模式识别和计算机视觉

HMM: Hidden Markov Model

隐马尔科夫模型

张振宇

南京大学智能科学与计算学院

2025

人脸识别上的应用

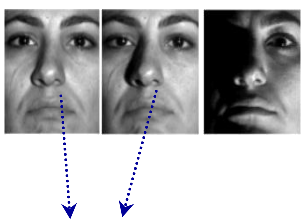
- 对全班150名同学采集人脸照片，每人100张，分辨率100*100
- 每张图片拉成向量，维度$p = 10000$
- 每人每张图片向量依次排列，得到字典项$k = 15000$
- 得到过完备字典D，大小$p \times k$
- 给定某个ID图像向量$\boldsymbol{x}_i$，求解

$$\min_{\alpha_i} \|\boldsymbol{x}_i - D\boldsymbol{\alpha}_i\|^2 + \lambda\|\boldsymbol{\alpha}_i\|_1$$

人脸识别上的应用

✓ 为什么求解$\min_{\alpha_i} \|\boldsymbol{x}_i - D\boldsymbol{\alpha}_i\|^2 + \lambda\|\boldsymbol{\alpha}_i\|_1$可以实现人脸识别？

- 同ID一致性假设
- 找到对应的稀疏系数所属ID
- 非零系数在某个ID内明显最多，测试图像则属于该类



$$\boldsymbol{A}_i = [\ \ |\ \ |\ \ \ldots\ ] \in \mathbb{R}^{m \times n_i}$$

$\boldsymbol{y}$  $\approx\ x_{i,1}$  $+\ x_{i,2}$  $+ \ldots +\ x_{i,n}$  $= \boldsymbol{A}_i \boldsymbol{x}_i$

字典学习

- 从数据集中学习一个字典D，使得所有样本$\{\boldsymbol{x}_i\}$能被字典中的原子稀疏表示，即

$$\forall i,\ \mathbf{x}_i \approx \mathbf{D}\alpha_i, \quad \|\alpha_i\|_0 \leq s.$$

- 需要联合优化字典和系数

$$\min_{\mathbf{D},\{\alpha_i\}} \sum_i \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \sum_i \|\alpha_i\|_1$$

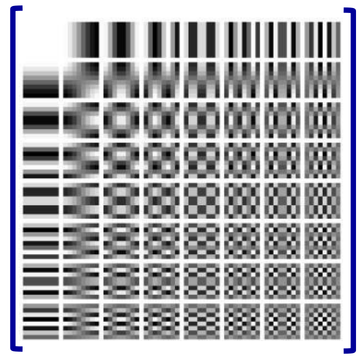
应用场景

✓ 图像压缩

Compression – JPEG

$\boldsymbol{y}$

(Patches of) …
input image



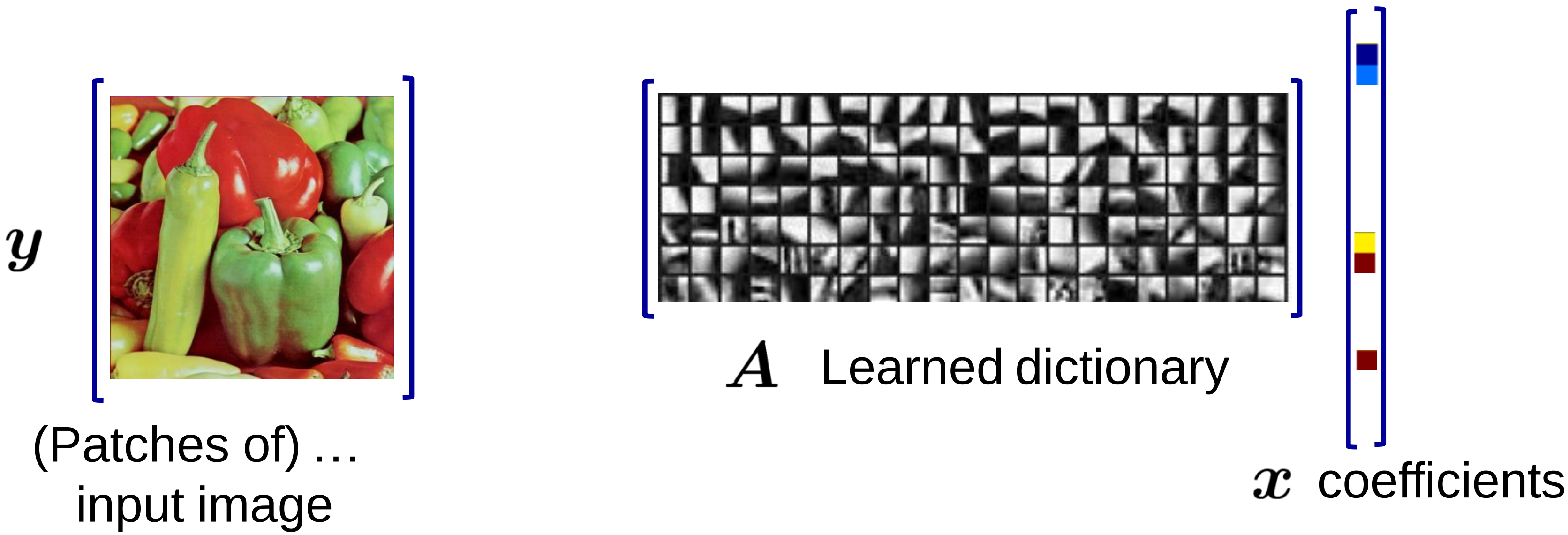
$\boldsymbol{A}$ DCT basis

$\boldsymbol{x}$ coefficients

应用场景

✓ 图像压缩

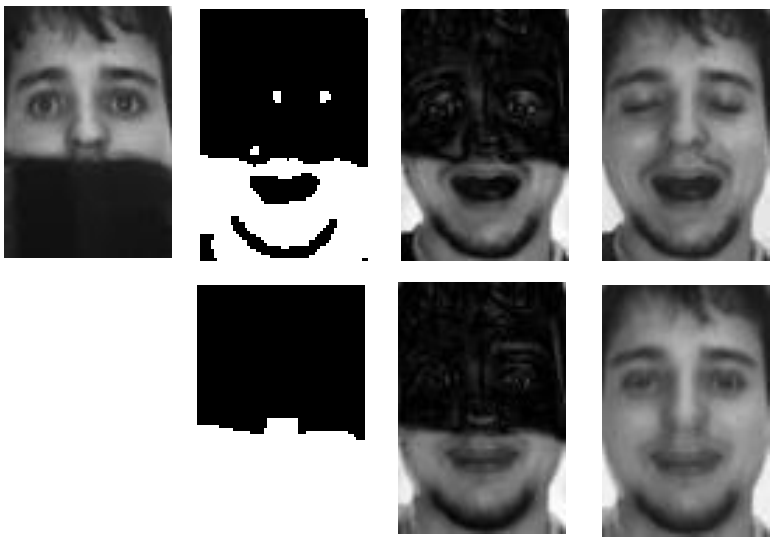
Compression – Learned dictionary



稀疏向量的推广

✓ 如果在二维结构上具有稀疏性，会有怎样的特点？

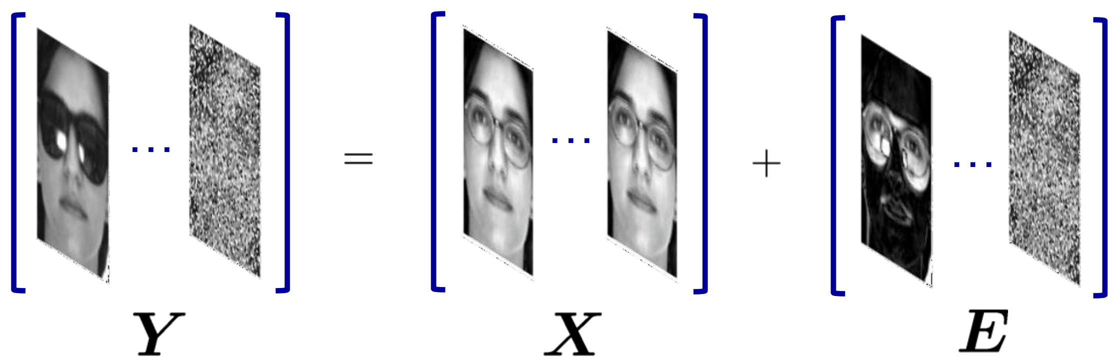
- 低秩矩阵（Low-Rank Matrix）



稀疏向量的推广

✓ 如果在二维结构上具有稀疏性，会有怎样的特点？

- 低秩矩阵（Low-Rank Matrix）

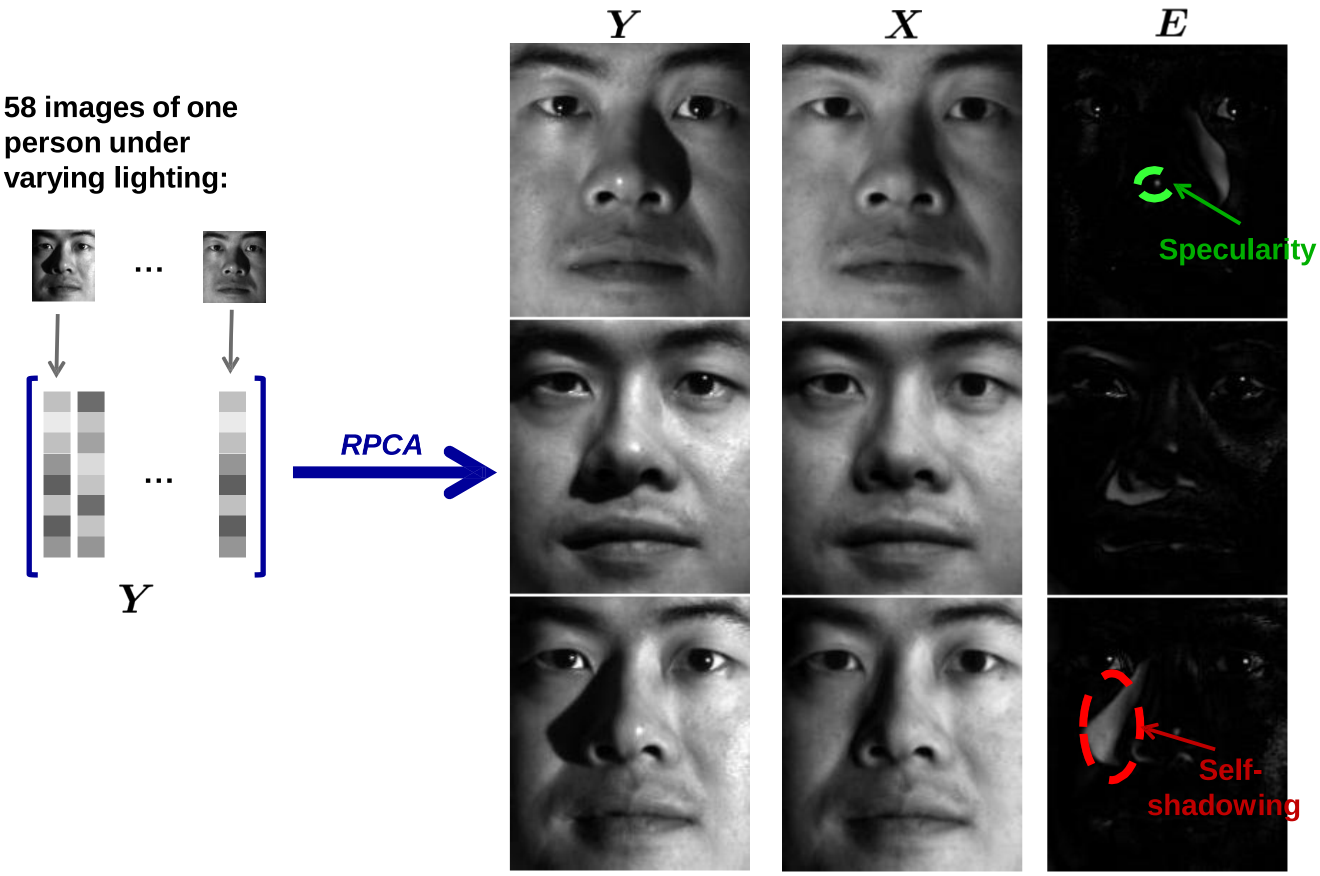


Given $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}$, with $\boldsymbol{X}$ low-rank, $\boldsymbol{E}$ sparse, recover $\boldsymbol{X}$.

稀疏向量的推广

✓ 如果在二维结构上具有稀疏性，会有怎样的特点？

- 低秩矩阵（Low-Rank Matrix）



目标

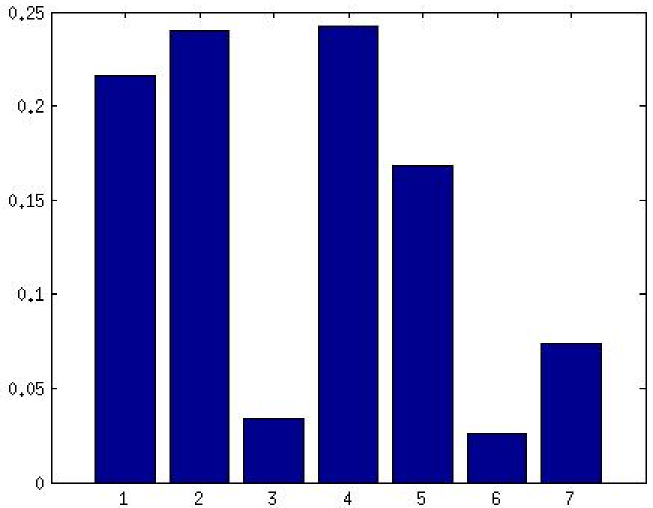
- ✓ 掌握随机过程和马尔科夫性质的基本概念
- ✓ 掌握隐马尔科夫模型（离散观测值）的应用条件和相关推理算法
- ✓ 了解隐马尔科夫模型（离散观测值）的学习算法

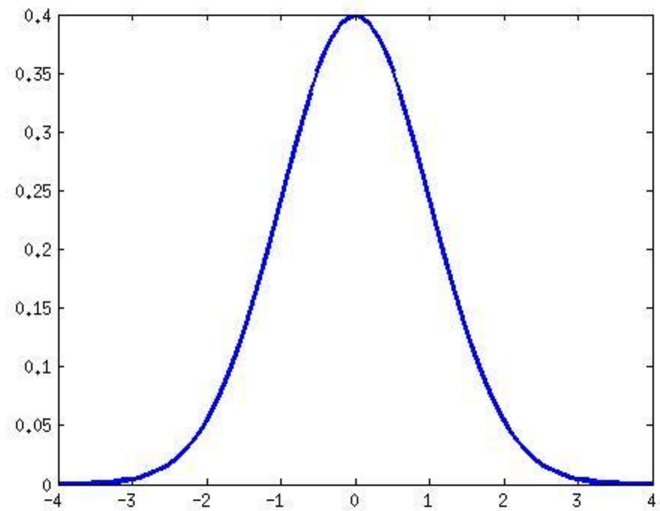
- ✓ 提高目标
  - 进一步能通过独立阅读、了解HMM的实际应用
  - 进一步能通过独立阅读、了解基本的图模型graphical model的概念、 belief propagation （BP）算法

Markovian

随机变量(Random variable)

- ✓ X: 可以是离散(discrete)、连续(continuous)、或者混合(hybrid)的





随机变量

- 目的是得到映射：$\mathcal{X} \mapsto \mathcal{Y}$
  - 数据分布$p(\mathcal{X})$
  - 先验分布 prior distribution $p(\mathcal{Y})$
    - *a priori*: Knowable without appeal to particular experience
    - a priori distribution: special meaning, do not misuse
  - 联合joint分布$p(\mathcal{X}, \mathcal{Y})$
  - 类条件分布$p(\mathcal{X}|y = i)$
  - 后验分布posterior distribution $p(y = i|\boldsymbol{x})$

贝叶斯参数估计

- Bayesian parameter estimation
  - MLE：视$\boldsymbol{\theta}$为固定的参数，假设存在一个最佳的参数（或参数的真实值是存在的），目的是找到这个值
  - MAP：将$p(\boldsymbol{\theta})$的影响代入MLE中，仍然假设存在最优的参数
  - 以上均称为点估计point estimation
- 在贝叶斯观点中，$\boldsymbol{\theta}$是一个分布/随机变量，所以估计应该是估计一个分布，而不是一个值（点）！
  - $p(\boldsymbol{\theta}|D)$：这是贝叶斯参数估计的输出，是一个完整的分布，而不是一个点

之前接触到的随机变量

- ✓ 静态的，没有时序索引概念
- ✓ 通常研究**独立同分布（i.i.d.）**或少量变量间的相关性（如协方差）
- ✓ 关注单个或有限个变量的分布性质
- ✓ 研究有限维联合分布（如二维正态分布）

HMM中的随机变量

- ✓ 随时间、空间索引不断变化
- ✓ 关注**时间或空间上的依赖结构**
- ✓ 研究**动态演化规律**和**极限行为**
- ✓ 需指定**所有有限维联合分布**的一致性

随机过程stochastic process

- A stochastic process $\{X(t), t \in T\}$ is a collection of random variables. That is, for each $t \in T$, $X(t)$ is a random variable. The index t is often interpreted as time and, as a result, we refer to $X(t)$ as the *state* of the process at time t.
  - The set T is called the *index* set of the process.
  - When T is a countable set … a discrete-time process.
  - If T is an interval of the real line, … a continuous-time process.
  - The state space (状态空间) of a SP is defined as the set of all possible values of that random variables $X(t)$ can assume.
- A SP is a family of random variables that describes the evolution through time of some (physical) process.

S. M. Ross, "Introduction to Probability Models", 8th ed.

时间序列Times Series

- ✓ 随机过程$\{X_1, X_2, \dots\}, X_i \in \mathcal{X}$
  - $\mathcal{X}$称为状态空间，我们假设$\mathcal{X} = \{1,2,\dots,N\}$
  - 假设对所有的i，$\mathcal{X}$都相同
  - 假设只处理时间序列，即i代表时间
  - 随机性的优缺点
- ✓ 目的是希望“过去”对“现在”有帮助
  - 即如果有对$X_1, \dots, X_{t-1}$的了解，能帮助确定X_t
  - Formally, $P(X_t | X_{1:t-1})$ vs. $P(X_t)$

Markov Property

- ✓ Curse of dimensionality
  - $P(X_2|X_1)$需要多少存储空间才能指定?
  - $P(X_3|X_2,X_1)$需要多少存储空间才能指定?
  - $P(X_t|X_{1:t-1})$需要多少存储空间才能指定?
    - N^t!
- ✓ Markov Property马尔科夫性质
  - 限定：$P(X_t|X_{1:t-1}) = P(X_t|X_{t-1})$，含义是?
  - 无记忆性memoryless
  - 这个假设有效吗?
  - 好处是什么?

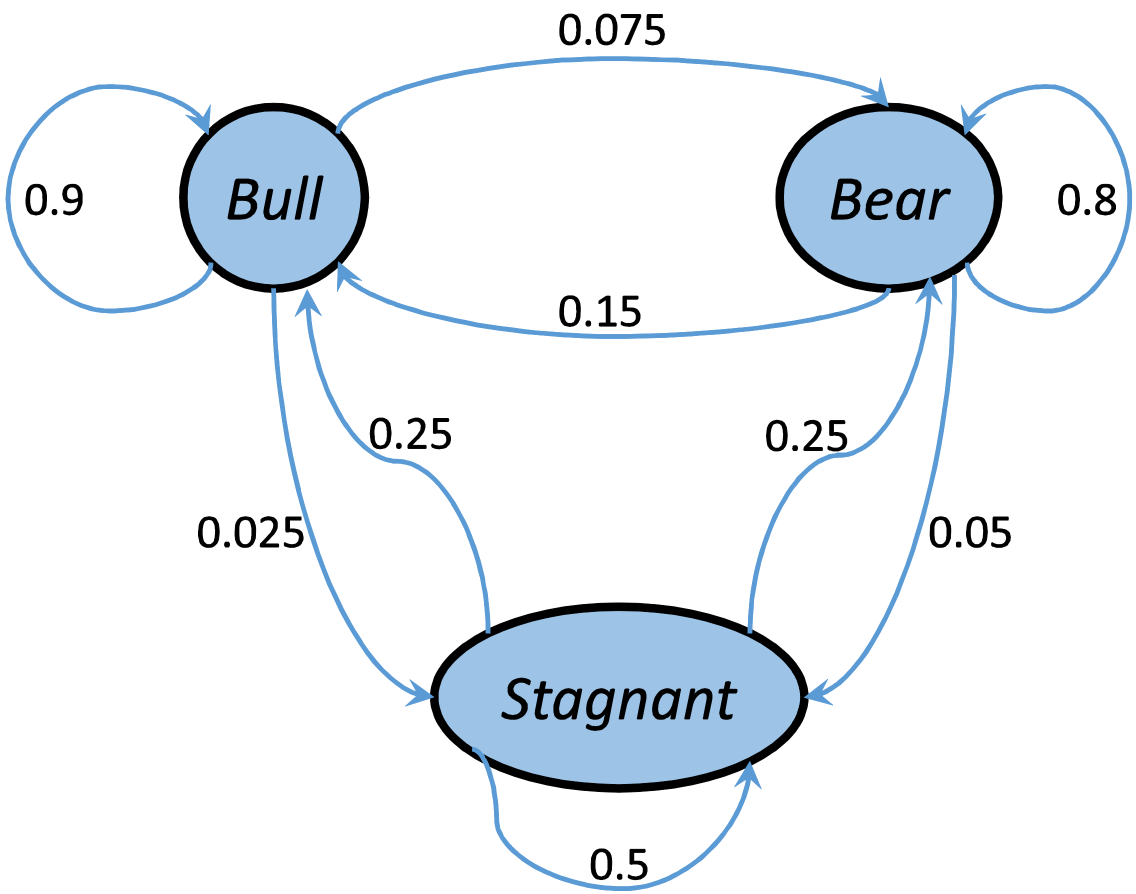
Andrey Markov

http://en.wikipedia.org/wiki/Andrey_Markov
Retrieved Jan 15 2014

- Chebyshev–Markov–Stieltjes inequalities
- Gauss–Markov theorem
- Gauss–Markov process
- Hidden Markov model
- Markov blanket
- Markov chain Monte Carlo
- Markov decision process
- Markov's inequality
- Markov information source
- Markov network
- Markov number
- Markov property
- Markov process
- Subjunctive possibility

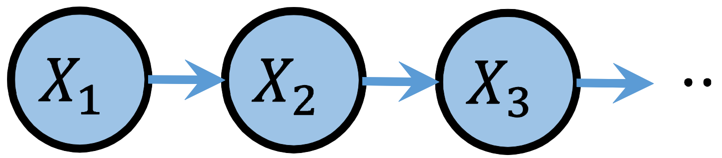
Markov Chain马尔科夫链

✓ Markov chain (discrete-time Markov chain or DTMC)



可视化和形式化

✓ 可视化：



- 注意填充的变量表示观察值（即随机变量值已知）

✓ 那么，如何形式化定义DTMC？需要哪些量？

- 系统初始化Initialization: $P(X_1)$或者$X_1 = x_1$
- Transition probability: $A = P(X_{t+1}|X_t)$
- 还需要别的吗？
- 两次运行结果会一样吗？

转移概率矩阵

- ✓ Transition probability matrix转移概率矩阵
  - A是一个$N \times N$的矩阵
  - $A_{ij} = P(X_t = j | X_{t-1} = i)$
  - 行和为1！

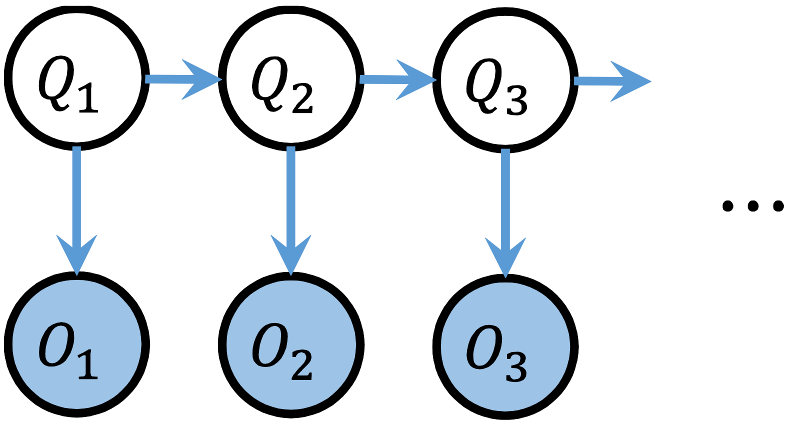
- ✓ 如果运行足够久（$t \to \infty$），那么X_t的分布在很多情况下将稳定下来，叫Stationary distribution，记为$\boldsymbol{\pi}$
  - $\boldsymbol{\pi} = A\boldsymbol{\pi}$
- ✓ Google PageRank的简化!

Hidden Markov Model

怎样在模式识别中发挥更大作用？

✓ 例如，在连续手写识别中，笔画stroke没有在DTMC里面用到，那么DTMC就没法用于连续手写识别

- 想办法把笔画加进去！
- 状态是什么？



形式化

- ✓ Q：隐变量(hidden variable)，不可观测的状态
- ✓ N：number of states 状态数,N个可能的状态为$\{S_1, \dots, S_N\}$
- ✓ $O(o)$：观察值(observations)，M个可能的观察值$\{V_1, V_2, \dots, V_M\}$
- ✓ T：时间序列的长度
- ✓ $\boldsymbol{\pi}$：初始化, $\pi_j = P\left(Q_1 = S_j\right)$
- ✓ A：transition probability matrix, $A_{ij} = P(Q_{t+1} = S_j | Q_t = S_i)$
- ✓ B：emission probability 发出观察值的概率（发射概率）
  - $b_j(k) = \Pr(O_t = V_k | Q_t = S_j)$
  - 假设B不随时间变化，当未知状态为j时观察到为k的概率
  - 那么，j，k的取值范围是？B的行和是？

HMM中要解决的问题

- 怎样设计状态？ -- 自动学习？
- 怎样设计观察值？ --根据问题的特点和实践反复设计

- 与具体问题无关的
  - 指定一个HMM需要的所有参数：$\lambda = (\boldsymbol{\pi}, A, B)$
  - 问题1： Evaluation估值
  - 问题2： Decoding解码
  - 问题3： Learning学习

Problem 1. Evaluation

- ✓ 输入
  - 一个完全指定的HMM模型，即$\lambda = (\boldsymbol{\pi}, A, B)$已知
  - 一个完全观测的输出序列$O_1 O_2 \cdots O_T$，或$\boldsymbol{O} = O_{1:T}$
- ✓ 输出
  - $P(\boldsymbol{O}|\lambda)$ – 含义是？
  - 在这个模型λ中观察到特定输出$\boldsymbol{O}$的概率
- ✓ 作用是？
  - 可以看出此模型对该观察序列的成绩score
  - 可以用来从多个模型中选择最适合的模型

Problem 2： Decoding

- 输入
  - 一个完全指定的HMM模型，即$\lambda = (\boldsymbol{\pi}, A, B)$已知
  - 一个完全观测的输出序列$O_1 O_2 \cdots O_T$，或$\boldsymbol{O} = O_{1:T}$
  - 某个标准criterion
- 输出
  - 一个完全指定的隐变量序列$X_{1:T}$的值
- 作用是？
  - 如，语音识别中状态可能有实际意义（各音节）
    - 唯一吗？
  - 可以用来观察模型结构，优化模型

Problem 3： Learning

- 输入
  - 网络结构，状态数、输出数
  - 若干观测序列$\{\boldsymbol{O}\}$
- 输出
  - 最优的参数$\lambda = (\boldsymbol{\pi}, A, B)$使得$P(\{\boldsymbol{O}\}|\lambda)$最大
- 作用
  - 显而易见
  - 最重要的问题
  - 有时候一个足够长的观测序列就够了

Evaluation

假设隐状态已知

- ✓ 已知$\lambda, o_{1:T}$，求$P(o_{1:T}|\lambda)$
- ✓ 若假设oracle已告知所有的隐变量的值$q_{1:T}$
  - $\Pr(o_{1:T}|\lambda, q_{1:T}) = \prod_{i=1}^{T} \Pr(o_t|q_t, \lambda) = \prod_{i=1}^{T} b_{q_i}(o_i)$
  - 证明？含义？
  - λ的存在只是表明概率的大小是基于该模型参数计算的，可以去除而不影响计算
- ✓ 关于各随机变量之间的独立性的判断，进一步参阅PRML第八章

一种naïve的计算方法

- ✓ 那么隐变量序列$q_{1:T}$的可能性多大呢？
  - $\Pr(q_{1:T}|\lambda) = \pi_{q_1} A_{q_1 q_2} A_{q_2 q_3} \cdots A_{q_{T-1} q_T}$
  - 含义？
- ✓ 用全概率公式对所有可能的$q_{1:T}$求和可以得到$\Pr(o_{1:T}|\lambda)$
  - $\Pr(o_{1:T}|\lambda) = \sum_{all\ Q} \Pr(o_{1:T}|\lambda, q_{1:T}) \Pr(q_{1:T}|\lambda)$ ，复杂度？
  - $O(T \times N^T)$
- ✓ 虽然不实用，但可以从中学到一种思考问题的方法
  - 后面EM学习算法用相似的思路

那么，如何快速计算？

- ✓ 动态规划！
- ✓ 只看最后一步（$t = T$)，该如何计算?
  1. 最后一步（$t = T$)时一共可能有N种状态：$q_T = S_1, \ldots, S_N$，其概率$\Pr(o_{1:T-1}, Q_T = S_i \mid \lambda) =?$
  2. 若最后一步状态为S_i, 那么观察到输出o_T的概率是多少？
  3. 所求的值是多少？(全概率公式)

$$\Pr(o_{1:T} \mid \lambda) = \sum_{i=1}^{N} \Pr(o_{1:T-1}, Q_T = S_i \mid \lambda) b_i(o_T)$$

- 只限于最后一步吗？

快速计算（2）

✓ 如何计算 $\Pr(o_{1:T-1}, Q_T = S_i|\lambda)$？

- 有N种可能，即$T-1$时刻状态为$q_{T-1} = S_j$，$j = 1,2,\ldots,N$，然后通过概率A_{ji}转移

- 全概率公式，again！

$$\Pr(o_{1:T-1}, Q_T = S_i|\lambda) = \sum_{j=1}^{N} \Pr\left(o_{1:T-1}, Q_{T-1} = S_j \middle| \lambda\right) A_{ji}$$

快速计算小结

- ✓ $\Pr(o_{1:T}|\lambda) = \sum_{i=1}^{N} \Pr(o_{1:T-1}, Q_T = S_i|\lambda\,)b_i(o_T) = \sum_{i=1}^{N}\left(b_i(o_T)\sum_{j=1}^{N} \Pr\left(o_{1:T-1}, Q_{T-1} = S_j|\lambda\right)A_{ji}\right)$
- ✓ 红色部分是什么？
  - 一个规模小一点的相同问题（$T-1$）
  - 但是需要对所有j的可能取值计算
  - 正如DTW中一样，可以通过动态规划解决，但是需要解决比原问题更多数目的小规模子问题
  - 但是，复杂的是，目前牵涉两个数值而不是一个：$\Pr(o_{1:T-1}, Q_T = S_i|\lambda\,)$和$P(o_{1:T}|\lambda)$
  - 计算的方向应该是什么？

进一步的阅读

✓ 如果对本章的内容感兴趣，可以参考如下文献

- HMM Tutorial: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=18626&tag=1
- HMM software: HTK in http://htk.eng.cam.ac.uk/
- PRML 13.1, 13.2
- Graphical model: PRML chapter 8, 9, 10, 11, 12, 13
- Graphical model: "Probabilistic Graphical Models: Principles and Techniques" by Daphne Koller and Nir Friedman; http://mitpress.mit.edu/books/probabilistic-graphical-models