Hilbert Space Embeddings of Distributions

By 张凯歌

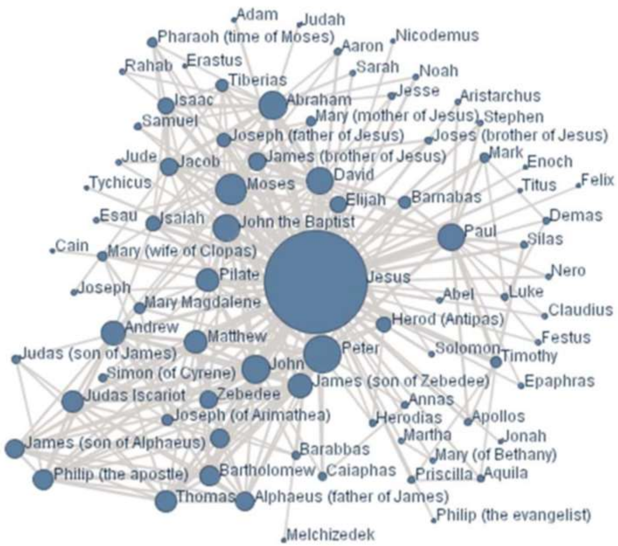
CONTENTS

/01

Introduction

Probabilistic Graphical Models
概率图模型

图



模型

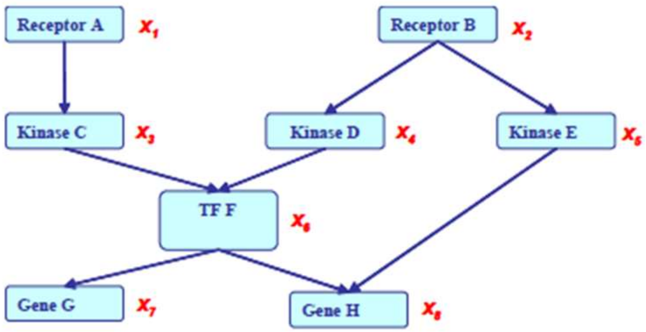
$$M$$

数据

$$D \equiv \{X_1^i, X_2^i, \ldots, X_m^i\}_{i=1}^N$$

Probabilistic Graphical Models

If X_i's are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,

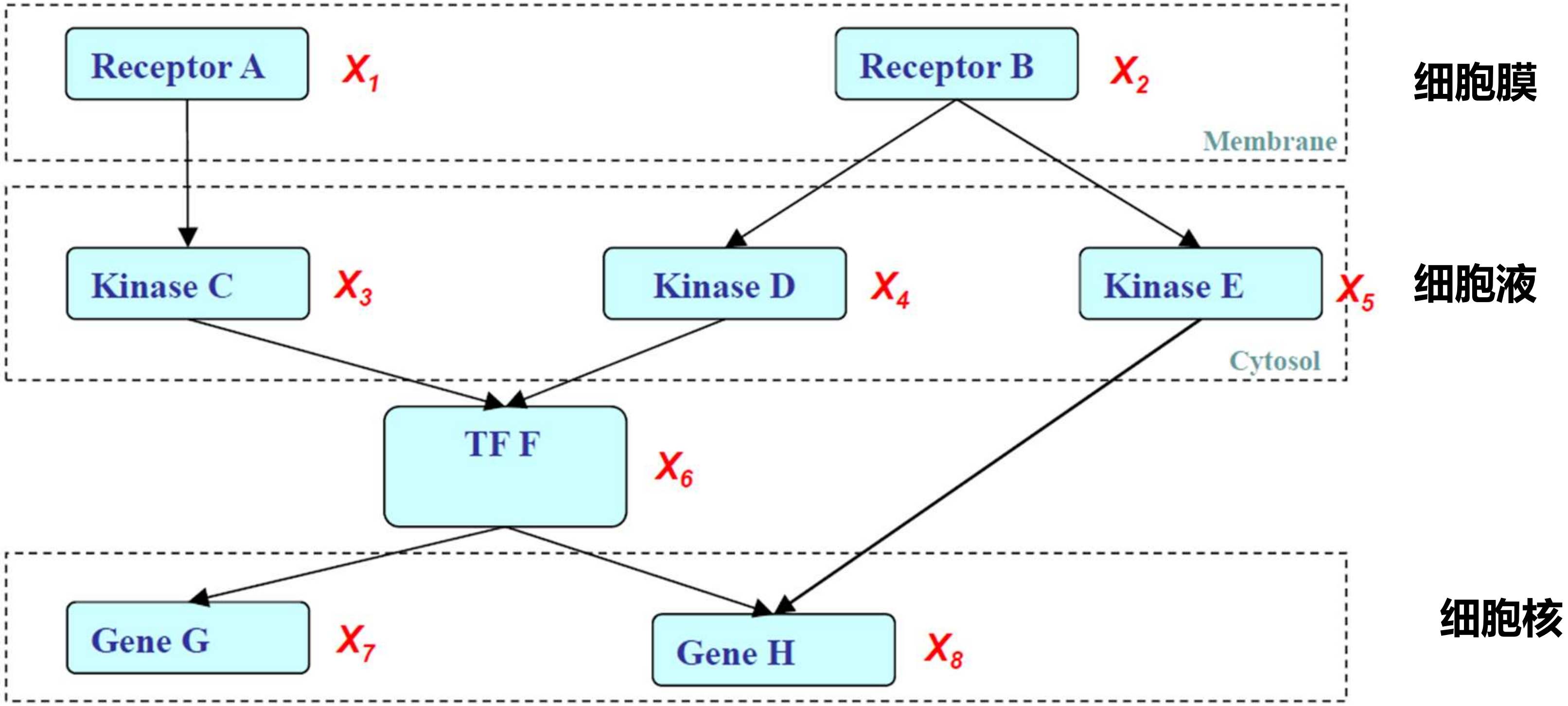


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)$$
$$P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

Why we may favor a PGM?

GM: Structure Simplifies Representation

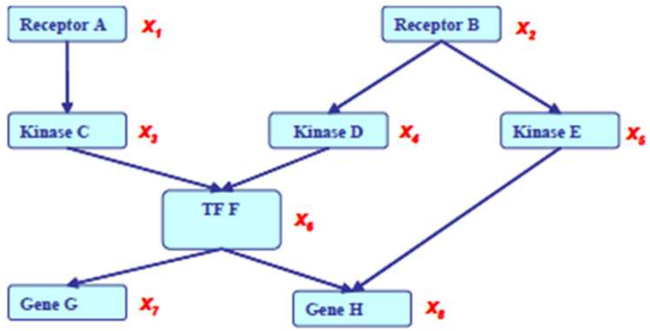
细胞内信号传输过程：变量之间的依赖关系



领域知识和因果(逻辑)结构的结合

Probabilistic Graphical Models

If X_i's are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,

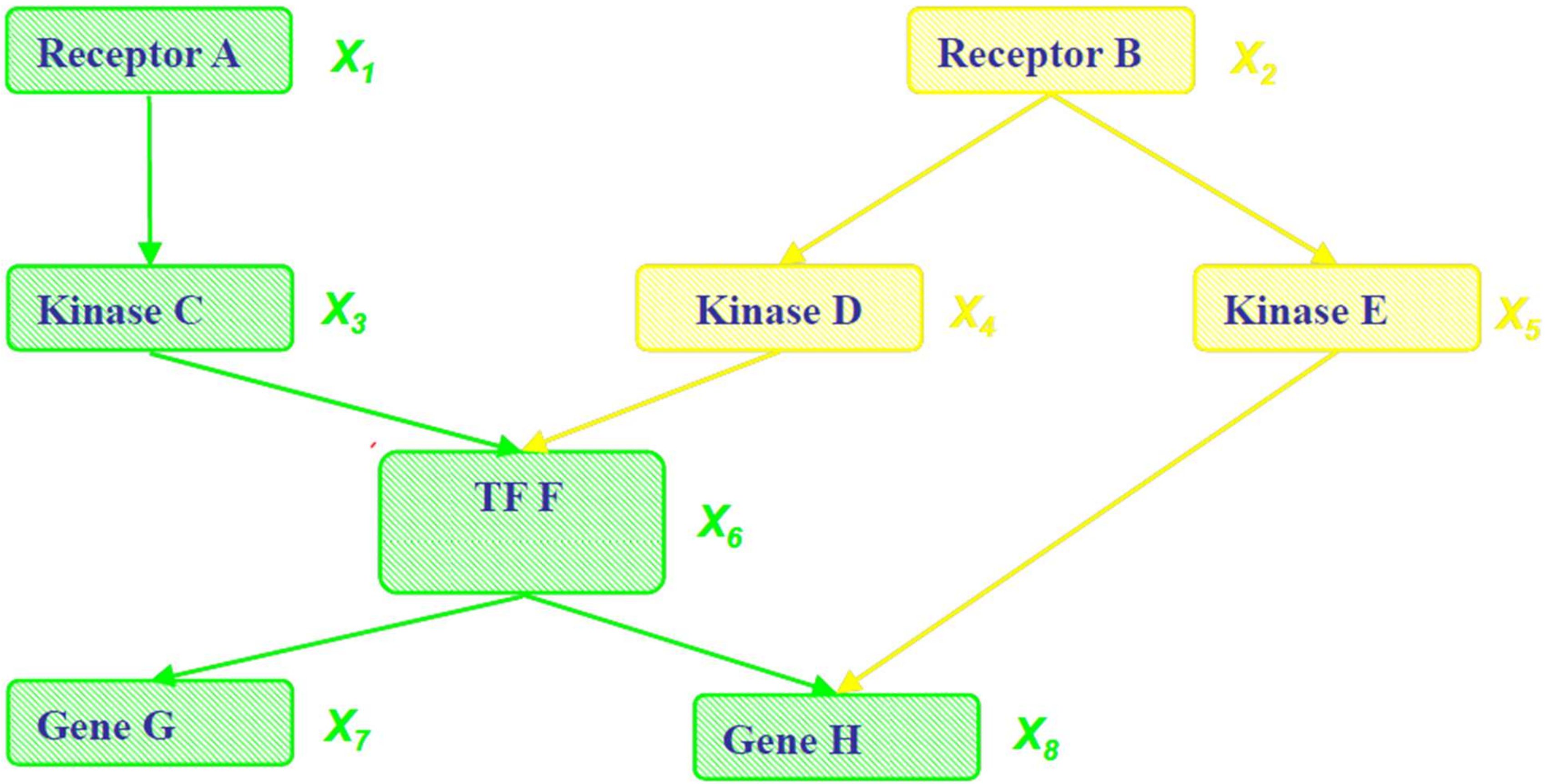


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)$$
$$P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

Why we may favor a PGM?

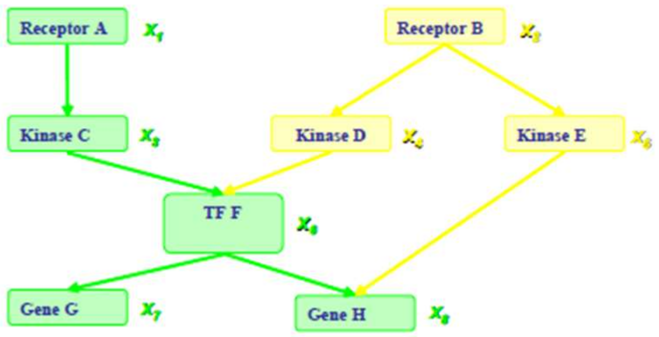
- 领域知识和因果(逻辑)结构的结合
  - 1+1+2+2+2+4+2+4=18 $2^8 \div 18 \approx 16$ 知识表示的成本降低了16倍

GM: Data Integration



Probabilistic Graphical Models

If X_i's are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_2)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_1)\, P(X_3 | X_1)$$
$$P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

Why we may favor a PGM?

- 领域知识和因果(逻辑)结构的结合
  - 1+1+2+2+2+4+2+4=18 $2^8 \div 18 \approx 16$ 知识表示的成本降低了16倍
- 异构数据融合的模块化组合

Rational Statistical Inference

The Bayes Theorem:

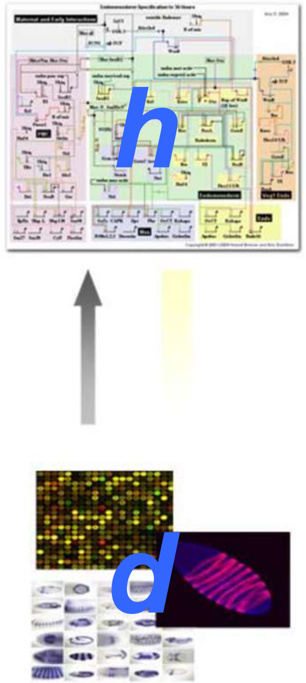
似然函数

后验概率

先验概率

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in H} p(d|h')p(h')}$$

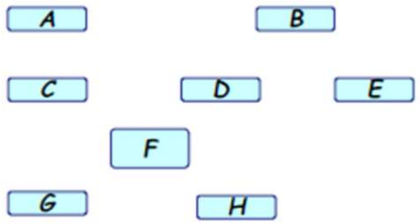
对假设空间求和



- 这允许我们以有原则的方式捕获模型的不确定性
- 但是我们如何指定和描述一个复杂的模型?
  - 通常，需要建模的基因数量是数千个

GM: MLE and Bayesian Learning

Probabilistic statements of Θ is conditioned on the values of the observed variables A_{obs} and prior $p(\ |\chi)$



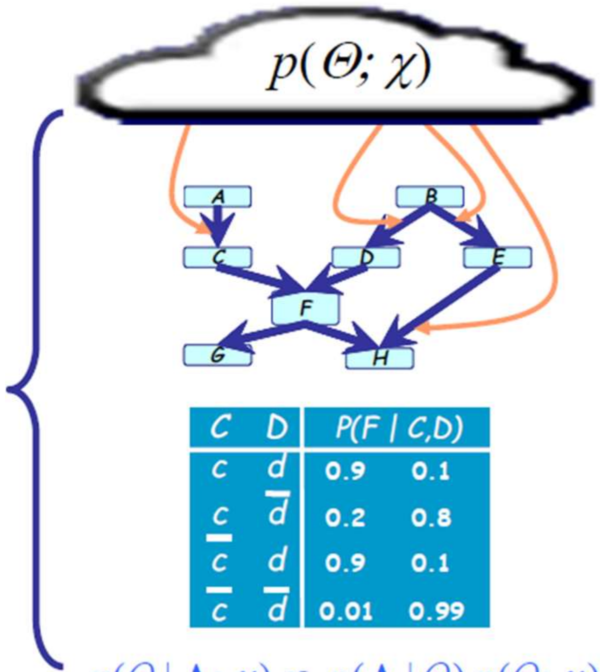
(A,B,C,D,E,…)=(T,F,F,T,F,…)

A= (A,B,C,D,E,…)=(T,F,T,T,F,…)

……..

(A,B,C,D,E,…)=(F,T,T,T,F,…)

$$\Theta_{Bayes} = \int \Theta p(\Theta|A,\chi)d\Theta$$



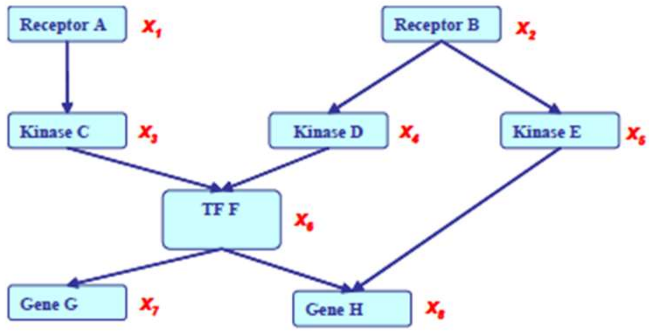
C	D	P(F | C,D)	
c	d	0.9	0.1
c	$\bar{d}$	0.2	0.8
$\bar{c}$	d	0.9	0.1
$\bar{c}$	$\bar{d}$	0.01	0.99

$$\underbrace{p(\Theta \mid \mathbf{A};\chi)}_{\text{posterior}} \propto \underbrace{p(\mathbf{A} \mid \Theta)}_{\text{likelihood}}\, \underbrace{p(\Theta;\chi)}_{\text{prior}}$$

凯歌1

Probabilistic Graphical Models

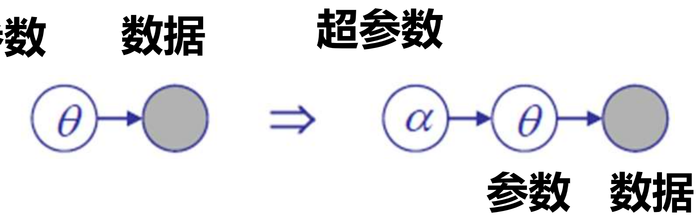
If X_i's are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)$$
$$P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

Why we may favor a PGM?

- 领域知识和因果(逻辑)结构的结合
  - ➢ 1+1+2+2+2+4+2+4=18 $2^8 \div 18 \approx 16$ 知识表示的成本降低了16倍
- 异构数据融合的模块化组合
- 贝叶斯理论
  - ➢ Knowledge meets data



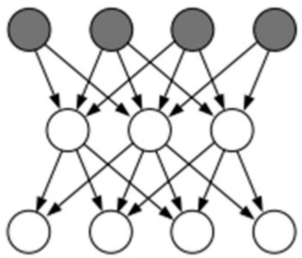
幻灯片 12

凯歌1 凯歌 张, 2019/6/25

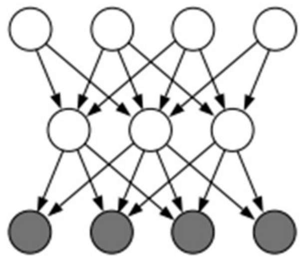
So What is a Graphical Model?

GM = Multivariate Statistics + Structure

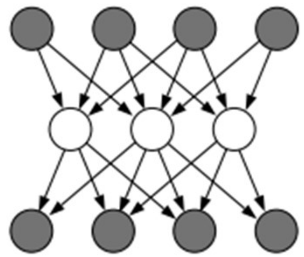
- **Some ways to use a graphical model**



预测



诊断、控制、优化

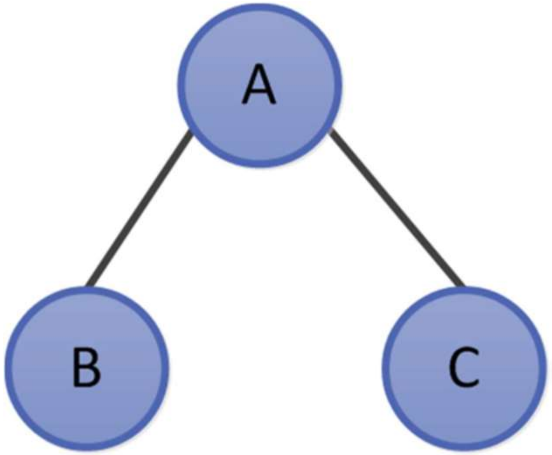


监督学习

/02

Problem Statement

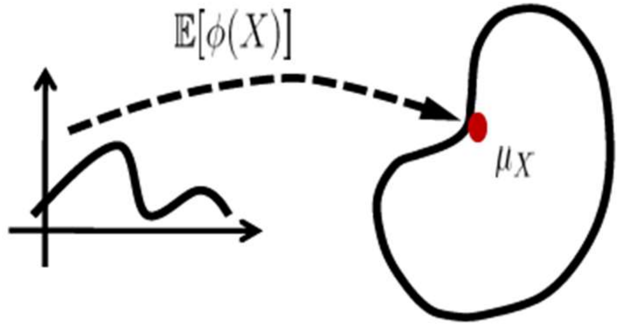
Optimization：Why do Gaussians Work?



均值、方差对与分布之间的双射

$(\mu_1, \sigma_1) \rightarrow N(\mu_1, \sigma_1)$

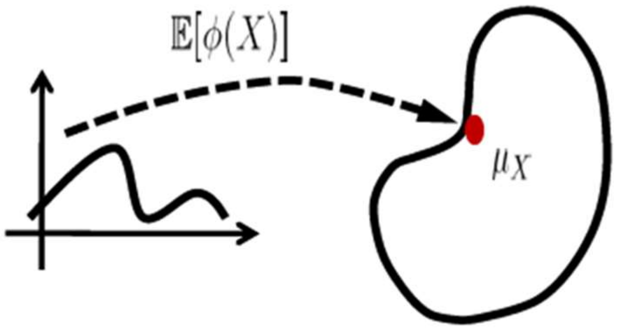
$(\mu_2, \sigma_2) \rightarrow N(\mu_2, \sigma_2)$

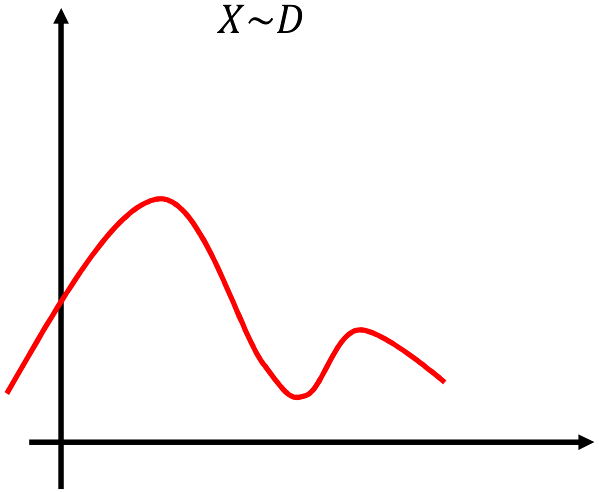


- 因为我们有参数(足够的统计数据)
- 容易获得边缘分布和条件分布信息

Create Sufficient Statistic for Arbitrary Distribution

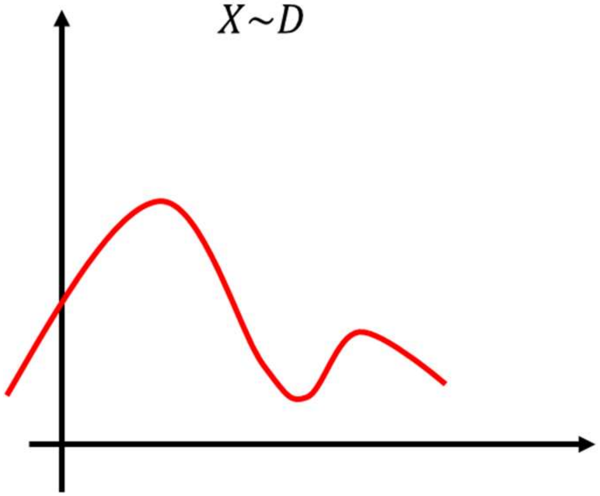
用向量$\boldsymbol{\mu_x}$来表示这个分布





$$\boldsymbol{\mu_x} = \begin{bmatrix} \cdots \\ \cdots \\ \cdots \end{bmatrix}$$

Take some Moments



$$\boldsymbol{\mu_x} = (\boldsymbol{E}(\boldsymbol{X}))$$

问题:很多分布都有相同的均值

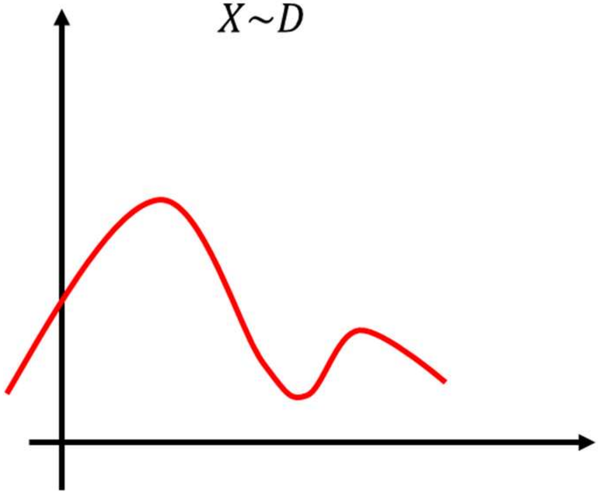
$$\boldsymbol{\mu_x} = \begin{pmatrix} \boldsymbol{E}(\boldsymbol{X}) \\ \boldsymbol{E}(\boldsymbol{X}^2) \end{pmatrix}$$

但很多分布仍然有相同的均值和方差

$$\boldsymbol{\mu_x} = \begin{pmatrix} \boldsymbol{E}(\boldsymbol{X}) \\ \boldsymbol{E}(\boldsymbol{X}^2) \\ \boldsymbol{E}(\boldsymbol{X}^3) \end{pmatrix}$$

但是很多分布仍然有相同的前三个矩

Better Idea:
Create Infinite Dimensional Statistic



$$\mu_x = \begin{pmatrix} E(X) \\ E(X^2) \\ E(X^3) \\ \dots \\ \dots \\ \dots \end{pmatrix}$$

当然，实际上是不可行的，因为存储或操作一个无限维度的向量是不可行的不可能的。

Kernel Trick

Primal Formulation:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + C\sum_j \xi$$

$$(\boldsymbol{w}^\top \phi(\boldsymbol{x}_j) + b)y_j \geqslant 1 - \xi_j \quad \forall j$$

$$\xi_j \geqslant 0 \quad \forall j$$

无限维，不能直接计算

内积可以计算

Dual Formulation:

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_i)$$

$$\sum_i \alpha_i y_i = 0$$

$$0 \leqslant \alpha_i \leqslant C \quad \forall i$$

Overview of
Hilbert Space Embedding

- 为一个分布创建一个无限维统计量。

- 两个条件
  - 从分布到统计的映射是 one-to-one
  - 虽然统计量是无限的，但它构造得很巧妙，可以应用内核技巧。

- 信念传递算法，将统计量看成条件概率表

- 引入希尔伯特空间的概念使这个结构更加正式

/03

Method

Hilbert Space

- 希尔伯特空间是向量空间的一个扩展。

$$\upsilon, \omega \in \mathcal{V} \Longrightarrow \alpha v + \beta \omega \in \mathcal{V}$$

$\upsilon, \omega \in \mathcal{V}$是有限维向量，其实也可以是函数

- 函数可以看成一个无限维的向量。

$$f = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}$$

- 希尔伯特空间是一个具有内积的完全向量空间。

$$\boxed{\langle f, g \rangle = \int f(x) g(x) dx}$$

两个函数的内积是一个数

- 希尔伯特空间中的内积必须考虑以下性质：

1.Symmetry: $\langle f, g \rangle = \langle g, f \rangle$
2.Linearity: $\langle \alpha f_1 + \beta f_2, g \rangle = \langle \alpha f_1, g \rangle + \langle \beta f_2, g \rangle$
3.Non-negativity: $\langle f, f \rangle \geq 0$
4.Zero: $\langle f, f \rangle = 0 \Longrightarrow f = 0$

Operators, Adjoints and the Outer Product

- An operator C maps a function f in one Hilbert Space to another function g in the same or another Hilbert Space. Mathematically this corresponds to:

$$g = Cf$$

线性性质:　　$\mathrm{C}(\alpha f + \beta g) = \alpha C f + \beta C g$

- the adjoint $C^T\colon \mathcal{G} \longrightarrow \mathcal{F}$ of an operator $\mathrm{C}\colon \mathcal{F} \longrightarrow \mathcal{G}$ is define such that the following always holds:

$$\langle g, Cf \rangle = \langle C^T g, f \rangle, \forall f \in \mathcal{F}, g \in \mathcal{G}$$

- Finally, also consider the Hilbert Space Outer Product$f \otimes g$, which is implicitly defined such that:

$$f \otimes g(h) = \langle g, h \rangle f$$

Reproducing Kernel Hilbert Spaces

- 再生核希尔伯特空间(RKHS)是一个希尔伯特空间，其空间的每一点都是一个连续的线性函数。

- RKHS是在Mercer内核的基础上构造的。
  - 一个Mercer核$K(x,y)$是两个变量的函数

$$\iint K(x,y)f(x)f(y)dxdy > 0, \forall f$$

- 最常用的核函数是高斯RBF核函数

$$\mathrm{K}(x,y) = exp\left(\frac{\|x-y\|_2^2}{\sigma^2}\right)$$

The Feature Function

- 考虑固定内核的一个元素。
- 结果是一个只有一个变量的函数，我们称之为feature function。
- feature function的集合称为 **feature map** 。

$$\phi_x := K(x, :)$$

- 对于高斯核函数，feature function是非归一化高斯函数。

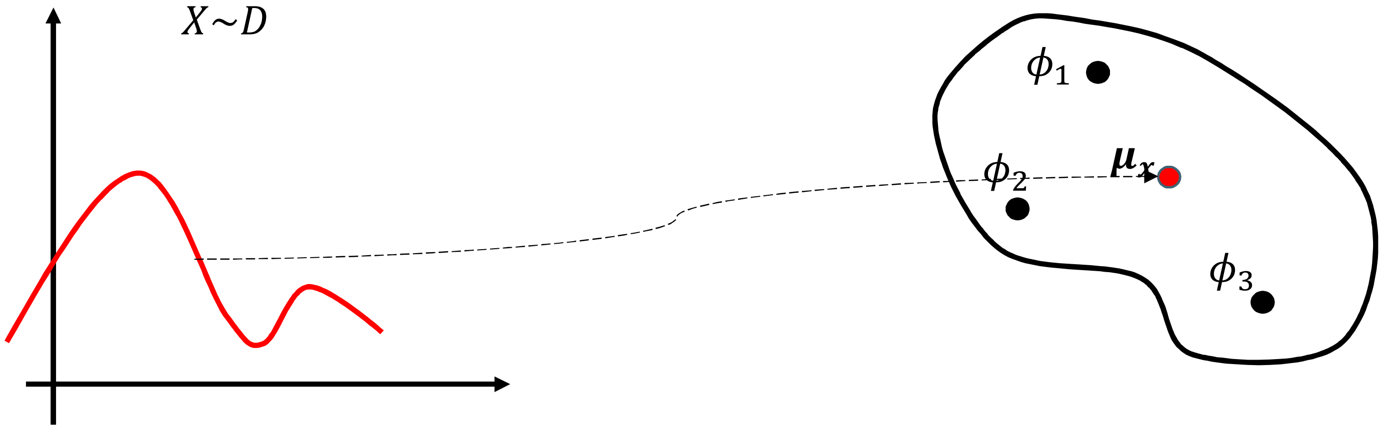
$$\phi_1(y) := exp\left(\frac{\|1 - y\|_2^2}{\sigma^2}\right)$$

$$\phi_{1.5}(y) := exp\left(\frac{\|1.5 - y\|_2^2}{\sigma^2}\right)$$

- RKHS中feature function的内积表示为:

$$\langle \phi_x, \phi_y \rangle = \langle K(x,\cdot), K(y,\cdot) \rangle = K(x, y)$$

Mean Map



The Hilbert Space Embedding of *X*

density

$$\mu_X(\cdot) = \mathbb{E}_{X \sim D}[\phi_X] = \int p_D(X)\phi_X(\cdot)dX$$

它直观地与数据的“经验估计”相对应。

Data point

$$\hat{\mu}_X = \frac{1}{N}\sum_{n=1}^{N} \phi_{x_n}$$

Example

- The finite dimensional case of an RKHS embedding for a distribution that takes on discrete values from 1 to 4.
- Now consider an RKHS mapping of the data into R^4, the feature functions in this RKHS are:

$$\phi_1 = \begin{pmatrix}1\\0\\0\\0\end{pmatrix},\ \phi_2 = \begin{pmatrix}0\\1\\0\\0\end{pmatrix},\ \phi_3 = \begin{pmatrix}0\\0\\1\\0\end{pmatrix},\ \phi_4 = \begin{pmatrix}0\\0\\0\\1\end{pmatrix},$$

- mean map is:

$$\mu_X = \mathbb{E}_X[\phi_X] = \mathbb{P}[X=1]\phi_1 + \mathbb{P}[X=2]\phi_2 + \mathbb{P}[X=3]\phi_3 + \mathbb{P}[X=4]\phi_4$$

$$\mu_X = \begin{pmatrix}\mathbb{P}[X=1]\\\mathbb{P}[X=2]\\\mathbb{P}[X=3]\\\mathbb{P}[X=4]\end{pmatrix}$$

Summary

- 希尔伯特空间嵌入提供了一种为任意分布创建足够统计量的方法。
- 可以在RKHS中嵌入边缘分布、联合分布和条件分布

References

- Smola, A. J., Gretton, A., Song, L., and Schölkopf, B., A Hilbert Space Embedding for Distributions, Algorithmic Learning Theory, E. Takimoto(Eds.), Lecture Notes on Computer Science, Springer, 2007.

- L. Song. Learning via Hilbert space embedding of distributions. PhD Thesis 2008.

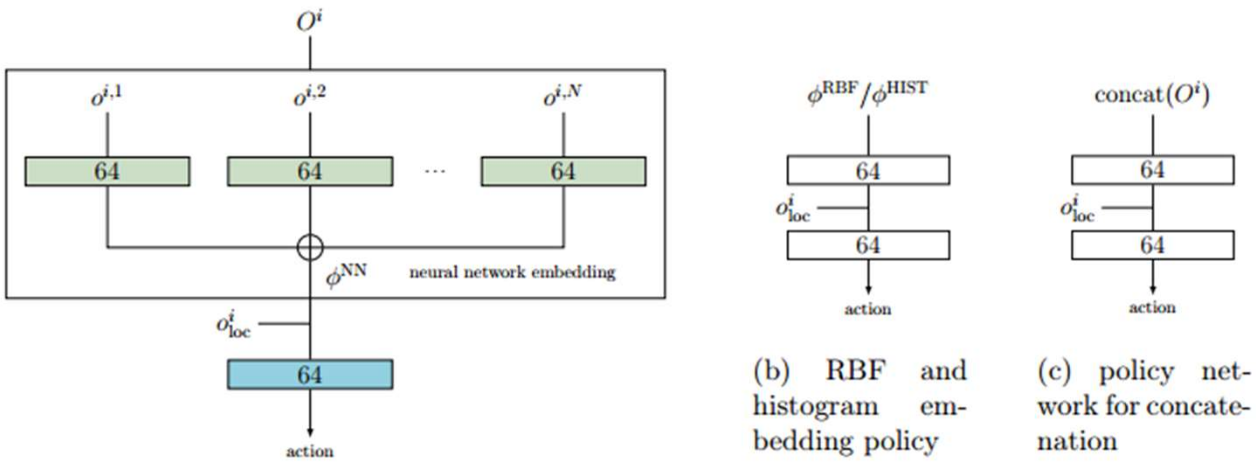
- Song, L., Huang, J., Smola, A., and Fukumizu, K., Hilbert space embeddings of conditional distributions, International Conference on Machine Learning, 2009.

/04

Future Work

Mean Embeddings as State Representations for Swarms

Mean embedding policy



(b) RBF and histogram embedding policy

(c) policy network for concatenation

(a) neural network embedding policy network

Algorithm 1 FTD-FALCON

Inputs: Environment, Flock(TD-FALCON), available action set $\mathcal{A}$

while Terminated == FALSE **do**
  for agt $\in$ Flock **do**
    /*Sense*/
    S $\leftarrow$ Sense(agt, Environment, s_t)

    /*Act*/
    $k_{sel} \leftarrow \epsilon$-Greedy Action Selection Strategy
    $(s_{t+1}) \leftarrow$ Act(agt, k_{sel}, Environment)

    /*Credit Assignment*/
    $\delta_{sep} \leftarrow$ Flock:ComputeSeparation(agt)
    $\delta_{coh} \leftarrow$ Flock:ComputeCohesion(agt)
    $\delta_{alg} \leftarrow$ Flock:ComputeAlignment(agt)
    $\delta_{fear} \leftarrow$ Flock:ComputeFear(agt)
    $\delta_{pur} \leftarrow$ Flock:ComputePreyTracking(agt)
    $r \leftarrow$ Flock:ComputeReward($\delta_{sep}, \delta_{coh}, \delta_{alg}, \delta_{fear}, \delta_{pur}$)

    agt **broadcast** *signal*(agt) to nearby agents

    /*Sense*/
    $S_{new} \leftarrow$ Sense(agt, Environment, s_{t+1})

    /*Estimate ΔQ-Value*/
    $Q \leftarrow$ PredictValue(S, k_{sel})
    $\Delta Q \leftarrow$ EstimateQDelta(S, k_{sel}, S_{new}, r, Q)
    $Q \leftarrow Q + \Delta Q$

    /*Learn*/
    $\mathbf{A} = (A_1, A_2, ..., A_n)$, where $A_k = 1, k = k_{sel}$ and $A_k = 0, \forall k \neq k_{sel}$
    FALCON::Learn(S, $\mathbf{A}$, Q)

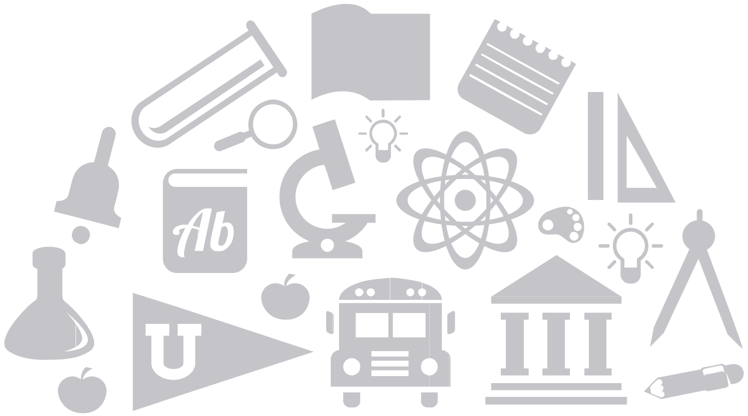
    $t \leftarrow t + 1$
  end for
end while

?

迁移学习

Maximum Mean Discrepancy (MMD)

Hilbert-Schmidt Independence Criterion (HSIC)



Thanks.

And Your Slogan Here.

Speaker name and title

www.islide.cc