised learning
- structured output learning
- preference learning and collaborative filtering*
- etc.

Motivations

Options

1. Probabilistic models
   - allow an estimation of uncertainty in the output space.
   - cannot deal with structured input/output directly.
   - can suffer from model misspecification.
2. Kernel-based methods
   - can handle structured input/output directly.
   - ignore an uncertainty in the output space.

Motivations

Options

1. Probabilistic models
   - allow an estimation of uncertainty in the output space.
   - cannot deal with structured input/output directly.
   - can suffer from model misspecification.
2. Kernel-based methods
   - can handle structured input/output directly.
   - ignore an uncertainty in the output space.

Goal

The learning algorithm which

- can deal with structured input/output.
- can handle an uncertainty in the output space directly.
- makes no assumption about the underlying distribution, i.e., nonparametric.

Motivations

Distribution Propagation

Conditional Mean Embedding

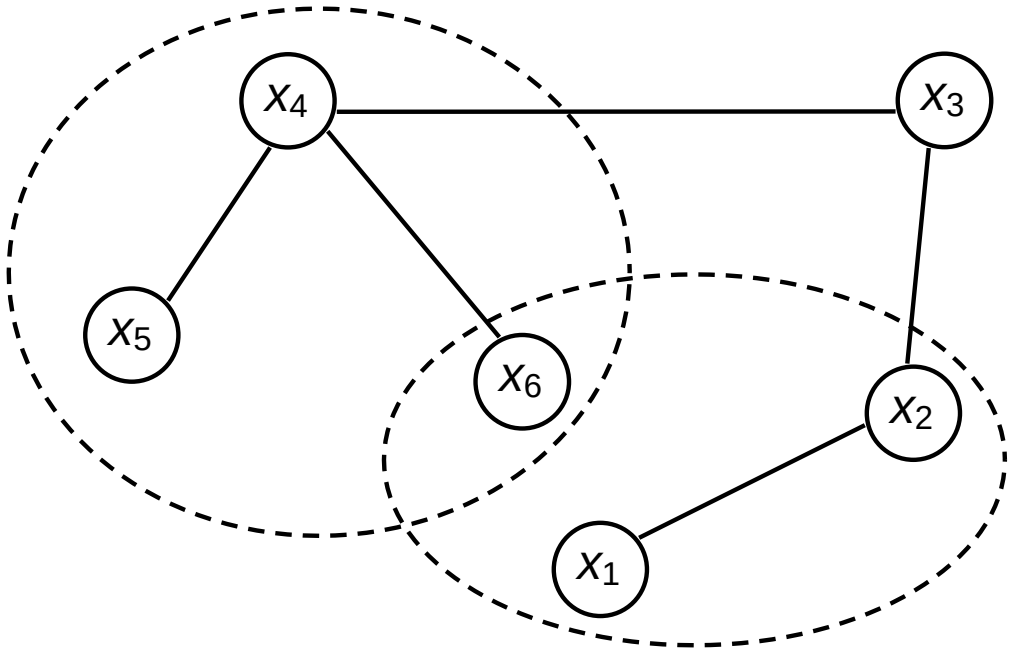
Conditioning Maximum Mean Discrepancy

Kernel Distribution Propagation

Conclusions

Hypergraphs

A hypergraph is a generalization of a graph wherein edges can connect more than two vertices and are called hyperedges.



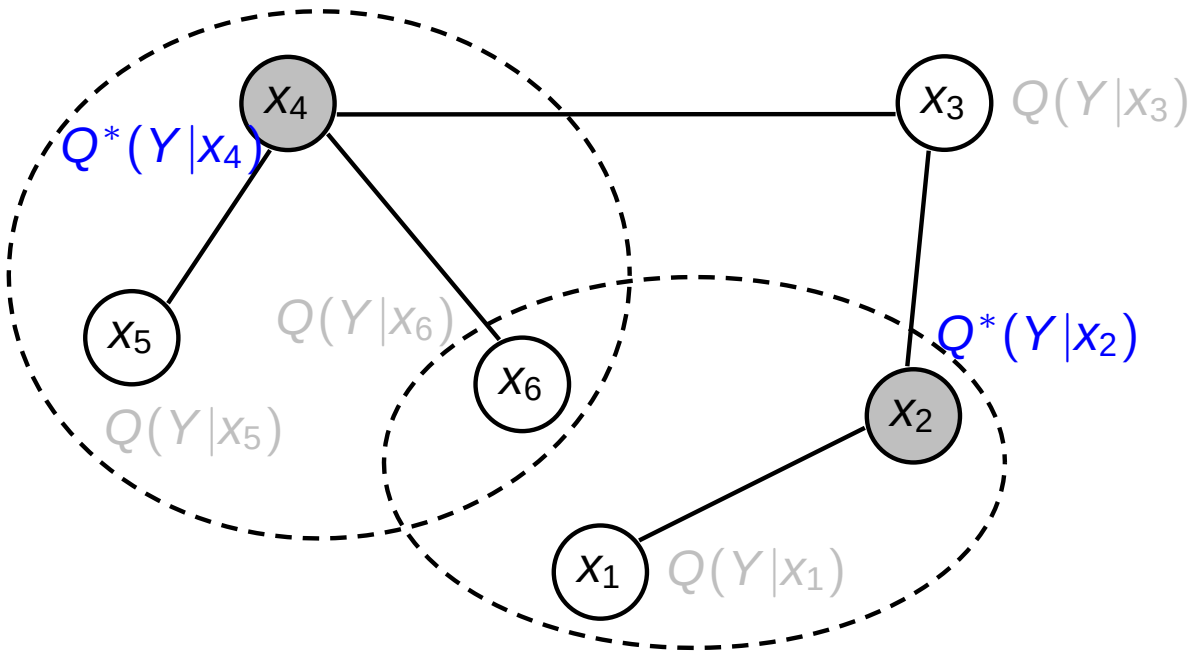
A hypergraph **G** is composed of a set of nodes $x_1, \ldots, x_l$ and a set of hyperedges $R_1, \ldots, R_m$. Furthermore, there are positive weights $\beta_1, \ldots, \beta_m$ associated with the hyperedges.

(Corduneanu and Jaakkola, 2004; Tsuda, 2005)

Distribution Propagation

Distribution propagation on hypergraphs

Given a hypergraph **G** and conditional distributions $Q(Y|x),\ \forall x \in S_I$, estimate the conditional distributions $Q(Y|x'),\ \forall x' \in S$.

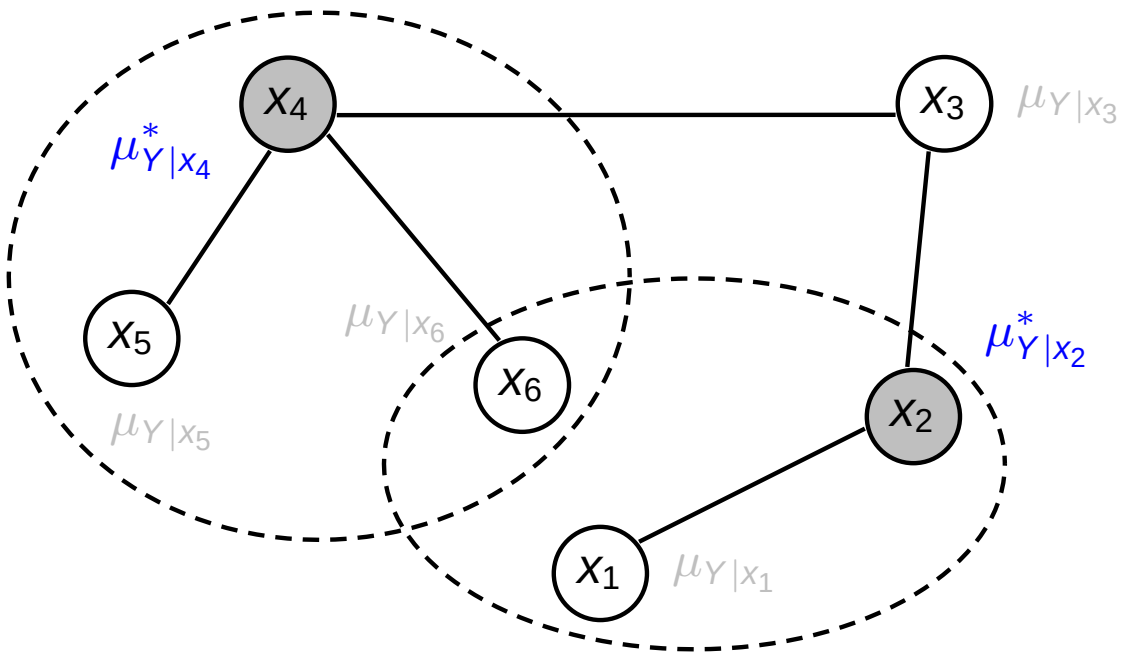


(Corduneanu and Jaakkola, 2004; Tsuda, 2005)

Distribution Propagation

Distribution propagation on hypergraphs

Given a hypergraph **G** and conditional distributions $Q(Y|x),\ \forall x \in S_l$, estimate the conditional distributions $Q(Y|x'),\ \forall x' \in S$.



Motivations

Distribution Propagation

Conditional Mean Embedding

Conditioning Maximum Mean Discrepancy

Kernel Distribution Propagation

Conclusions

Conditional Mean Embedding

Cross-covariance operator

For reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ and $\mathcal{G}$, the cross-covariance operator $\mathcal{C}_{XY} : \mathcal{G} \to \mathcal{F}$ is defined as

$$\mathcal{C}_{XY} = \mathbb{E}_{XY}[\varphi(X) \otimes \phi(Y)] - \mu_X \otimes \mu_Y$$

where $\otimes$ is the tensor product. Given two functions, $f \in \mathcal{F}$ and $g \in \mathcal{G}$, their cross-covariance, $\mathrm{Cov}_{XY}[f(X), g(Y)] := \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)]$, is computed as

$$\langle f, \mathcal{C}_{XY} g \rangle_{\mathcal{F}} \text{ or equivalently } \langle f \otimes g, \mathcal{C}_{XY} \rangle_{\mathcal{F} \otimes \mathcal{G}}$$

(Baker, 1971; Fukumizu et al., 2004; Song et al., 2009)

Conditional Mean Embedding

Using a relation

$$\mathcal{C}_{XX}\mathbb{E}_{Y|X}[g(Y)|X] = \mathcal{C}_{XY}g \ ,$$

the conditional embedding $\mathcal{U}_{Y|X}$ distributions of the form $Q(Y|X)$ into an RKHS can be defined in terms of cross-covariance operators as

$$\mathcal{U}_{Y|X} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$$

Assuming that $\mathbb{E}_{Y|X}[g(Y)|X] \in \mathcal{F}$, the conditional embedding $\mathcal{U}_{Y|X}$ satisfies

$$\begin{aligned} \mu_{Y|x} &= \mathbb{E}_{Y|x}[\phi(Y)|x] = \mathcal{U}_{Y|X}k(x,\cdot) \\ \mathbb{E}_{Y|x}[g(Y)|x] &= \langle g, \mu_{Y|x}\rangle_{\mathcal{G}} \end{aligned}$$

(Baker, 1971; Fukumizu et al., 2004; Song et al., 2009)

Conditional Mean Embedding

Using a relation

$$\mathcal{C}_{XX}\mathbb{E}_{Y|X}[g(Y)|X] = \mathcal{C}_{XY}g \ ,$$

the conditional embedding $\mathcal{U}_{Y|X}$ distributions of the form $Q(Y|X)$ into an RKHS can be defined in terms of cross-covariance operators as

$$\mathcal{U}_{Y|X} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$$

Assuming that $\mathbb{E}_{Y|X}[g(Y)|X] \in \mathcal{F}$, the conditional embedding $\mathcal{U}_{Y|X}$ satisfies

$$\begin{aligned} \mu_{Y|x} &= \mathbb{E}_{Y|x}[\phi(Y)|x] = \mathcal{U}_{Y|X}k(x,\cdot) \\ \mathbb{E}_{Y|x}[g(Y)|x] &= \langle g, \mu_{Y|x}\rangle_{\mathcal{G}} \end{aligned}$$

Theorem

Let $\Upsilon = (\varphi(x_1), \ldots, \varphi(x_m))$, $\Phi = (\phi(y_1), \ldots, \phi(y_m))$, $\mathbf{K} = \Upsilon^{\mathsf{T}}\Upsilon$, $\mathbf{H} = \mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^{\mathsf{T}}$, and $k_x = \Upsilon^{\mathsf{T}}\varphi(x)$. Then $\hat{\mu}_{Y|x}$ can be estimated as:

$$\hat{\mu}_{Y|x} = \Phi(\mathbf{H}\mathbf{K} + \lambda m\mathbf{I})^{-1}\mathbf{H}k_x$$

(Baker, 1971; Fukumizu et al., 2004; Song et al., 2009)

Motivations

Distribution Propagation

Conditional Mean Embedding

Conditioning Maximum Mean Discrepancy

Kernel Distribution Propagation

Conclusions

Conditioning Maximum Mean Discrepancy

Let $\mathcal{F}$ and $\mathcal{G}$ be RKHSs with reproducing kernel k and l, respectively. Given samples $(x_1, y_1), \ldots, (x_m, y_m)$, the conditioning maximum mean discrepancy (cMMD) and its empirical estimate are defined as

$$\begin{aligned} cMMD^2(Q(Y|x), Q(Y|x')) &= \|\mu_{Y|x} - \mu_{Y|x'}\|^2_{\mathcal{G}} \\ \widehat{cMMD}^2(Q(Y|x), Q(Y|x')) &= k_x^\mathsf{T} \mathbf{C} k_x - 2k_x^\mathsf{T} \mathbf{C} k_{x'} + k_{x'}^\mathsf{T} \mathbf{C} k_{x'} \end{aligned}$$

where $\mathbf{C} = \mathbf{H}(\mathbf{HK} + \lambda m \mathbf{I})^{-1} \mathbf{L} (\mathbf{HK} + \lambda m \mathbf{I})^{-1} \mathbf{H}$ and $\mathbf{L} = \Phi^\mathsf{T} \Phi$.

In collaboration with Kun Zhang

Conditioning Maximum Mean Discrepancy

Let $\mathcal{F}$ and $\mathcal{G}$ be RKHSs with reproducing kernel k and l, respectively. Given samples $(x_1, y_1), \ldots, (x_m, y_m)$, the conditioning maximum mean discrepancy (cMMD) and its empirical estimate are defined as

$$\begin{aligned} cMMD^2(Q(Y|x), Q(Y|x')) &= \|\mu_{Y|x} - \mu_{Y|x'}\|^2_{\mathcal{G}} \\ \widehat{cMMD}^2(Q(Y|x), Q(Y|x')) &= k_x^\mathsf{T} \mathbf{C} k_x - 2k_x^\mathsf{T} \mathbf{C} k_{x'} + k_{x'}^\mathsf{T} \mathbf{C} k_{x'} \end{aligned}$$

where $\mathbf{C} = \mathbf{H}(\mathbf{HK} + \lambda m \mathbf{I})^{-1} \mathbf{L} (\mathbf{HK} + \lambda m \mathbf{I})^{-1} \mathbf{H}$ and $\mathbf{L} = \Phi^\mathsf{T} \Phi$.

Conditioning MMD $\neq$ Conditional MMD

Conditioning MMD The distance between $Q(Y|x)$ and $Q(Y|x')$.

Conditional MMD The distance between $P(Y|X)$ and $Q(Y|X)$.

In collaboration with Kun Zhang

Motivations

Distribution Propagation

Conditional Mean Embedding

Conditioning Maximum Mean Discrepancy

Kernel Distribution Propagation

Conclusions

Kernel Distribution Propagation

Kernel Distribution Propagation (KDP)

Given a hypergraph **G** and conditional distributions $Q(Y|x)$, $x \in S_l$, the KDP algorithm find $Q(Y|x')$, $x' \in S$ by minimizing

$$\ell(Q;\gamma) = \sum_{x \in \mathbf{L}} \|\mu^*_{Y|x} - \mu_{Y|x}\|^2_{\mathcal{G}} + \gamma \sum_{k=1}^{m} \beta_k \sum_{x \in R_k} \|\mu_{Y|x} - \mu_{R_k}\|^2_{\mathcal{G}}$$

where

$$\mu_{R_k} = \frac{1}{|R_k|} \sum_{x \in R_k} \mu_{Y|x}$$

and $\mu^*_{Y|x}$ are mean embeddings of $Q(Y|x)$ for $x \in S_l$.

Kernel Distribution Propagation

Kernel Distribution Propagation (KDP)

Given a hypergraph **G** and conditional distributions $Q(Y|x)$, $x \in S_l$, the KDP algorithm find $Q(Y|x')$, $x' \in S$ by minimizing

$$\ell(Q;\gamma) = \sum_{x \in \mathbf{L}} \|\mu^*_{Y|x} - \mu_{Y|x}\|^2_{\mathcal{G}} + \gamma \sum_{k=1}^{m} \beta_k \sum_{x \in R_k} \|\mu_{Y|x} - \mu_{R_k}\|^2_{\mathcal{G}}$$

where

$$\mu_{R_k} = \frac{1}{|R_k|} \sum_{x \in R_k} \mu_{Y|x}$$

and $\mu^*_{Y|x}$ are mean embeddings of $Q(Y|x)$ for $x \in S_l$.

Conjecture

$$\begin{aligned} \mu_{R_k} &= \frac{1}{|R_k|} \sum_{x \in R_k} \mu_{Y|x} = \frac{1}{|R_k|} \sum_{x \in R_k} \mathcal{C}_{YX} \mathcal{C}_{XX}^{-1} k(x,\cdot) \\ &= \mathcal{C}_{YX} \mathcal{C}_{XX}^{-1} \frac{1}{|R_k|} \sum_{x \in R_k} k(x,\cdot) = \mathcal{C}_{YX} \mathcal{C}_{XX}^{-1} \mathbf{m}_{R_k} \end{aligned}$$

Kernel Distribution Propagation

Unlabeled Nodes

Let $\mathbf{k}_x = \Upsilon^\mathsf{T} k(x, \cdot)$ and $\mathbf{k}_{R_k} = \Upsilon^\mathsf{T} \mathbf{m}_{R_k}$. For unlabeled nodes in the hypergraph, we have

$$\begin{aligned}
\hat{\ell}(Q;\gamma) &= \gamma \sum_{k=1}^{m} \beta_k \sum_{x \in R_k} \|\hat{\mu}_{Y|x} - \hat{\mu}_{R_k}\|_{\mathcal{G}}^2 \\
&= \gamma \sum_{k=1}^{m} \beta_k \sum_{x \in R_k} \|\Phi(\mathbf{HK} + \lambda m \mathbf{I})^{-1} \mathbf{Hk}_x - \Phi(\mathbf{HK} + \lambda m \mathbf{I})^{-1} \mathbf{Hk}_{R_k}\|_{\mathcal{G}}^2 \\
&= \gamma \sum_{k=1}^{m} \beta_k \sum_{x \in R_k} [\mathbf{h}_x^\mathsf{T} \mathbf{L} \mathbf{h}_x - 2\mathbf{h}_x^\mathsf{T} \mathbf{L} \mathbf{h}_{R_k} + \mathbf{h}_{R_k}^\mathsf{T} \mathbf{L} \mathbf{h}_{R_k}]
\end{aligned}$$

where $\mathbf{h}_x = (\mathbf{HK} + \lambda m \mathbf{I})^{-1} \mathbf{Hk}_x$ and $\mathbf{h}_{R_k} = (\mathbf{HK} + \lambda m \mathbf{I})^{-1} \mathbf{Hk}_{R_k}$.

Kernel Distribution Propagation

Unlabeled Nodes

Let $\mathbf{k}_x = \Upsilon^\top k(x, \cdot)$ and $\mathbf{k}_{R_k} = \Upsilon^\top \mathbf{m}_{R_k}$. For unlabeled nodes in the hypergraph, we have

$$\begin{aligned}
\hat{\ell}(Q;\gamma) &= \gamma \sum_{k=1}^{m} \beta_k \sum_{x \in R_k} \|\hat{\mu}_{Y|x} - \hat{\mu}_{R_k}\|_{\mathcal{G}}^2 \\
&= \gamma \sum_{k=1}^{m} \beta_k \sum_{x \in R_k} \|\Phi(\mathbf{HK} + \lambda m \mathbf{I})^{-1}\mathbf{Hk}_x - \Phi(\mathbf{HK} + \lambda m \mathbf{I})^{-1}\mathbf{Hk}_{R_k}\|_{\mathcal{G}}^2 \\
&= \gamma \sum_{k=1}^{m} \beta_k \sum_{x \in R_k} [\mathbf{h}_x^\top \mathbf{L}\mathbf{h}_x - 2\mathbf{h}_x^\top \mathbf{L}\mathbf{h}_{R_k} + \mathbf{h}_{R_k}^\top \mathbf{L}\mathbf{h}_{R_k}]
\end{aligned}$$

where $\mathbf{h}_x = (\mathbf{HK} + \lambda m \mathbf{I})^{-1}\mathbf{Hk}_x$ and $\mathbf{h}_{R_k} = (\mathbf{HK} + \lambda m \mathbf{I})^{-1}\mathbf{Hk}_{R_k}$.

Labeled Nodes

For labeled nodes in the hypergraph, we have

$$\hat{\ell}(Q;\gamma) = \underbrace{\sum_{x \in \mathbf{L}} \|\mu^*_{Y|x} - \hat{\mu}_{Y|x}\|_{\mathcal{G}}^2}_{?} + \gamma \sum_{k=1}^{m} \beta_k \sum_{x \in R_k} \|\hat{\mu}_{Y|x} - \hat{\mu}_{R_k}\|_{\mathcal{G}}^2$$

Motivations

Distribution Propagation

Conditional Mean Embedding

Conditioning Maximum Mean Discrepancy

Kernel Distribution Propagation

Conclusions

Conclusions

Kernel Distribution Propagation on hypergraphs using Hilbert space embedding:

- can deal with structured input/output.
- can handle an uncertainty in the output space directly.
- makes no assumption about the underlying distribution, i.e., nonparametric.

Acknowledgement

- Kun Zhang
- David Balduzzi

Questions?

References

Baker, Charles R. (1971). *Joint Measures and Cross-Covariance Operators*.

Corduneanu, Adrian and Tommi Jaakkola (2004). “Distributed Information Regularization on Graphs”. In: *NIPS*.

Fukumizu, Kenji et al. (2004). “Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces”. In: *Journal of Machine Learning Research* 5, p. 2004.

Song, Le et al. (2009). “Hilbert space embeddings of conditional distributions with applications to dynamical systems”. In: *ICML ’09: Proceedings of the 26th Annual International Conference on Machine Learning*. New York, NY, USA: ACM, pp. 961–968.

Tsuda, Koji (2005). “Propagating distributions on a hypergraph by dual information regularization”. In: *In Proc. 22th Intl. Conf. on Machine Learning*.