自然语言处理

第6讲：对数线性模型

刘洋

内容提要

Logistic回归模型

最大熵模型

条件随机场

支持向量机的缺点

- 支持向量机可以对数据进行分类，但是无法表示类别的概率。

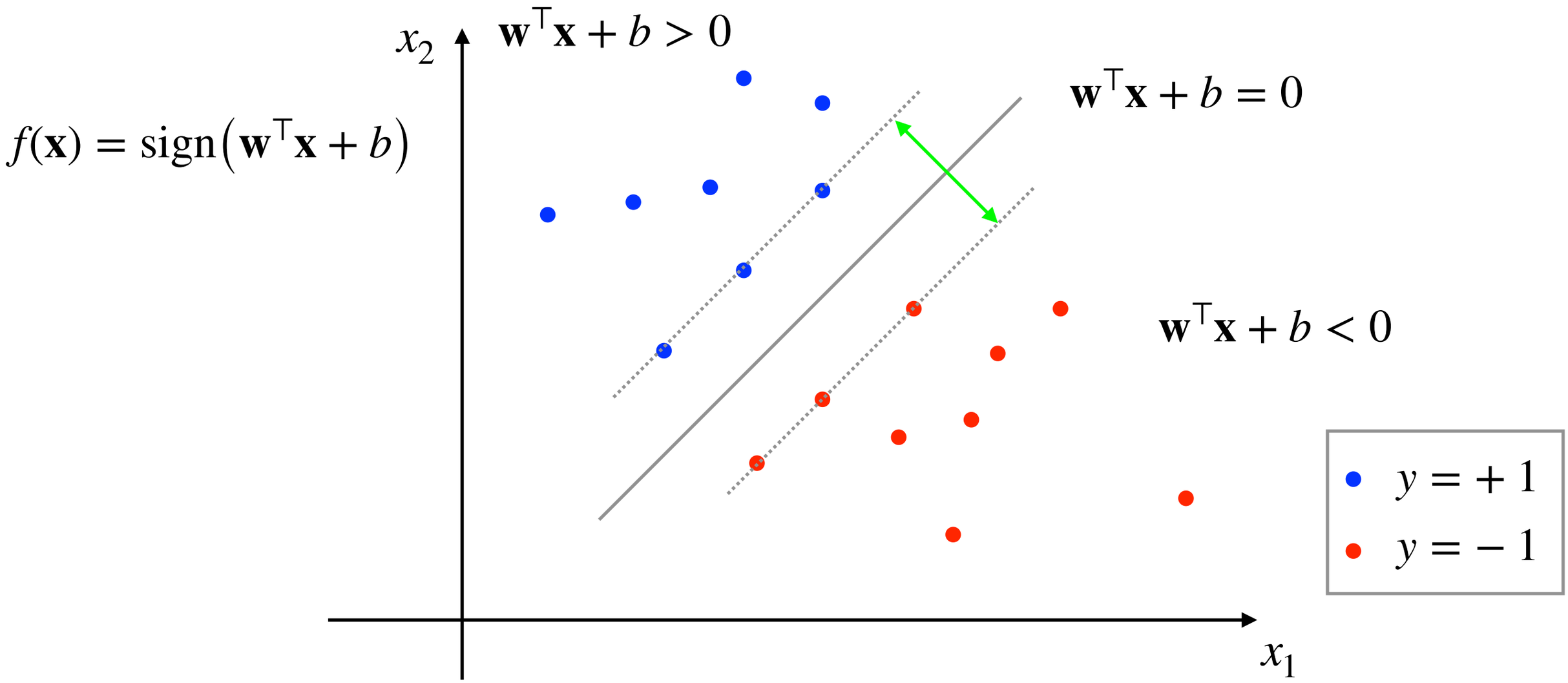

图1：支持向量机。

支持向量机的缺点

- 支持向量机可以对数据进行分类，但是无法表示类别的概率。

$$f(\mathbf{x}) = \text{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right)$$

$$P(y = +1 \,|\, \mathbf{x}) = ?$$

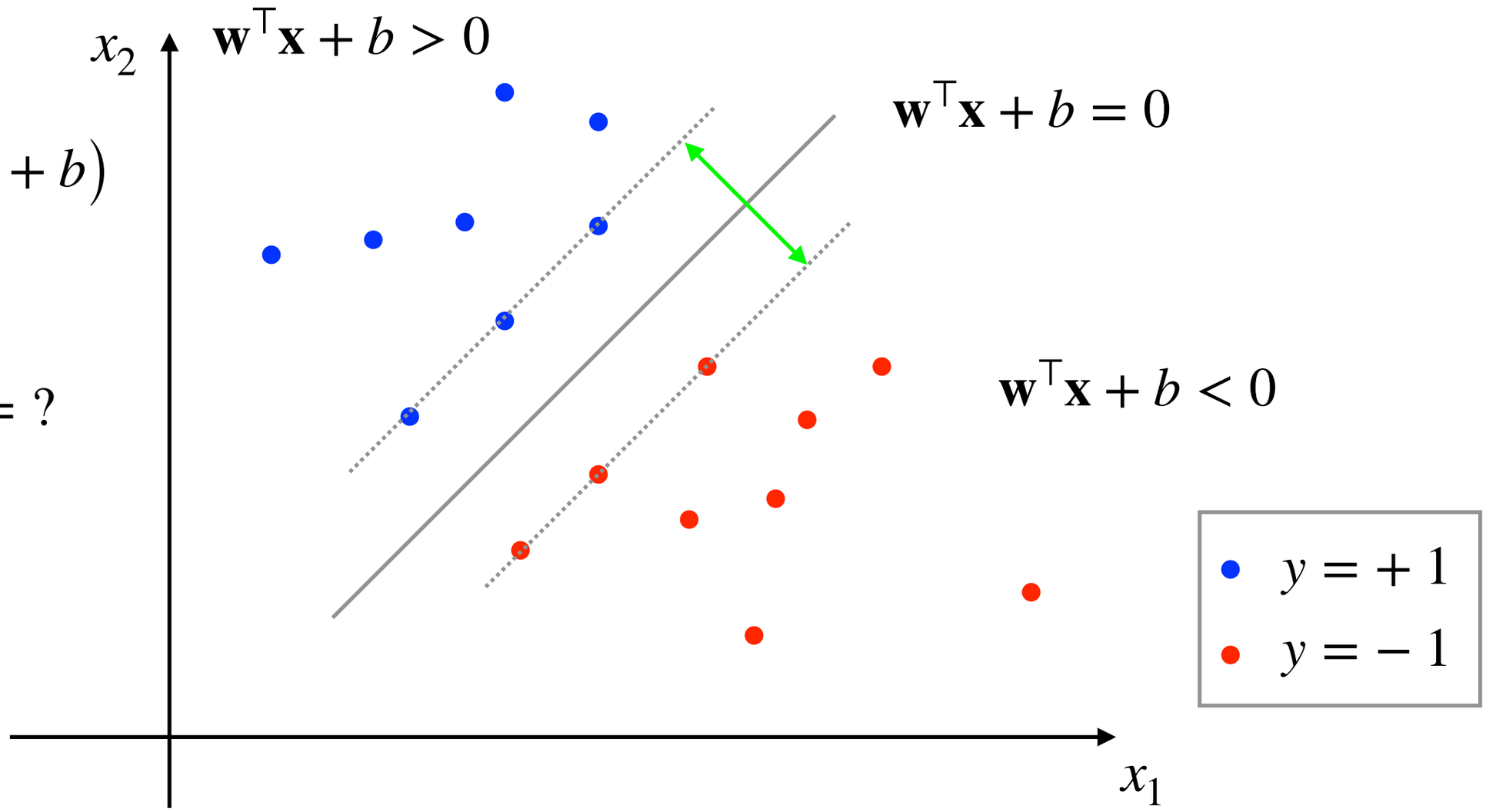

图1：支持向量机。

支持向量机的缺点

- 支持向量机可以对数据进行分类，但是无法表示类别的概率。

$$f(\mathbf{x}) = \text{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right)$$

$$P(y = +1 \,|\, \mathbf{x}) = ?$$

$$P(y = -1 \,|\, \mathbf{x}) = ?$$

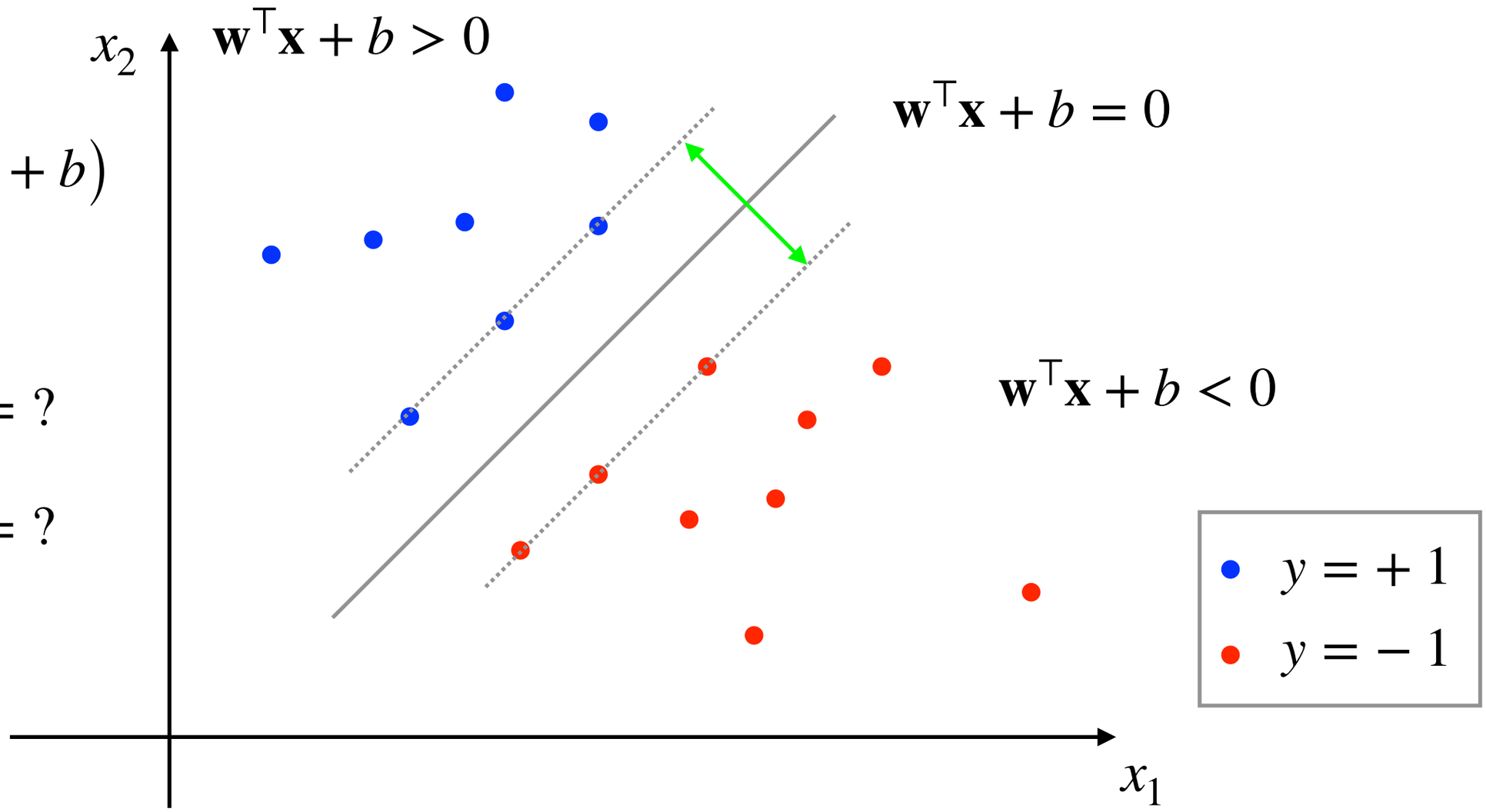

图1：支持向量机。

如何获得分类概率?

从直观上理解，对于数据点$\mathbf{x}$，当$\mathbf{w}^\top\mathbf{x}+b>0$时，我们希望

$$P(y=+1\,|\,\mathbf{x})>P(y=-1\,|\,\mathbf{x})$$

当$\mathbf{w}^\top\mathbf{x}+b<0$时，我们希望

$$P(y=+1\,|\,\mathbf{x})<P(y=-1\,|\,\mathbf{x})$$

由于分类概率必须满足非负性和归一性，而指数函数恰好满足所需的特性，因此分类概率可以定义为：

$$P(y=+1\,|\,\mathbf{x})=\frac{\exp(\mathbf{w}^\top\mathbf{x}+b)}{1+\exp(\mathbf{w}^\top\mathbf{x}+b)}$$

$$P(y=-1\,|\,\mathbf{x})=\frac{1}{1+\exp(\mathbf{w}^\top\mathbf{x}+b)}$$

通常可以通过变换将偏置项b消去，因此后面为了简便我们将不使用偏置项。

二项Logistic回归模型

二项Logistic回归模型是一种二元分类模型，能够给出分类概率：

$$P(y=1|\mathbf{x}) = \frac{\exp(\mathbf{w}^\top \mathbf{x})}{1+\exp(\mathbf{w}^\top \mathbf{x})}$$

$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp(\mathbf{w}^\top \mathbf{x})}$$

其中，$\mathbf{x} \in \mathbb{R}^{M\times 1}$是输入数据点，$y \in \{1, 0\}$是输出类别，$\mathbf{w} \in \mathbb{R}^{M\times 1}$是权重向量。注意：为了表示上的方便，我们重新定义了类别的标签。

二项Logistic回归模型的参数主要由权重向量组成：

$$\boldsymbol{\theta} = \{\mathbf{w}\}$$

有时候，我们也可以在概率中显式写出参数：$P(y|\mathbf{x};\boldsymbol{\theta})$。

参数估计

给定包含N个训练样本的训练集$D = \left\{\langle \mathbf{x}^{(i)}, y^{(i)} \rangle\right\}_{i=1}^{N}$，其中$\mathbf{x}^{(i)} \in \mathbb{R}^{M\times 1}$是第$i$个输入数据点，$y^{(i)} \in \{0, 1\}$是第$i$个输出分类。为了简便，令

$$P(y = 1 \,|\, \mathbf{x}; \boldsymbol{\theta}) = \pi(\mathbf{x}, \boldsymbol{\theta})$$

$$P(y = 0 \,|\, \mathbf{x}; \boldsymbol{\theta}) = 1 - \pi(\mathbf{x}, \boldsymbol{\theta})$$

则似然函数可以表示为

$$\prod_{i=1}^{N} \Big(\pi(\mathbf{x}^{(i)}, \boldsymbol{\theta})\Big)^{y^{(i)}} \Big(1 - \pi(\mathbf{x}^{(i)}, \boldsymbol{\theta})\Big)^{1-y^{(i)}}$$

相应的对数似然函数为

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \Big(y^{(i)} \mathbf{w}^{\top} \mathbf{x}^{(i)} - \log\big(1 + \exp(\mathbf{w}^{\top} \mathbf{x}^{(i)})\big)\Big)$$

参数估计

可以通过梯度上升法来求对数似然的极值：

$$\hat{\mathbf{w}} = \text{argmax}_{\mathbf{w}} \Big\{ L(\boldsymbol{\theta}) \Big\}$$

其基本思想是使用以下公式迭代更新参数：

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \lambda \nabla L\Big(\boldsymbol{\theta}^{(t-1)}\Big)$$

对于Logistic回归模型而言，偏导计算如下：

$$\frac{\partial L(\boldsymbol{\theta})}{\partial w_m} = \sum_{i=1}^{N} \Big(y^{(i)} - P(y^{(i)} \,|\, \mathbf{x}^{(i)}; \boldsymbol{\theta}) \Big) x_m^{(i)}$$

其中，w_m表示权重向量$\mathbf{w}$中的第m个元素。从直观上看，当模型预测$P(y^{(i)} \,|\, \mathbf{x}^{(i)}; \boldsymbol{\theta})$与标准答案$y^{(i)}$相等时，偏导为0，达到极值点。

模型推断

获得模型参数$\hat{\mathbf{w}}$之后，给定一个输入数据点$\mathbf{x}$，可以得到分类概率：

$$P(y=1|\mathbf{x};\hat{\boldsymbol{\theta}}) = \frac{\exp(\hat{\mathbf{w}}^{\top}\mathbf{x})}{1+\exp(\hat{\mathbf{w}}^{\top}\mathbf{x})}$$

$$P(y=0|\mathbf{x};\hat{\boldsymbol{\theta}}) = \frac{1}{1+\exp(\hat{\mathbf{w}}^{\top}\mathbf{x})}$$

我们可以很容易地就得到分类结果。如果

$$P(y=1|\mathbf{x};\hat{\boldsymbol{\theta}}) > P(y=0|\mathbf{x};\hat{\boldsymbol{\theta}})$$

分类结果为$y=1$。此时，实际上等价于

$$\exp(\hat{\mathbf{w}}^{\top}\mathbf{x}) > 1 \quad \Leftrightarrow \quad \hat{\mathbf{w}}^{\top}\mathbf{x} > 0$$

反之，分类结果为$y=0$。由于在推断时，分类结果实际上取决于取对数后的线性模型，我们可以将Logistic回归模型称为是一种对数线性模型。

多项Logistic回归模型

我们可以将二项Logistic回归模型扩展到多项Logistic回归模型，从而可以处理多元分类问题。

对于K项多项Logistic回归模型，当$k = 1, \ldots, K-1$时，分类概率为

$$P(y = k | \mathbf{x}) = \frac{\exp(\mathbf{w}_k^\top \mathbf{x})}{1 + \sum_{k=1}^{K-1} \exp(\mathbf{w}_k^\top \mathbf{x})}$$

当$k = K$时，分类概率为

$$P(y = k | \mathbf{x}) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\mathbf{w}_k^\top \mathbf{x})}$$

需要注意的是，该模型是归一的。多项Logistic回归模型的参数估计和推断与二项Logistic回归模型类似。

内容提要

Logistic回归模型

最大熵模型

条件随机场

熵

如果X是一个离散型随机变量，其概率分布为$P(X=x)=p(x)$，$x\in\mathcal{X}$。其中，$\mathcal{X}$表示随机变量所有取值的集合，则该随机变量的熵为：

$$H(X)=-\sum_{x\in\mathcal{X}}p(x)\log_2 p(x)$$

我们约定$0\log_2 0=0$。

当X服从均匀分布时，熵最大。假设$P(X=0)=0.1$，$P(X=1)=0.9$，则熵计算为

$$H(X)=-0.1\times\log_2 0.1-0.9\times\log_2 0.9=0.4690$$

如果$P(X=0)=0.5$，$P(X=1)=0.5$，则熵计算为

$$H(X)=-0.5\times\log_2 0.5-0.5\times\log_2 0.5=1.0000$$

最大熵原理

假设天气作为一个随机变量有三个取值：晴天、阴天和雨天。已知明天的天气为阴天的概率为0.4，问明天的天气为晴天和雨天的概率分别是多少？由于我们没有任何额外信息作为依据来判定明天的天气是否是晴天或雨天，因此最稳妥的做法是假设晴天和雨天都是等可能的，即都为0.3。

最大熵原理的基本思想是：知之为知之，不知为不知。对于未知的多种选择，最公平的方案是假设每种选择都是等可能性的，即熵最大。最大熵原理任务熵最大的模型是最好的模型。

通常用约束条件来确定概率模型的集合，所以最大熵原理也可以表述为在满足约束条件的模型集合中选择熵最大的模型。例如，在上面的例子，“明天的天气为阴天的概率为0.4”就是一个约束条件。

真实分布与经验分布

由于随机变量的真实分布通常是不可获知的，可以使用相对频度在训练集训练集$\mathscr{D} = \{\langle x^{(i)}, y^{(i)} \rangle\}_{i=1}^{N}$上估计联合概率分布$P(x, y)$的经验分布：

$$\tilde{P}(x, y) = \frac{c(x, y)}{N}$$

其中，$c(x, y)$表示训练数据中样本$\langle x, y \rangle$出现的频次。

所谓经验分布，是指从实际的训练数据中估计得到的概率分布。一般而言，当训练数据集规模越大，经验分布越接近于真实分布。因此，通常需要足够的训练数据才可以保证模型参数估计的可靠性。

类似地，我们可以估计边缘概率分布$P(x)$的经验分布：

$$\tilde{P}(x) = \frac{c(x)}{N}$$

特征

- 盲人摸象：一个物体应该具备哪些特征才是大象？

图片来源：www.91experience.com

图2：盲人摸象。

特征函数

给定输入x和输出y，我们可以使用特征函数$f(x,y)$来刻画能够反映输入和输出之间关联的某个特性。

特征函数通常可以定义为二值函数，即$f(x,y) \in \{0,1\}$：

$$f(x,y) = \begin{cases} 1, & x\text{与}y\text{满足某一事实} \\ 0, & \text{否则} \end{cases}$$

例如，我们可以把“大象的耳朵像扇子”定义为一个特征函数。通过特征函数的方式，我们可以把人类所观察到的各种规律都表示出来并加入概率模型。

特征函数除了二值函数，通常也可以定义为连续的实数：$f(x,y) \in \mathbb{R}$。

约束条件

一个表达能力强的模型应满足什么条件？我们认为一个表达能力强的模型应当使得特征函数关于经验分布和模型的期望值相等：

$$\mathbb{E}_P[f] = \mathbb{E}_{\tilde{P}}[f]$$

只有如此，模型才有可能从训练集中获取有用的信息。

特征函数$f(x, y)$关于经验分布$\tilde{P}(x, y)$的期望值可以表示为：

$$\mathbb{E}_{\tilde{P}}[f] = \sum_{x,y} \tilde{P}(x, y) f(x, y)$$

特征函数$f(x, y)$关于模型$P(y \mid x; \boldsymbol{\theta})$与经验分布$\tilde{P}(x)$的期望值可以表示为：

$$\mathbb{E}_P[f] = \sum_{x,y} \tilde{P}(x) P(y \mid x; \boldsymbol{\theta}) f(x, y)$$

最大熵模型

给定训练集$\mathcal{D} = \{\langle x^{(i)}, y^{(i)} \rangle\}_{i=1}^{N}$和特征函数$f_k(x, y)$（$k = 1, \ldots, K$），在条件概率分布$P(y \mid x)$上的条件熵可定义为

$$H(P) = -\sum_{x,y} \tilde{P}(\mathbf{x}) P(y \mid x) \log P(y \mid x)$$

最大熵模型是在满足约束条件的情况下熵最大的模型

$$\max_{P} H(P)$$

$$\text{s.t. } \mathbb{E}_P[f_k] = \mathbb{E}_{\tilde{P}}[f_k], k = 1, \ldots, K \qquad \text{(公式1)}$$

$$\sum_{y} P(y \mid x) = 1$$

因此，最大熵模型的学习等价于一个受限优化问题。

拉格朗日乘子法

可以使用拉格朗日乘子法将问题转化为无约束优化：

$$L(P,\mathbf{w}) = -H(P) - w_0\left(\sum_y P(y\,|\,x) - 1\right) - \sum_{k=1}^{K} w_k\left(\mathbb{E}_P[f_k] - \mathbb{E}_{\tilde{P}}[f_k]\right)$$

$$= \sum_{x,y} \tilde{P}(x)P(y\,|\,x)\log P(y\,|\,x)$$

$$-w_0\left(\sum_y P(y\,|\,x) - 1\right) \qquad \text{(公式2)}$$

$$-\sum_{k=1}^{K} w_k\left(\sum_{x,y} \tilde{P}(x)P(y\,|\,x)f_k(x,y) - \sum_{x,y} \tilde{P}(x,y)f_k(x,y)\right)$$

注意：我们将$P(y\,|\,x)$看成是自变量，另一个自变量是拉格朗日乘子$\mathbf{w}$。

原始问题

定义一个只包含自变量$P(y|x)$的函数

$$g(P) = \max_{\mathbf{w}} L(P, \mathbf{w})$$

如果$P(y|x)$违反了约束条件，即存在某个$P(y|x)$使得$\mathbb{E}_P[f_k] \neq \mathbb{E}_{\tilde{P}}[f_k]$或者$\sum_y P(y|x) \neq 1$，那么一定可以找到某个$w_k$使得$g(P)$的取值为正无穷大。

因此，函数$g(P)$实际上等价于

$$g(P) = \begin{cases} -H(P) & P\text{满足约束条件} \\ +\infty & \text{否则} \end{cases}$$

因此，原始问题可以表述为：

$$\min_P g(P) = \min_P \max_{\mathbf{w}} L(P, \mathbf{w})$$

原始问题与对偶问题

虽然与支持向量机不一样，最大熵模型中没有不等式约束，只有等式约束，我们依然可以将其描述为min-max优化问题，这样便使得将原始问题转换为对偶问题成为可能。

最大熵的原始问题是：

$$\min_{P} \max_{\mathbf{w}} L(P, \mathbf{w})$$

与之对应的对偶问题是：

$$\max_{\mathbf{w}} \min_{P} L(P, \mathbf{w})$$

由于$L(P, \mathbf{w})$函数满足KKT条件，原始问题的解与对偶问题的解是等价的，因此可以通过求解对偶问题来获得原始问题的解。

对偶问题求解

首先考虑内层的求极值问题：$\min_{P} L(P, \mathbf{w})$。计算$L(P, \mathbf{w})$对于$P(y|x)$的偏导，可以得到

$$\frac{\partial L(P, \mathbf{w})}{\partial P} = \sum_{x,y} \tilde{P}(x)\Big(\log P(y|x) + 1\Big) - \sum_{y} w_0 - \sum_{x,y}\left(\tilde{P}(x)\sum_{k=1}^{K} w_k f_k(x, y)\right)$$

$$= \sum_{x,y} \tilde{P}(x)\left(\log P(y|x) + 1 - w_0 - \sum_{k=1}^{K} w_k f_k(x, y)\right) \qquad \text{(公式3)}$$

令公式3取值为0，可以得到P与$\mathbf{w}$之间的关系：

$$P(y|x) = \frac{\exp\left(\sum_{k=1}^{K} w_k f_k(x, y)\right)}{\exp(1 - w_0)} \qquad \text{(公式4)}$$

对偶问题求解

由于$P(y|x)$作为概率必须满足归一性，因此可以将公式4重写为

$$P(y|x) = \frac{1}{Z(x, \mathbf{w})} \exp\left(\sum_{k=1}^{K} w_k f_k(x, y) \right) \qquad \text{(公式5)}$$

其中，配分函数或归一化因子$Z(x, \mathbf{w})$定义为

$$Z(x, \mathbf{w}) = \sum_{y'} \exp\left(\sum_{k=1}^{K} w_k f_k(x, y) \right) \qquad \text{(公式6)}$$

公式5所描述的就是最大熵模型的常见形式。由此可以看出，最大熵模型本质上是一个对数线性模型，有两个重要组成部分：

① 特征函数：即$f_k(x, y)$，刻画了输入x和输出y之间的特性，人工设计。

② 特征权重：即w_k，体现了对应特征的重要性，自动学习。

对偶问题求解

然后考虑外层的求极值问题：$\max_{\mathbf{w}} \min_{P} L(P, \mathbf{w})$，将公式5代入公式2，可以得到一个自变量为$\mathbf{w}$的函数：

$$\mathbf{\Psi}(\mathbf{w}) = \sum_{x,y} \tilde{P}(x) P(y \mid x; \mathbf{w}) \log P(y \mid x; \mathbf{w})$$

$$- \sum_{k=1}^{K} w_k \Big(\sum_{x,y} \tilde{P}(x) P(y \mid x; \mathbf{w}) f_k(x, y) - \sum_{x,y} \tilde{P}(x, y) f_k(x, y) \Big)$$

$$= \sum_{x,y} \tilde{P}(x, y) \sum_{k=1}^{K} w_k f_k(x, y) + \sum_{x,y} \tilde{P}(x) P(y \mid x; \mathbf{w}) \Big(\log P(y \mid x; \mathbf{w}) - \sum_{k=1}^{K} w_k f_k(x, y) \Big)$$

$$= \sum_{x,y} \tilde{P}(x, y) \sum_{k=1}^{K} w_k f_k(x, y) - \sum_{x} \tilde{P}(x) \log Z(x, \mathbf{w}) \qquad \text{(公式7)}$$

为了体现自变量$\mathbf{w}$的依赖关系，我们将模型写成$P(y \mid x; \mathbf{w})$。

对偶问题求解

计算$\boldsymbol{\psi}(\mathbf{w})$关于第$k$个特征权重$w_k$的偏导：

$$\frac{\partial \boldsymbol{\psi}(\mathbf{w})}{\partial w_k} = \sum_{x,y} \tilde{P}(x,y) f_k(x,y) - \sum_{x,y} \tilde{P}(x) P(y \mid x; \mathbf{w}) f_k(x,y)$$

$$= \mathbb{E}_{\tilde{P}}[f_k] - \mathbb{E}_P[f_k] \qquad \text{(公式8)}$$

之后，可以利用梯度下降法等优化方法求解极值。

由公式8可以看出，在计算一个特征权重w_k的偏导时，需要计算相应特征函数$f_k(x,y)$的两个期望的差值。第一个期望$\mathbb{E}_{\tilde{P}}[f_k]$在标注数据上计算，也就是可以观察到所有的$x$和$y$。第二个期望$\mathbb{E}_P[f_k]$则只使用标注数据中的$x$，而需要枚举所有可能的$y$。

由于期望的计算涉及到指数级空间上的求和，如果特征函数本身具备良好的局部性质，便可以根据问题特性设计高效的动态规划算法（类似于HMM中的前向算法和后向算法）进行精准计算。如果特征函数是非局部的，通常要采用采样等技术进行近似计算。

训练与推断

在求出对偶问题的解之后，即

$$\hat{\mathbf{w}} = \text{argmax}_{\mathbf{w}} \Big\{ \boldsymbol{\Psi}(\mathbf{w}) \Big\}$$

我们便可以得到原始问题的解：

$$\hat{P}(y \,|\, x; \hat{\mathbf{w}}) = \frac{\exp\Big(\sum_{k=1}^{K} \hat{w}_k f_k(x, y) \Big)}{Z(x, \hat{\mathbf{w}})}$$

在做推断时，给定输入x，可以使用以下公式计算最优输出$\hat{y}$

$$\hat{y} = \text{argmax}_y \Big\{ \hat{P}(y \,|\, x; \hat{\mathbf{w}}) \Big\}$$

$$= \text{argmax}_y \left\{ \sum_{k=1}^{K} \hat{w}_k f_k(x, y) \right\}$$

最大熵模型与支持向量机、Logistic回归模型一样，都是对数线性模型。

内容提要

Logistic回归模型

最大熵模型

条件随机场

概率有向图模型

隐马尔科夫模型是一个典型的概率有向图模型，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 \,|\, z_1) \times \prod_{t=2}^{T} p(z_t \,|\, z_{t-1}) \times p(x_t \,|\, z_t)$$

图2：隐马尔科夫模型是一个有向图模型。

概率有向图模型

隐马尔科夫模型是一个典型的概率有向图模型，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 \mid z_1) \times \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \times p(x_t \mid z_t)$$

图2：隐马尔科夫模型是一个有向图模型。

概率有向图模型

隐马尔科夫模型是一个典型的概率有向图模型，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 \mid z_1) \times \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \times p(x_t \mid z_t)$$

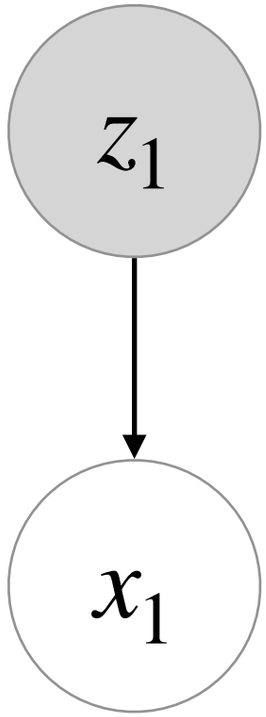

图2：隐马尔科夫模型是一个有向图模型。

概率有向图模型

隐马尔科夫模型是一个典型的概率有向图模型，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 \mid z_1) \times \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \times p(x_t \mid z_t)$$

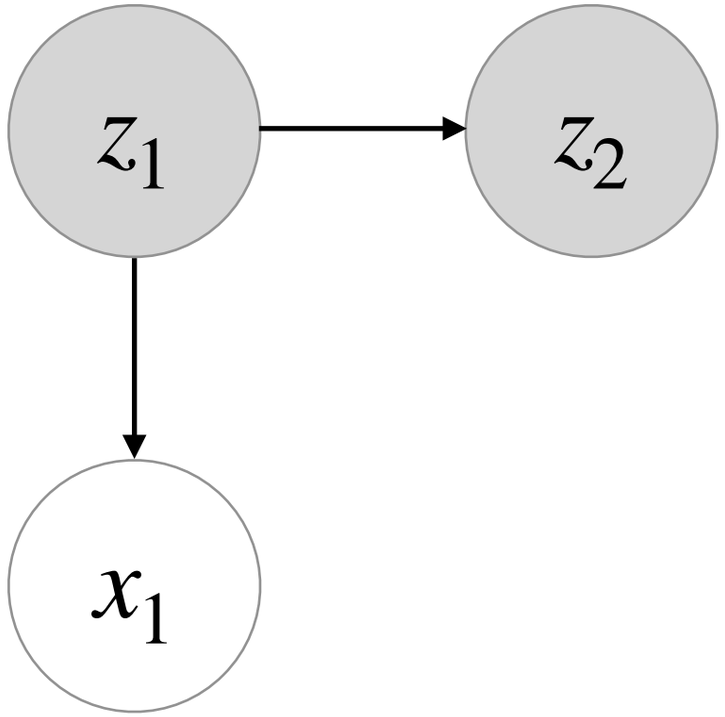

图2：隐马尔科夫模型是一个有向图模型。

概率有向图模型

隐马尔科夫模型是一个典型的概率有向图模型，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 \mid z_1) \times \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \times p(x_t \mid z_t)$$

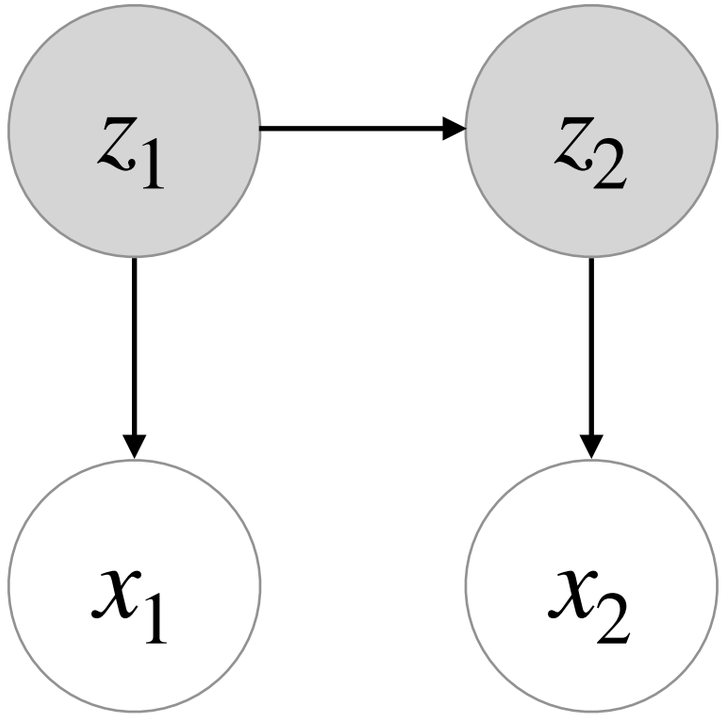

图2：隐马尔科夫模型是一个有向图模型。

概率有向图模型

隐马尔科夫模型是一个典型的概率有向图模型，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 \mid z_1) \times \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \times p(x_t \mid z_t)$$

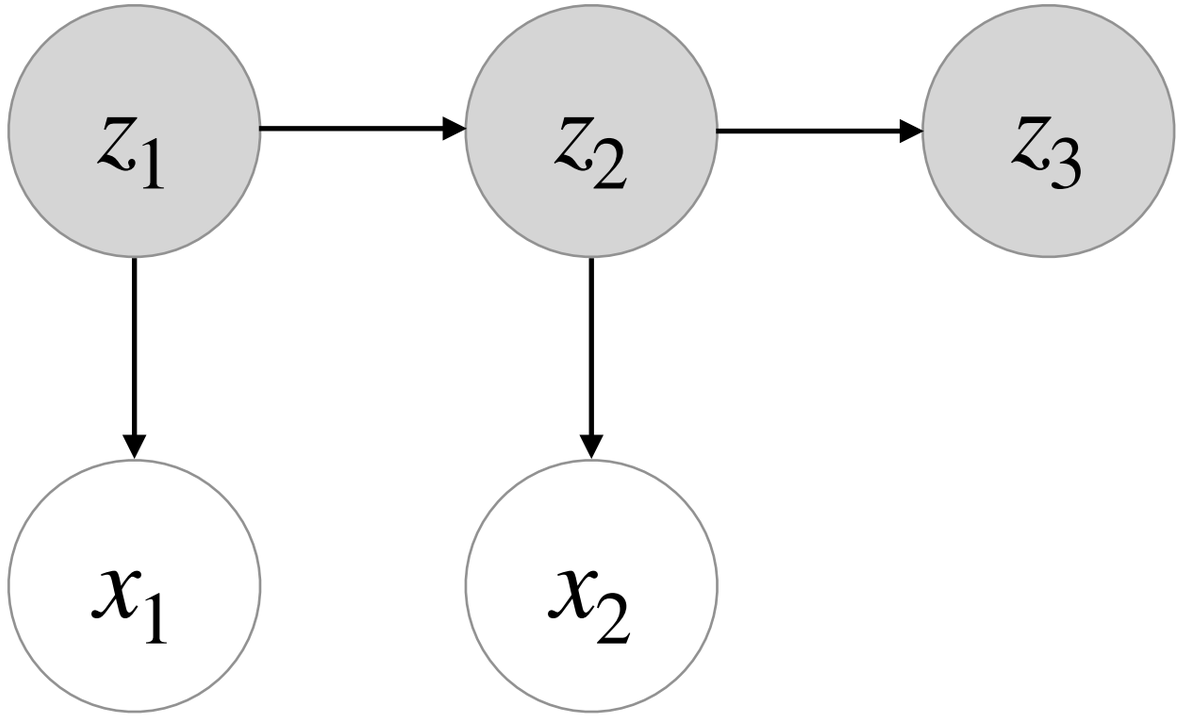

图2：隐马尔科夫模型是一个有向图模型。

概率有向图模型

隐马尔科夫模型是一个典型的概率有向图模型，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 \mid z_1) \times \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \times p(x_t \mid z_t)$$

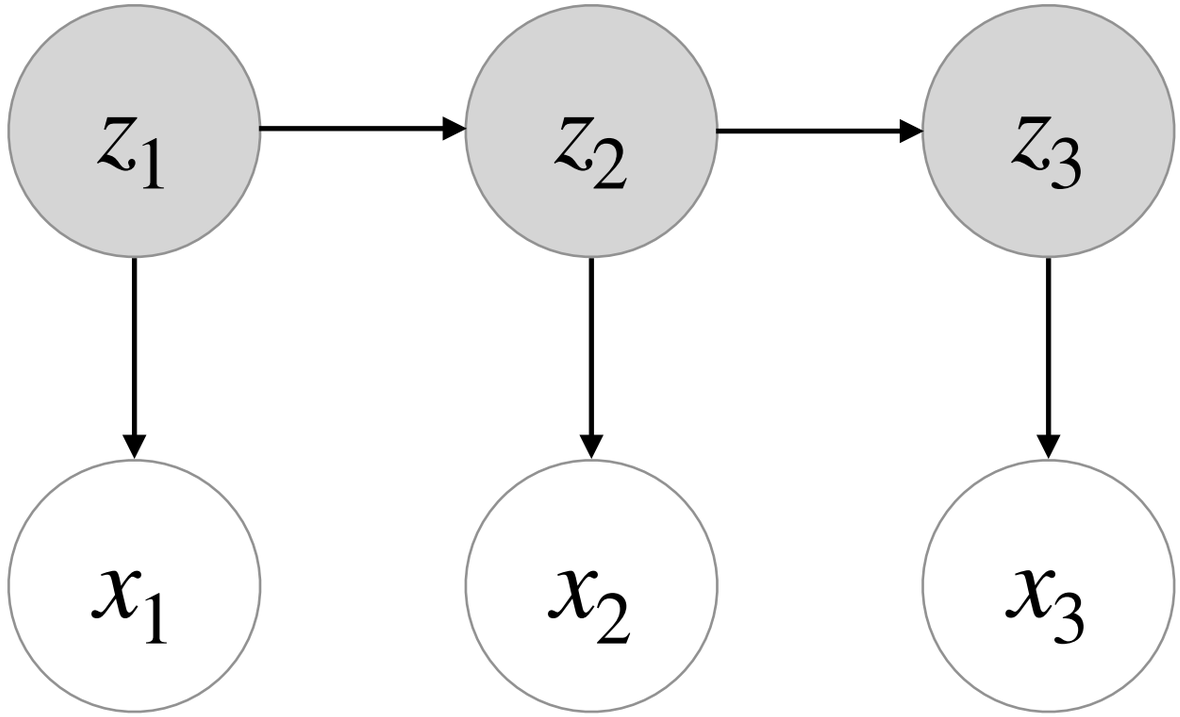

图2：隐马尔科夫模型是一个有向图模型。

概率有向图模型

隐马尔科夫模型是一个典型的概率有向图模型，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 \mid z_1) \times \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \times p(x_t \mid z_t)$$

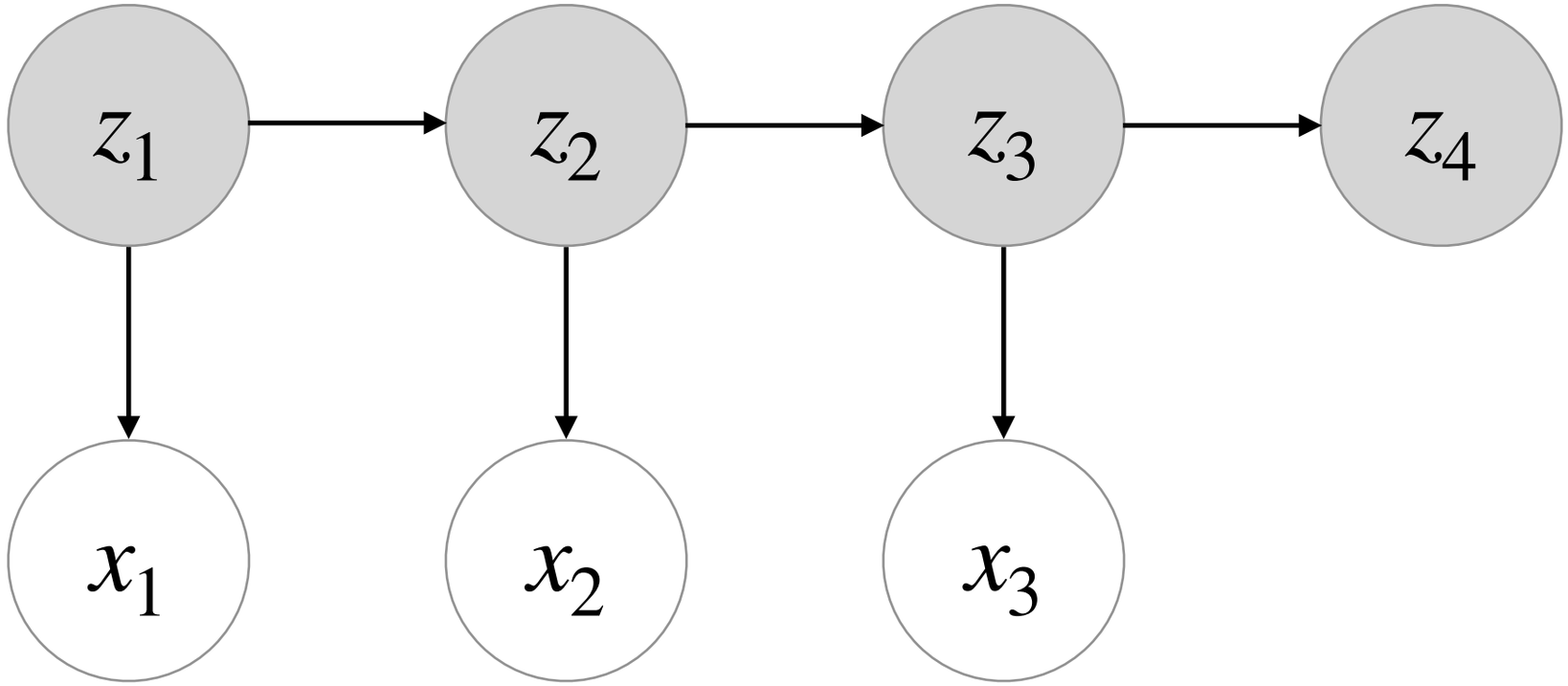

图2：隐马尔科夫模型是一个有向图模型。

概率有向图模型

隐马尔科夫模型是一个典型的概率有向图模型，边的箭头决定了节点生成的先后顺序：

$$P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(z_1) \times p(x_1 \mid z_1) \times \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \times p(x_t \mid z_t)$$

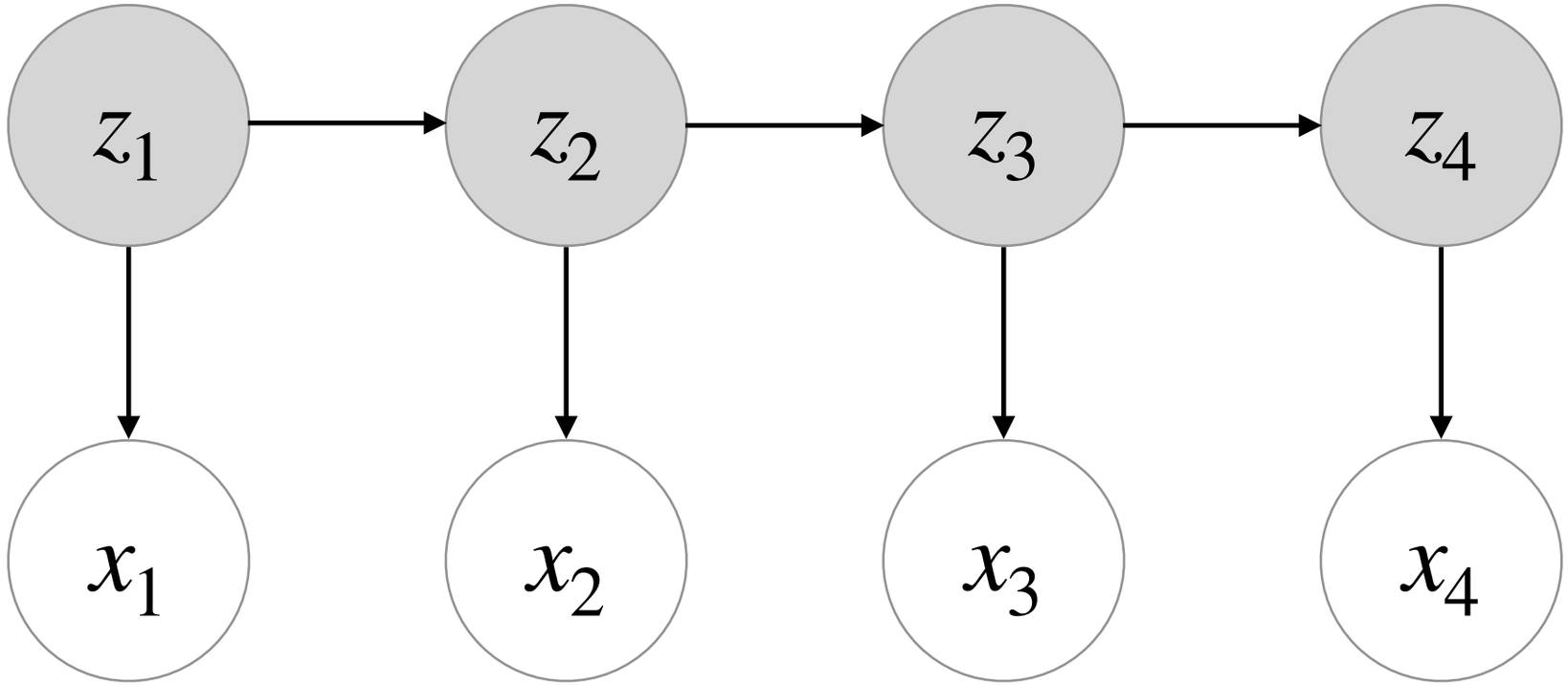

图2：隐马尔科夫模型是一个有向图模型。

概率无向图模型

给定一个无向图$G = (V, E)$，V表示节点的集合，每个节点表示一个随机变量，E表示边的集合，每条边表示随机变量之间的依赖关系。如果这些随机变量的联合概率分布满足成对、局部或全局马尔科夫性，就称此联合概率分布为概率无向图模型或马尔科夫随机场。

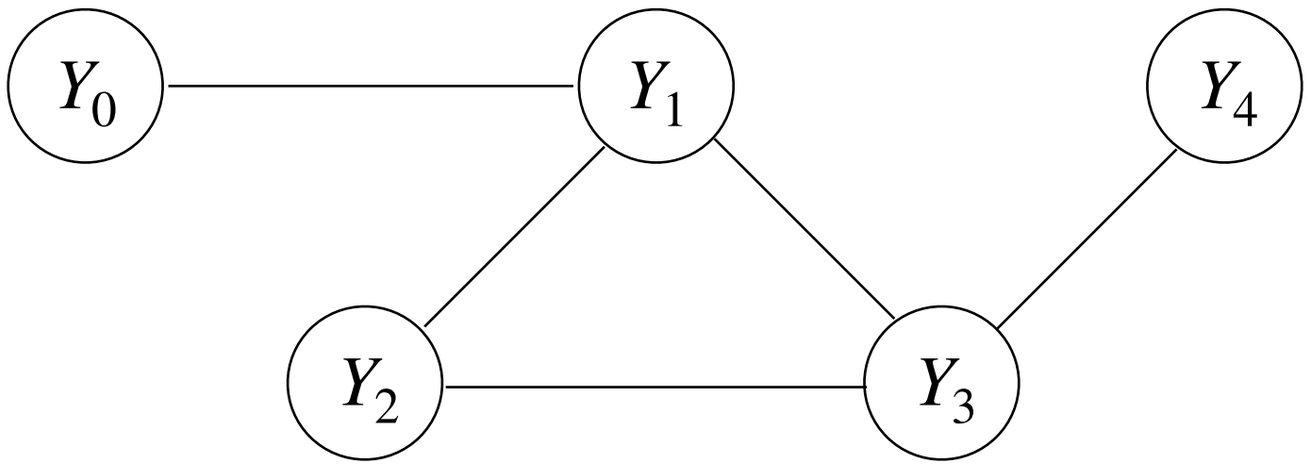

图3：概率无向图模型。

成对马尔科夫性

设u和v是无向图G任意两个没有边连接的点，其余所有节点为O。成对马尔科夫性是指给定随机变量组Y_O的条件下，随机变量Y_u和Y_v是条件独立的：

$$P(Y_u, Y_v | Y_O) = P(Y_u | Y_O) P(Y_v | Y_O)$$

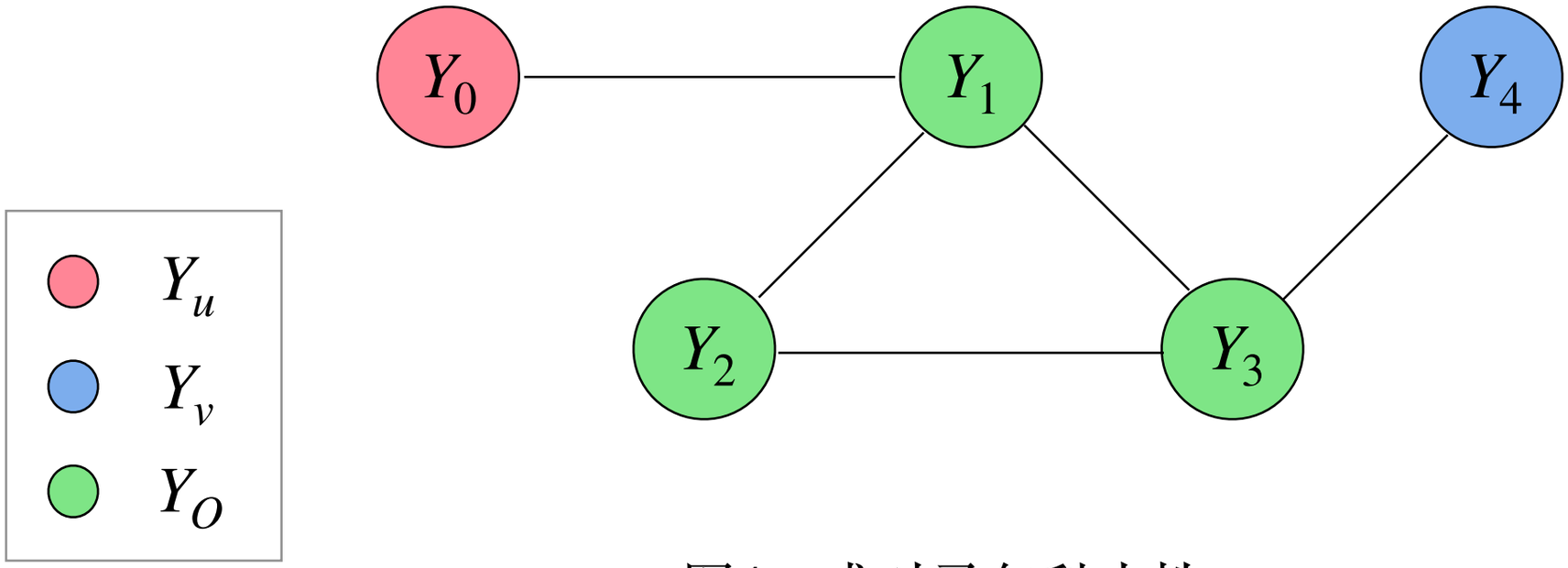

图4：成对马尔科夫性。

局部马尔科夫性

设u是无向图G中任意一个节点，W是与u有边连接的所有节点，O是u和W以外的其他所有节点。局部马尔科夫性是指给定随机变量组Y_W的条件下，随机变量Y_u和随机变量组Y_O是条件独立的：

$$P(Y_u, Y_O | Y_W) = P(Y_u | Y_W) P(Y_O | Y_W)$$

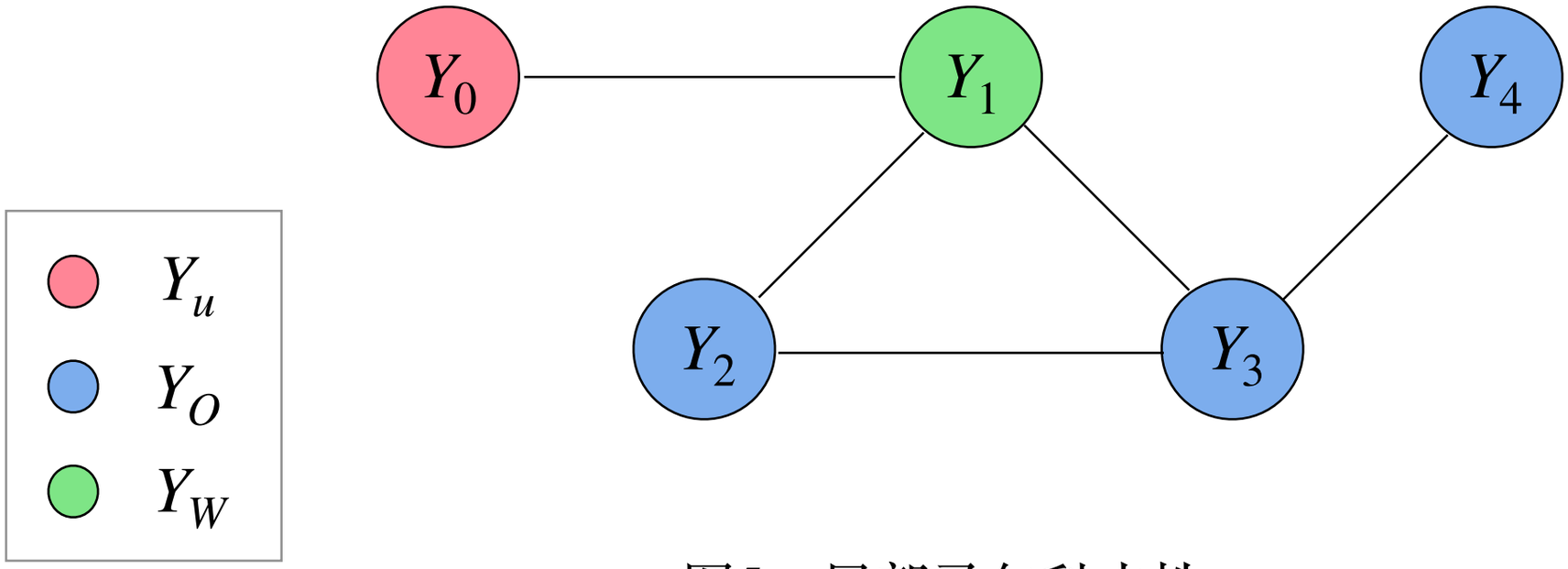

图5：局部马尔科夫性。

全局马尔科夫性

设节点集合A和B是在无向图G中被节点集合C分开的任意节点集合，全局马尔科夫性是指给定随机变量组Y_C的条件下，随机变量组Y_A和随机变量组Y_B是条件独立的：

$$P(Y_A, Y_B | Y_C) = P(Y_A | Y_C)P(Y_B | Y_C)$$

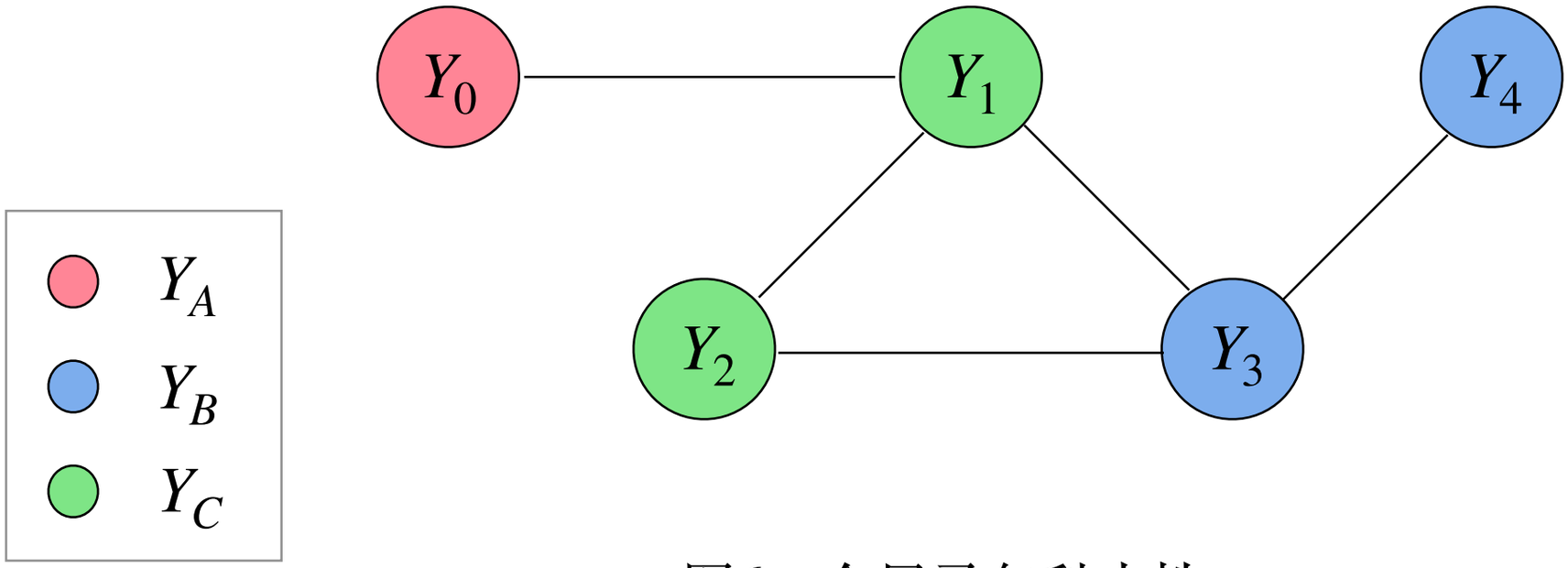

图6：全局马尔科夫性。

最大团

无向图中任意两个节点均有边连接的节点子集成为团。如果C是无向图G的一个团，而且不能再加入任何一个节点使其成为一个更大的团，则称C为最大团。

例如，在图7中，共有三个最大团：$\{Y_0, Y_1\}, \{Y_1, Y_2, Y_3\}, \{Y_3, Y_4\}$。

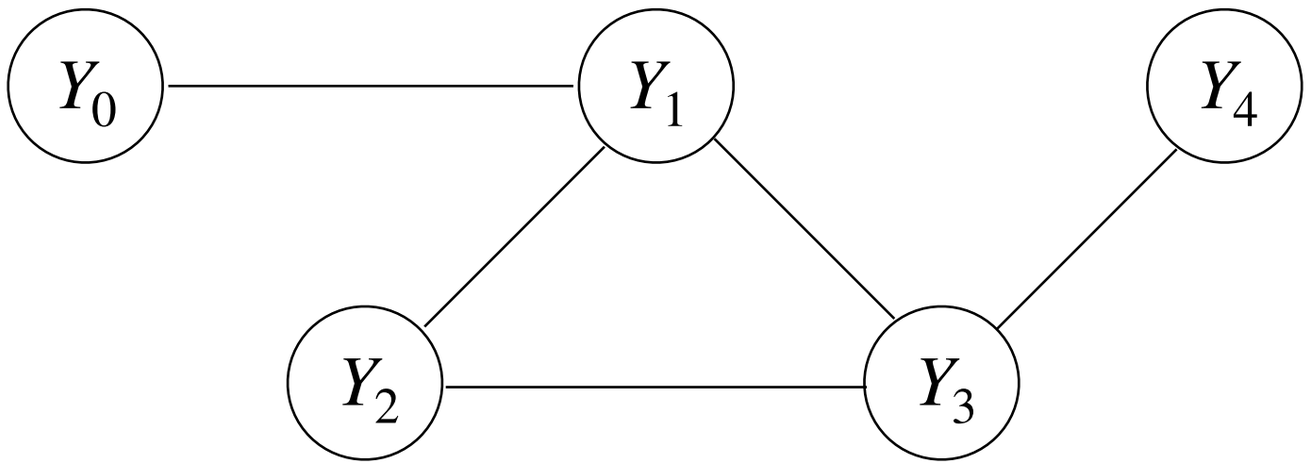

图7：最大团。

Hammersley-Clifford定理

概率无向图模型的联合概率分布$P(Y)$可以定义为如下形式：

$$P(Y)=\frac{1}{Z}\prod_{C}\Psi_C(Y_C)$$

归一化因子定义如下：

$$Z=\sum_{Y}\prod_{C}\Psi_C(Y_C)$$

其中，C是无向图中的最大团，Y_C是C中节点对应的随机变量组，$\Psi_C(Y_C)$是在C上定义的严格正函数，被称为势函数。乘积是在无向图上所有的最大团上进行的。

势函数的定义可根据实际任务决定，只需满足函数取值严格为正即可。指数函数经常被用于定义势函数。

Hammersley-Clifford定理

以图7为例，联合概率分布可计算如下

$$P(Y) = \frac{1}{Z} \times \Psi_{(0,1)}(Y_0, Y_1) \times \Psi_{(1,2,3)}(Y_1, Y_2, Y_3) \times \Psi_{(3,4)}(Y_3, Y_4)$$

势函数可以根据具体任务进行定义。

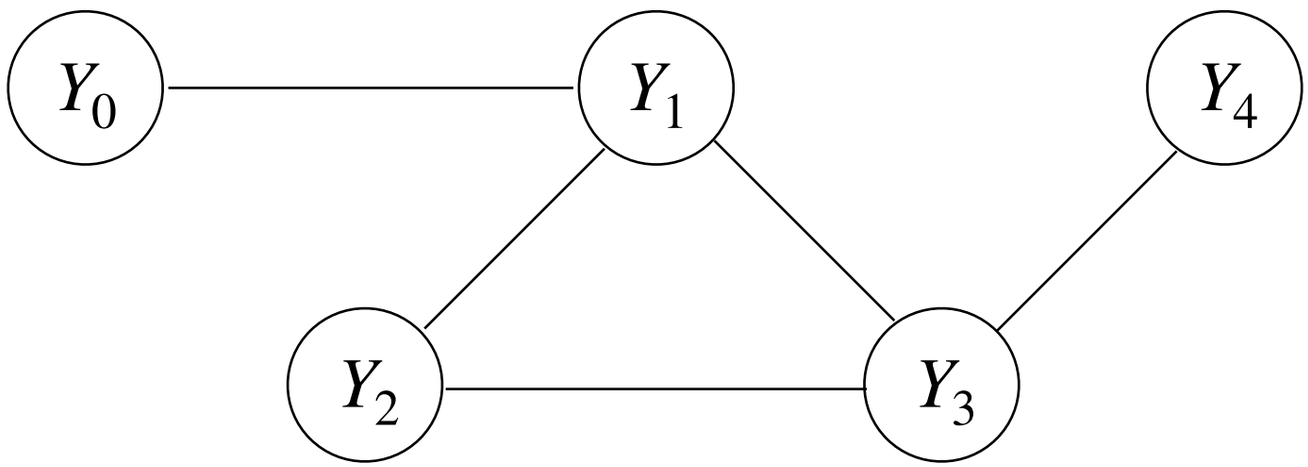

图7：最大团。

条件随机场

条件随机场是研究给定一组输入随机变量，另一组输出随机变量的条件概率分布的模型，其特点是假设输出随机变量构成概率无向图模型。

设X和Y是随机变量，$P(Y|X)$是给定X的条件下Y的条件概率分布。如果随机变量Y构成一个由无向图$G=(V,E)$表示的马尔科夫随机场，即

$$P(Y_u|X,Y_O)=P(Y_u|X,Y_W) \quad \text{(公式9)}$$

对于任意节点u成立，则称条件概率分布$P(Y|X)$为条件随机场。

在公式9中，$O=V-\{u\}$表示图中除了节点u以外所有节点的集合，Y_O表示与之对应的随机变量组，$W=\mathcal{N}(u)$表示图中所有与节点u有边相连的节点（即邻居节点）的集合，Y_W表示与之对应的随机变量组。

需要注意的是，输入变量X变不属于无向图的一部分。

线性链条件随机场

设$X=(X_1,\ldots,X_n)$与$Y=(Y_1,\ldots,Y_n)$均为线性链表示的随机变量序列，若在给定随机变量序列X的条件下，随机变量序列Y的条件概率分布$P(Y|X)$构成条件随机场，即满足马尔科夫性：

$$P(Y_i|X,Y_1,\ldots,Y_{i-1},Y_{i+1},\ldots,Y_n)=P(Y_i|X,Y_{i-1},Y_{i+1})$$

则称$P(Y|X)$为线性链条件随机场。

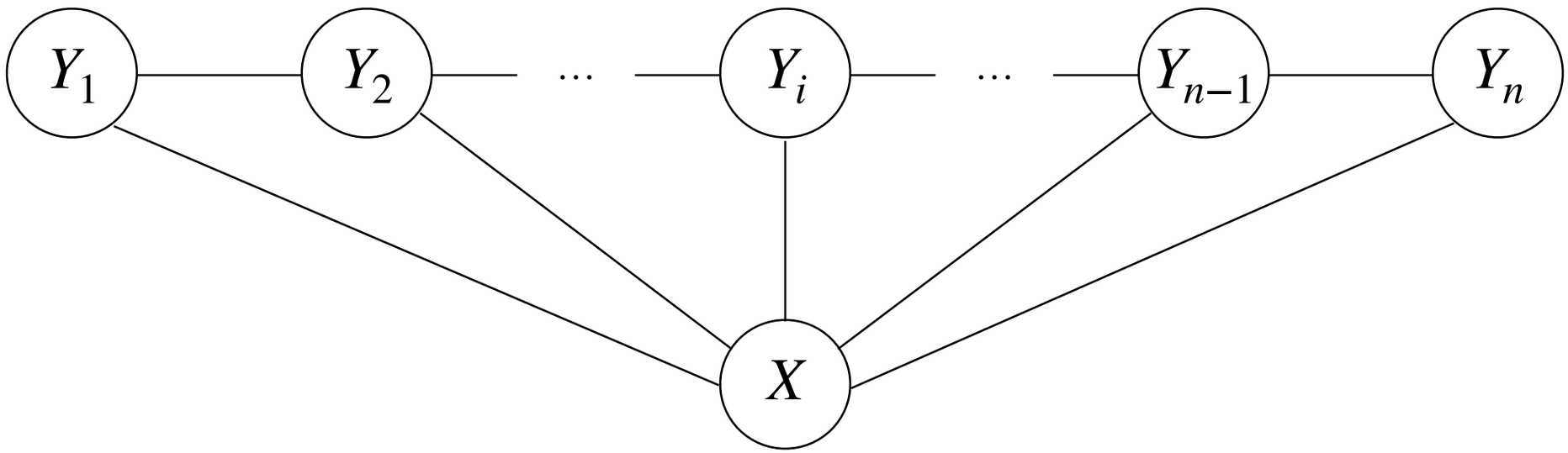

图8：线性链条件随机场。

线性链条件随机场的参数化形式

设$P(Y|X)$为线性链条件随机场，则在随机变量X取值为x的条件下，随机变量Y取值为y的条件概率具有以下形式：

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x_i) \right)$$

其中，归一化因子定义如下：

$$Z(x) = \sum_{y} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i) \right)$$

$t_k(y_{i-1}, y_i, x_i)$是定义在边上的转移特征，依赖于当前位置和前一个位置，$s_l(y_i, x, i)$是定义在节点上的状态特征，依赖于当前位置，λ_k是转移特征权重，μ_l是状态特征权重。

条件随机场的简化形式

由于条件随机场中的同一特征在各个位置上都有定义，可以对同一特征在各个位置求和，将局部特征函数转化为一个全局特征函数，这样就可以把条件随机场写成权重向量与特征向量的内积形式：

$$P(y \mid x; \mathbf{w}) = \frac{\exp(\mathbf{w} \cdot F(x, y))}{Z(x, \mathbf{w})}$$

其中归一化因子可定义为：

$$Z(x, \mathbf{w}) = \sum_{y} \exp\Big(\mathbf{w} \cdot F(x, y)\Big)$$

我们使用$F(x, y)$来表示全局特征向量。从上式可以看出，条件随机场本质上也是一个对数线性模型，在推断时依然基于线性模型$\mathbf{w} \cdot F(x, y)$。虽然与最大熵模型不同，条件随机场使用局部特征，但是局部特征依然可以转化为等价的全局特征。

总结

- 与以隐马尔科夫模型为代表的生成式模型不同，支持向量机和对数线性模型都属于判别式模型，其核心是特征权重向量与特征函数向量的内积。
- 支持向量机只能给出分类结果，无法给出分类概率。对数线性模型通过指数函数可以给出分类概率。此外，最大熵模型和条件随机场显式给出了特征函数的概念，为向模型中注入人类先验知识提供了良好的接口。
- 对数线性模型的关键之处在于需要根据具体任务的特点设计特征函数，如何确保人工设计的特征函数（即特征工程）能够覆盖所有的现象和规律成为最大的挑战。

谢谢