CSDN
不 止 于 代 码

第二讲：多元随机变量及其分布

CSDN学院

2017年7月

AI

▶复习

- 1. 怎样识别数据中是否有outliers? Outliers可以怎么处理？
- 2、直方图（histogram）和箱体图（boxplot）有什么区别？

Roadmap

- 随机向量的分布
  - 联合概率、边缘概率、
  - 随机变量之间的关系：独立、条件概率、条件独立、相关性（协方差）、互信息
  - 案例：随机变量之间的关系分析
- 多元正态分布
  - 线性判别分析
- 概率图模型
  - 朴素贝叶斯、Markov 链、HMM、MRF、CRF

一、随机向量及其分布

多元随机向量的分布

- 我们可以在多个随机变量组成的向量上定义分布，称之为多元随机向量的分布。

- 机器学习中我们的数据集通常由多元随机向量分布的样本组成。每一列为一个随机变量。

联合分布

- 对D维随机向量$(X_1,...,X_D)$
- 若X_j为离散型随机变量，则定义联合概率质量函数(pmf)为：

$$p(x_1,...,x_D)=\mathbb{P}(X_1=x_1,...,X_D=x_D)$$

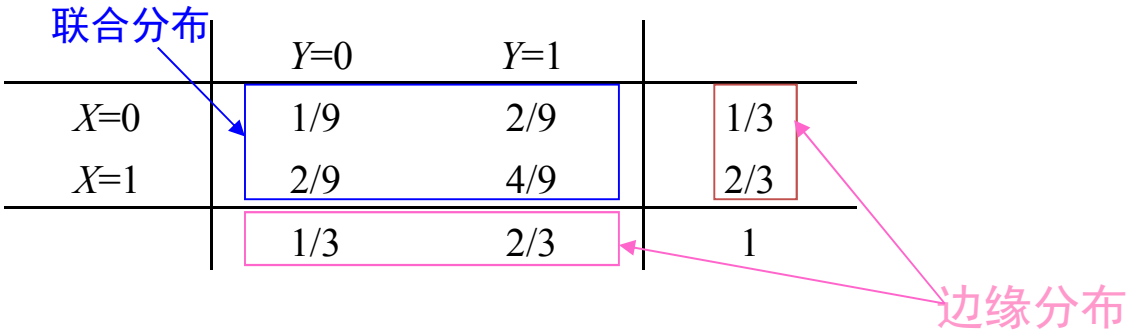
- 联合概率分布函数(CDF)为：

$$F(x_1,...,x_D)=\mathbb{P}(X_1\le x_1,...,X_D\le x_D)$$

- 若单维随机变量X_j需要的参数为K_j，则描述联合分布需要的参数为：$\prod_{j=1}^{D}K_j$

▶例：

- 例：如下有两维随机向量$(X,\ Y)$, 其中取值为0或1，



- 则 $p(1,1)=\mathbb{P}(X=1,Y=1)=4/9$

联合分布

- 对D维随机向量$(X_1,...,X_D)$
- 若X_j为连续型随机变量，则定义联合概率密度函数(pdf)为
  - $p(x_1,...,x_D) \geq 0,\ \forall x_1,...,x_D$
  - $\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty} p(x_1,...,x_D)dx_1...dx_D = 1$
  - 对任意集合 $A \subset \mathbb{R}^D, \mathbb{P}((X_1,...,X_D) \in A) = \int...\int_A p(x_1,...,x_D)dx_1...dx_D$
- 联合概率分布函数(CDF)为：

$$F(x_1,...,x_D) = \mathbb{P}(X_1 \leq x_1,...,X_D \leq x_D)$$

边缘分布

- 对离散型随机变量，如果$(X_1,...,X_D)$有联合密度函数$p(x_1,...,x_D)$，则X_j的边缘密度函数定义为

$$p(x_j)=\mathbb{P}(X_j=x_j)=\sum_{x_k,k=1,...D,k\neq j}\mathbb{P}(X_1=x_1,...,X_k=x_k,...,X_D=x_D)$$

- 对连续情况：

$$p(x_j)=\int...\int p(x_1,...,x_D)dx_1...dx_{j-1}dx_{j+1}...dx_D$$

联合分布包含了随机向量概率分布的信息
联合分布唯一确定了边缘分布，但反之通常不成立

条件分布

- 对二维随机变量(X,Y), 当 $p(y)>0$ 时，定义给定$Y=y$时X的条件分布为

$$p(x|y)=\frac{p(x,y)}{p(y)}$$

- 注意：$p(x,y)=p(x|y)p(y)=p(y|x)p(x)$

条件概率 → 链规则（Chain Rule）

- $p(x,y)=p(x|y)p(y)$
- 链规则: $p(x,y,z)=p(x|y,z)p(y,z)$
  $=p(x|y,z)p(y|z)p(z)$
- 或 $p(x,y,z)=p(x)p(y|x)p(z|x,y)$
- 一般地，$p(x_1,...,x_D)=p(x_1)\times p(x_2,...,x_D|x_1)$
  $=p(x_1)\times p(x_2|x_1)\times p(x_3,...,x_D|x_1,x_2)$
  $=p(x_1)\prod_{j=2}^{N}p(x_i|x_1,...,x_{i-1})$

贝叶斯规则

- 全概率公式：如果Y可以取值$y_1,\ldots,y_K$，x为X的一个取值，

$$p(x)=\sum_j p(x|y_j)p(y_j)$$

- 因此，有贝叶斯规则

$$p(y_k|x)=\frac{p(x|y_k)p(y_k)}{\sum_j p(x|y_j)p(y_j)}$$

似然　先验

后验

分子、分母形式相同

- 连续情况：$p(y|x)=\frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$

例：贝叶斯公式应用-医疗诊断

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$$

- 例：对疾病D的医学测试结果输出为+和-。
- 现有比较灵敏的测试，在得病的情况下检测为+的概率为

$$p(x=1 \mid y=1)=0.8, \quad p(x=1 \mid y=0)=0.1$$

检验相当正确

- 其中x=1表示测试结果为+，y=1表示得病。
- 假设得病的先验为$p(y=1)=0.004$
- 则假设检测为+，得病的概率为

$$p(y=1 \mid x=1)=\frac{p(x=1,y=1)}{p(x=1)}=\frac{p(x=1 \mid y=1)\,p(y=1)}{p(x=1 \mid y=1)\,p(y=1)+p(x=1 \mid y=0)\,p(y=0)}$$

$$=\frac{0.8\times 0.004}{0.8\times 0.004+0.1\times 0.996}=0.031$$

得病概率很小
不要相信直觉！

例：产生式分类器

- 用于分类任务的分类器可分为两大类：（根据模型中是否用到类条件概率）
  - 产生式分类器: $p(y=c|x)=\frac{p(x|y=c)p(y=c)}{\sum_{c'}p(x|y=c')}$
    - 有了类条件概率 $p(x|y=c)$ 和类先验 $p(y)$，可以得到后验分布，从而产生后验样本
  - 判别式分类器
    - 直接做出判别

独立

- 若对所有的x, y，有 $\mathbb{P}(X=x, Y=y)=\mathbb{P}(X=x)\mathbb{P}(Y=y)$
- 或 $p(x,y)=p(x)p(y)$ PDF可以因式分解
- 则称X与Y独立，记为 $X \perp Y$
  - 不独立：随机变量之间的关系用条件分布描述

- 但无条件独立通常很少→条件独立。

条件独立

- 若对所有的x, y, z，有

$$\mathbb{P}(X=x, Y=y \mid Z=z)=\mathbb{P}(X=x \mid Z=z)\mathbb{P}(Y=y \mid Z=z)$$

- 或

$$p(x,y \mid z)=p(x \mid z)\,p(y \mid z)$$

- 则称X与Y独立，记为$X \perp Y \mid Z$

- 条件独立是贝叶斯网络/概率图模型的基础。

协方差

- 如果随机变量之间不独立，可用协方差/相关系数刻画两个随机变量之间关系强弱

X,Y分别是具有均值μ_X,μ_Y、标准差σ_X,σ_Y的随机变量，

定义X与Y的协方差为

$$Cov(X,Y)=\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]$$

相关系数为

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

$$Cov(X,X)=\mathbb{V}(X)=\sigma_X^2$$

协方差的绝对值很大意味着变量值变化很大，且它们同时距离各自的均值很远

协方差(covariance) /相关系数

协方差满足：

$$Cov(X,Y)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$$

相关系数满足：

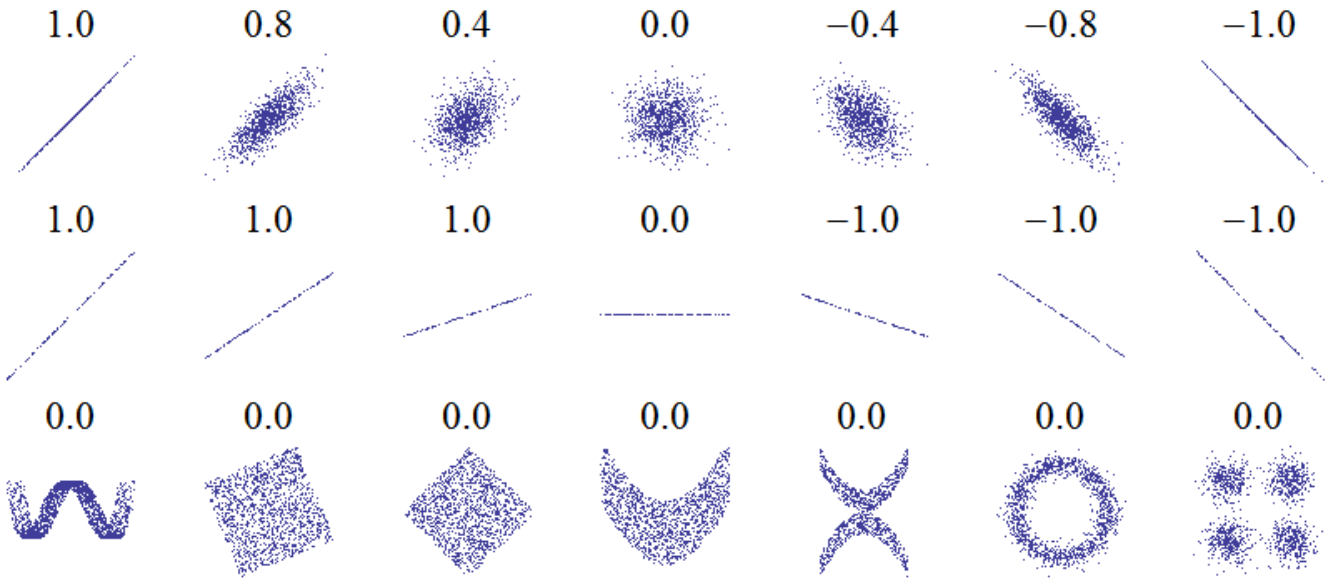
$$-1\le\rho(X,Y)\le 1$$

- X、Y独立，则X、Y不（线性）相关：

$$\mathbb{E}(XY)=\mathbb{E}(X)\mathbb{E}(Y)\ \Rightarrow\ Cov(X,Y)=0$$

  - 但反过来不成立！

相关系数举例



相关系数只对两个实数型随机变量有定义

一种更通用的随机变量之间的关系的度量：度量联合分布 $p(X,Y)$和因式分解形式 $p(X)p(Y)$之间的相似度：互信息

协方差的性质

- 对任意两个随机变量X和Y，有

$$\mathbb{V}(X+Y)=\mathbb{V}(X)+\mathbb{V}(Y)+2Cov(X,Y)$$
$$\mathbb{V}(X-Y)=\mathbb{V}(X)+\mathbb{V}(Y)-2Cov(X,Y)$$

- 当X、Y独立时: $\mathbb{V}(X+Y)=\mathbb{V}(X-Y)=\mathbb{V}(X)+\mathbb{V}(Y)$

- 推广到多个随机变量：

$$\mathbb{V}\left(\sum_i \alpha_i X_i\right)=\sum_i \alpha_i^2 \mathbb{V}(X_i)+2\sum\sum_{i<j}\alpha_i\alpha_j Cov(X_i,Y_j)$$

方差-协方差矩阵

- 令随机向量X的形式为：$X=\begin{pmatrix} X_1 \\ ... \\ X_D \end{pmatrix}$
- 则方差—协方差矩阵记为Σ

$$\Sigma=\begin{pmatrix} \mathbb{V}(X_1) & \mathrm{Cov}(X_1,X_2) & ... & \mathrm{Cov}(X_1,X_D) \\ \mathrm{Cov}(X_2,X_1) & \mathbb{V}(X_2) & ... & \mathrm{Cov}(X_2,X_D) \\ ... & ... & ... & ... \\ \mathrm{Cov}(X_D,X_1) & \mathrm{Cov}(X_D,X_2) & ... & \mathbb{V}(X_D) \end{pmatrix}$$

  - 当各个成分变量独立时，协方差矩阵是什么样子？

21

信息论

- 信息论起源于通讯 / 数据压缩
  - 用最少的比特传递 / 存储信息

- 机器学习：选择最简单的、能表示数据产生规律的模型
  - 模型选择（最小描述长度准则）
  - 特征选择：选择与目标最相关的特征

熵

- 熵是一种不确定度量。假设随机变量X的分布为p，则该随机变量的熵定义为

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k)$$

- 例：若$X \in \{1,\ldots,5\}$, 概率分布为$p = [0.25, 0.25, 0.2, 0.15, 0.15]$
- 则熵 $\mathbb{H}(X) = -2\times 0.25\log_2 0.25 - 0.2\log_2 0.2 - 2\times 0.15\log_2 0.15 = 2.2855$

- 问题：熵最大的分布是？什么时候熵最小？

Kullback-Leibler divergence (KL divergence)

- 一种度量两个分布p和q之间的差异为KL散度：

$$\mathbb{KL}\left(p||q\right) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

- 亦被称为相对熵（ relative entropy ）
- Kl散度还可以写成：

$$\mathbb{KL}\left(p||q\right) = \sum_{k} p_k \log p_k - \sum_{k} p_k \log q_k = -\mathbb{H}\left(p\right) + \mathbb{H}\left(p,q\right)$$

- 其中交叉熵（ cross entropy ）为 $\mathbb{H}\left(p,q\right) \triangleq -\sum_{k} p_k \log q_k$

► 互信息（Mutual information）

- 一种更通用的随机变量之间的关系的度量：度量联合分布 $p(X,Y)$ 和因式分解形式 $p(X)p(Y)$ 之间的相似度：互信息

$$\mathbb{I}(X;Y) \triangleq \mathbb{KL}(p(X,Y)||p(X)p(Y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- 可以证明：$\mathbb{I}(X;Y) \geq 0$
- 当且仅当 $p(X,Y) = p(X)p(Y)$ 即X与Y独立时，互信息 $\mathbb{I}(X;Y) = 0$.

► 互信息（cont.）

- 互信息还可写成

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X)$$

- 其中$\mathbb{H}(Y|X)$为条件熵，定义为：$\mathbb{H}(Y|X) = \sum_x p(x)\mathbb{H}(Y|X=x)$
- 所以互信息可以解释为在观测到X后Y的不确定的减少。

- 在特征选择时，可以通过计算特征与目标之间的互信息，选择与目标互信息最大的那些特征，抛弃与目标关系不大的特征。

最大信息系数

- 连续变量的互信息计算不方便
  - 需先离散化（互信息的结果对离散化的方式敏感）
- 最大信息系数（maximal information coefficient，MIC）寻找最优的离散化方式，并将互信息取值转换成到 [0,1]

$$m(x,y)=\frac{\max_{G\in\mathcal{G}(x,y)}\mathbb{I}(X(G);Y(G))}{\log\min(x,y)},\ \mathrm{MIC}\triangleq\max_{x,y:xy<B}m(x,y)$$

  - 其中$X(G);Y(G)$为某种离散方式，B建议为$N^{0.6}$为（箱子的大小），N为与本数目
- 例：$y=x^2$，MIC算出来的互信息值为1(最大的取值)。

最大信息系数（cont.）

工具包minepy实现了最大信息系数计算

- from sklearn.feature_selection import SelectKBest
- from minepy import MINE

- #由于MINE的设计不是函数式的，定义mic方法将其为函数式的，返回一个二元组，二元组的第2项设置成固定的P值0.5
- def mic(x, y):
    m = MINE()
- m.compute_score(x, y)
- return (m.mic(), 0.5)

- #选择K个最好的特征，返回特征选择后的数据
- SelectKBest(lambda X, Y: array(map(lambda x:mic(x, Y), X.T)).T, k=2).fit_transform(X_train, y_train)

https://github.com/minepy/minepy

特征选择：互信息

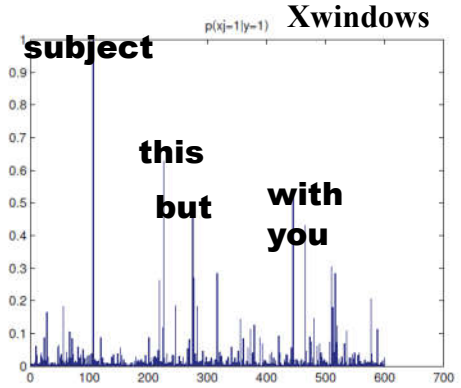
- 例：在分类任务中，多个特征中通常并不是所有的特征都对分类有贡献

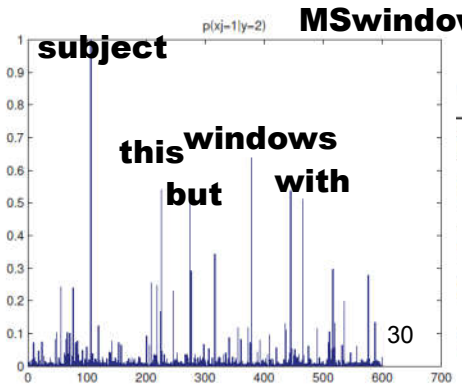
- 解决方案：特征选择
  - 去除不相关的特征，只选择最重要的特征
  - 最简单的方案：单独评估每个特征与目标的相关性，选择最相关的K个特征(**ranking**, **filtering**, or **screening**)

$$I(X,Y) = \sum_{x_j} \sum_{y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)}$$

► 例：特征选择

- 给定文档分类任务，将文档分成class 1 (X windows) and class 2 (MS windows)，特征为600个二维特征（600个词语分别是否在文档中出现）





highest MI	MI
windows	0.215
microsoft	0.095
dos	0.092
motif	0.078
window	0.067

– 互信息高的词语（windows, microsoft）更有判别性

二、多元正态分布

► 多元正态分布（multivariate normal, MVN）

- 正态分布的一般形式：$\mathbf{x}\in\mathbb{R}^D, \mathbf{\Sigma}\in\mathbb{R}^{D\times D}$

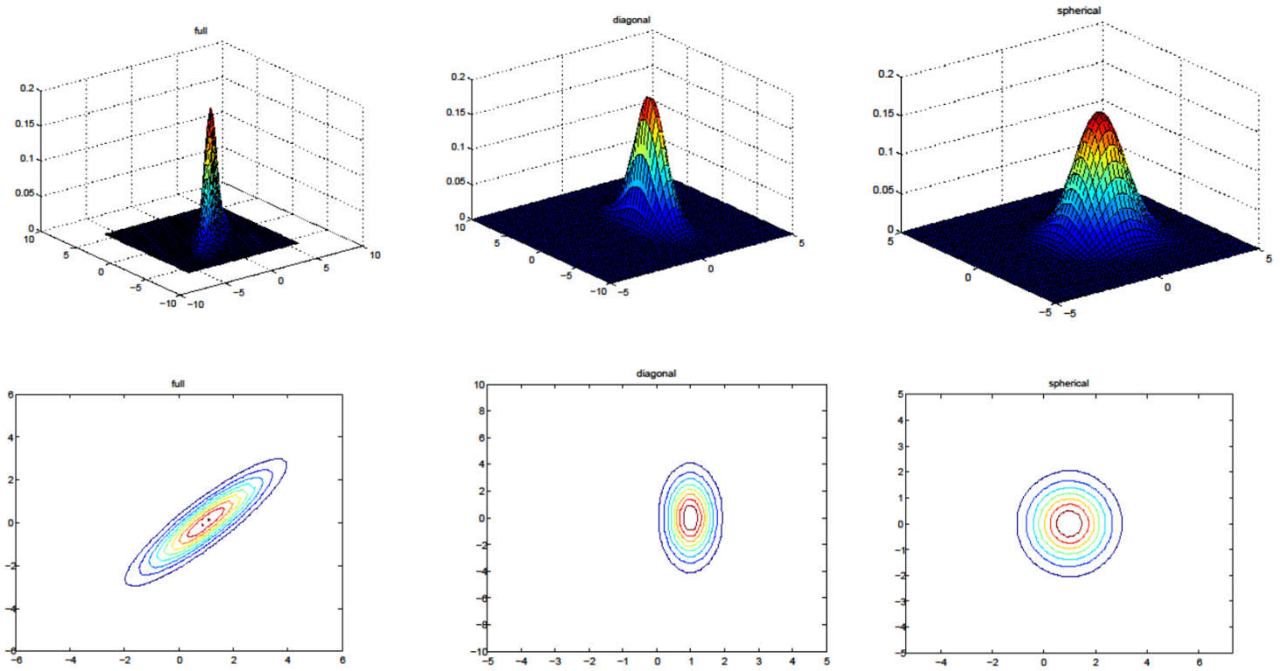
$$\mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu},\mathbf{\Sigma})=\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- $\boldsymbol{\mu}=\mathbb{E}(\mathbf{x}),\quad \mathbf{\Sigma}=\mathbb{E}\left((\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})\right)$
- 协方差(Covariance)还可写成(eg, D=2)

$$\mathbf{\Sigma}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$$

- 协方差矩阵有 $D\times(D+1)/2$ 个独立元素，是正定矩阵
- 协方差的逆 = 精度(precision)

2D高斯分布的pdf



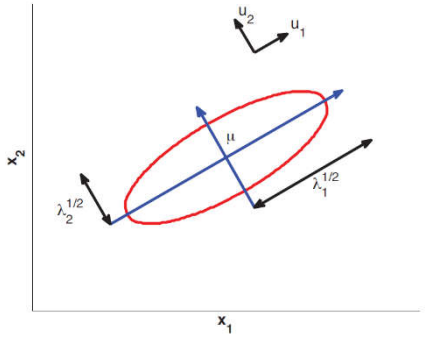
一般情况

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}$$

$$\boldsymbol{\Sigma}=\sigma^2\mathbf{I}$$

协方差的特征值分解

特征值分解：
$\Sigma = U\Lambda U^T$



水平集(Level set):
Mahalanobis dist = const

高斯分布的概率密度等高线轮廓为椭圆曲线

$$\frac{y_1^2}{\lambda_1}+\frac{y_2^2}{\lambda_2}=1$$

$$\Sigma^{-1} = U^{-T}\Lambda^{-1}U^{-1} = U\Lambda^{-1}U^T = \sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T$$

- Mahalanobis 距离：等于在翻转坐标系中的欧氏距离

$$(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^T\sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})$$

$$= \sum_{i=1}^{D}\frac{1}{\lambda_i}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i\mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu}) = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}$$

where $y_i = \mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})$

根据特征向量旋转　　将中心平移到均值

MVN的白化

- 令 $\mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma})$，则 $\mathbf{y} = \Lambda^{-\frac{1}{2}}\mathbf{U}^T\mathbf{x}$ 的分布为标准正态分布 $\mathcal{N}(0, \mathbf{I})$
- 称$\boldsymbol{y}$ 已经被白化(whitened)了。

- 常用的数据预处理方法之一

高斯判别分析

- 高斯分布的一个重要应用是在产生式分类器 (称为高斯的判别分析，GDA)

$$p(y=c|\mathbf{x},\boldsymbol{\theta}) \propto p(\mathbf{x}|y=c,\boldsymbol{\theta})\, p(y=c|\boldsymbol{\theta})$$

- 中作为类条件分布：$p(\mathbf{x}|y=c,\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$
  - 当协方差矩阵为对角阵时，为朴素贝叶斯分类器（在给定类别的情况下，各特征独立）

- Recall：类先验为 $p(y|\boldsymbol{\theta}) = \mathrm{Cat}(y|\boldsymbol{\pi})$

高斯判别分析(cont.)

$$p(y=c|\mathbf{x},\boldsymbol{\theta}) \propto p(\mathbf{x}|y=c,\boldsymbol{\theta})\, p(y=c|\boldsymbol{\theta})$$

- 决策规则为MAP： $\hat{y}(\mathbf{x}) = \arg\max_y p(y|\mathbf{x},\boldsymbol{\theta})$

$$= \arg\max_y \left[\log p(y=c|\boldsymbol{\pi}) + \log p(\mathbf{x}|\boldsymbol{\theta}_c)\right]$$

  - 计算 $p(\mathbf{x}|\boldsymbol{\theta}_c)$ ：需计算x到类中心 $\boldsymbol{\mu}_c$ 的Mahalanobis距离

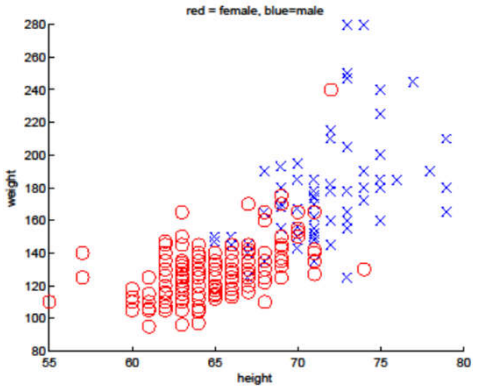
$$\Delta_c = (\mathbf{x}-\boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\mathbf{x}-\boldsymbol{\mu}_c)$$

    - 选择距离最短的类，称为nearest centroids classifier

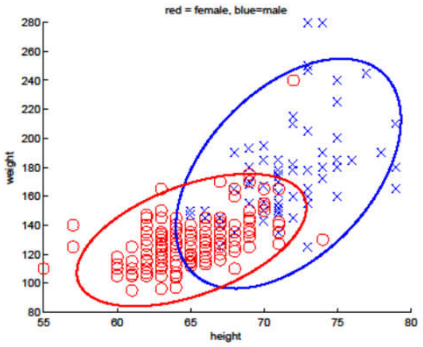
  - 类先验：改变发放的阈值
    - 当类先验为均匀先验时，$\hat{y}(\mathbf{x}) = \arg\min_y \left[(\mathbf{x}-\boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\mathbf{x}-\boldsymbol{\mu}_c)\right]$

高斯判别分析(cont.)

- 例：男生和女生身高和体重
  - 特征身高和体重相关
    - 身高高的人通常体重较重



特征



每个类的高斯估计
椭圆为95%概率

决策边界

- 利用判别MVN进行分类

$$p(y=c|\mathbf{x},\theta)=\frac{\pi_c|2\pi\boldsymbol{\Sigma}_c|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right]}{\boldsymbol{\Sigma}_{c'}\pi_{c'}|2\pi\boldsymbol{\Sigma}_{c'}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{c'})^T\boldsymbol{\Sigma}_{c'}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{c'})\right]}$$

- 假设协方差相同：$\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_c$

$$p(y=c|\mathbf{x},\theta)\propto\pi_c\exp\left[\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c\right]$$

$$=\exp\left[\underbrace{\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}}_{\boldsymbol{\beta}^T}\mathbf{x}-\underbrace{\frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c+\log\pi_c}_{\gamma}\right]\exp\left[-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right]$$

$$p(y=c|\mathbf{x},\theta)=\frac{e^{\boldsymbol{\beta}_c^T\mathbf{x}+\gamma_c}}{\sum_{c'}e^{\boldsymbol{\beta}_{c'}^T\mathbf{x}+\gamma_{c'}}}$$

同Logstic回归的形式相同

与c无关

决策边界(cont.)

$$p(y=c|\mathbf{x},\theta)=\frac{e^{\boldsymbol{\beta}_c^T\mathbf{x}+\gamma_c}}{\sum_{c'}e^{\boldsymbol{\beta}_{c'}^T\mathbf{x}+\gamma_{c'}}}$$

$$\boldsymbol{\beta}_c=\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1},\quad \gamma_c=-\frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c+\log\pi_c$$

- 类别c & c'之间的决策边界:

$$p(y=c|\mathbf{x},\theta)=p(y=c'|\mathbf{x},\theta)$$

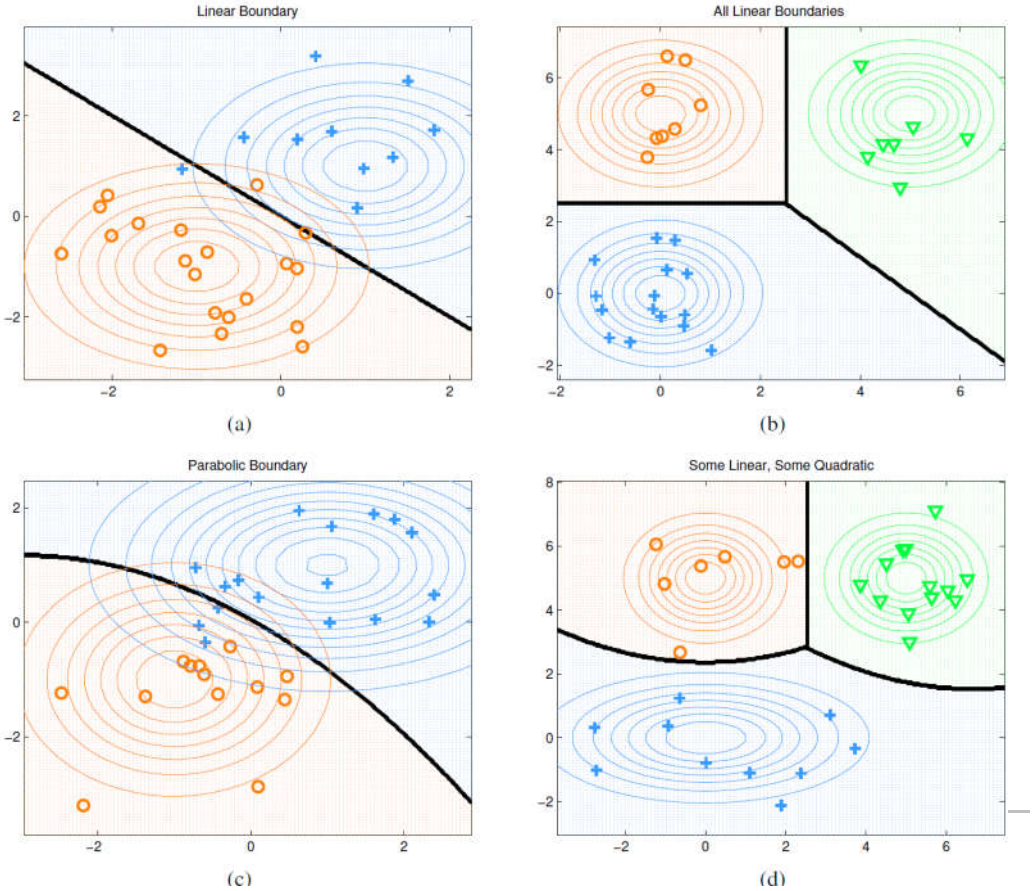
$$\boldsymbol{\beta}_c^T\mathbf{x}+\gamma_c=\boldsymbol{\beta}_{c'}^T\mathbf{x}+\gamma_{c'} \qquad (\Sigma\text{相同})$$

$$\mathbf{x}^T(\boldsymbol{\beta}_{c'}-\boldsymbol{\beta})=\gamma_{c'}-\gamma_c$$

  - 称为线性判别分析(Linear Discriminant Analysis, LDA)
    - 与$\mathbf{x}$呈线性
  - 一般情况下，决策边界为二次曲线，称为二次判别分析(Quadratic Discriminant Analysis, QDA)

$$p(y=c|\mathbf{x},\theta)\propto\exp\left[\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}_c^{-1}\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c+\log\pi_c\right]\exp\left[-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_c^{-1}\mathbf{x}\right]$$

决策边界举例


Linear Boundary
(a)
All Linear Boundaries
(b)
Parabolic Boundary
(c)
Some Linear, Some Quadratic
(d)

高斯判别分析（总结）

- 当产生式分类器 $p(y=c|\mathbf{x}) \propto p(\mathbf{x}|y=c)p(y=c)$ 中的类条件分布为高斯分布 $p(\mathbf{x}|y=c)=\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$
  - 当协方差矩阵为对角阵时，为朴素贝叶斯（在给定类别的情况下，各特征独立）
  - 当所有 $\boldsymbol{\Sigma}_c$ 都相等时，判别边界为线性，称为线性判别分析（Linear Discriminant Analysis, LDA）
  - 一般情况下，判别边界为二次曲线
  - 协方差决定了模型的复杂度（参数的数目）

三、概率图模型

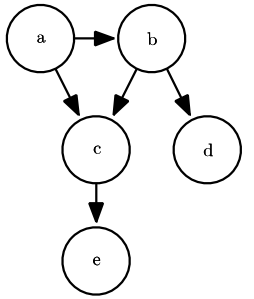
概率图模型

- 机器学习算法经常会涉及多元随机向量的概率分布。
- 如果采用单个函数来描述整个随机向量的联合分布是非常低效的(无论是计算上还是统计上)，因为这些随机变量中涉及到的直接相互作用通常只介于非常少的变量之间的。
- 利用随机变量之间的条件独立关系，可以将随机向量的联合分布分解为一些因式的乘积，得到简洁的概率概率表示。
- 我们可以采用图论中的 “图”的概率来表示这种分解，得到概率图模型：图中的节点表示随机变量，边表示随机变量之间的直接作用。
- 有向图和无向图均可用于概率表示。

有向图

- 有向图模型（directed graphical models，DGMs）使用带有有向边的图，用条件概率分布来表示分解：每个随机变量x_i都包含着一个影响因子，这些影响的影响因子被称为x_i的父节点，记为Pa (x_i)：

$$p(\mathbf{x}) = \prod_i p\left(x_i \mid Pa\left(x_i\right)\right)$$



- 例：右图对应的概率分布可以分解为

$$p(a,b,c,d,e) = p(a)\,p(b \mid a)\,p(c \mid a,b)\,p(d \mid b)\,p(e \mid c)$$

  - 从图模型可以快速看出此分布的一些性质：如a和c 直接相互影响，但a和e只有通过c间接相互影响

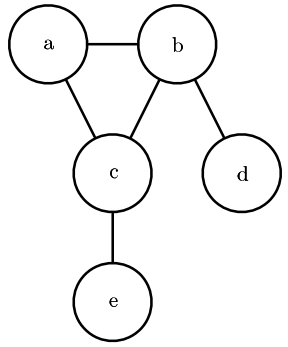
无向图

- 无向图模型（undirected graphical model，UGM）使用带有无向边的图，它将联合概率表示分解成一组函数的乘积
- 图中任何满足两两之间有边连接的顶点的集合被称为团（ clip ）。每个团$\mathcal{C}^i$都伴随着一个因子$\phi^i\left(\mathcal{C}^i\right)$
  - 每个因子的输出都必须是非负的
  - 但不像概率分布中那样要求因子的和/积分为1
- 随机向量的联合概率与所有这些因子的乘积成比例：

$$p\left(\mathbf{x}\right)=\frac{1}{Z}\prod_i \phi^i\left(\mathcal{C}^i\right)$$

- 其中归一化常数Z被定义为ϕ函数乘积的所有状态的求和或积分，使得这些乘积的求和为1（$p(\mathbf{x})$为一个合法的概率分布）

无向图（ cont. ）



- 例：右图对应的概率分布可以分解为

$$p(a,b,c,d,e)=\frac{1}{Z}\phi^{1}(a,b,c)\phi^{2}(b,d)\phi^{3}(c,e)$$

- 从图中可以快速看出此分布的一些性质：如a和c 直接相互影响，但a和e只有通过c间接相互影响

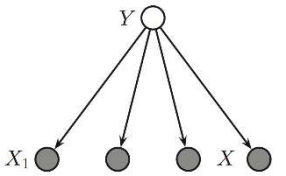
- 注意：这些图模型表示的分解仅仅是描述概率分布的一种语言，它们不是互相排斥的概率分布族，任何概率分布都可以用这两种方式进行描述。

特殊的概率图模型

- 下面我们学习一些常用的、特殊的概率图模型
  - 一般的概率图模型超出了这次课程的范围

- 朴素贝叶斯分类器
- Markov链
- 隐马尔科夫模型（HMM）
- Markov随机场（MRF）
- 条件随机场（CRF）

► 朴素贝叶斯 (Naive Bayes Classifier, NBC)



- 假设要处理分类任务，共有 C 个类别 $y \in (1,2,\ldots,C)$
- 每个样本有特征 $\mathbf{x} = (x_1, x_2, \ldots, x_D)$
- 朴素贝叶斯分类器是比较简单也很常用的分类器
  - 简单/朴素：假设各维特征在给定类别标签的情况下条件独立

$$p(\mathbf{x} \mid y = c, \theta) = \prod_{j=1}^{D} p(x_j \mid y = c, \theta)$$

  - 通常即使特征条件独立的假设不满足，NBC在实际应用中性能也不错（模型简单，不容易过拟合）
  - 预测：$p(y = c \mid \mathbf{x}) \propto p(\mathbf{x} \mid y = c)\, p(y = c)$

例：Titanic生存预测

背景：1912年4月15日，泰坦尼克号撞上冰山后沉没。
一个海难导致生命损失的原因是没有足够的救生艇的乘客和船员。
一些群体的人更可能生存其他群体更容易生还，比如妇女，儿童和上层阶级等。

任务：具备什么特征的人可能生存
训练集：891个样本
测试集：418个样本

PassengerId	旅客ID	这条数据应该没啥用
Survived	是否活下来了，1:yes 0:no	目标
Pclass	旅客等级 1 2 3 分别代表不同的等级	
Name	名字	
Sex	性别	
Age	年龄	
SibSp	有多少兄弟姐妹/配偶同船 Number of Siblings/Spouses Aboard	
Parch	有多少父母/子女同船 Number of Parents/Children Aboard	
Ticket	船票号码？	
Fare	船票收费	
Cabin	所在小屋	
Embarked	登船城市 Port of Embarkation	C Q S 分别代表不同的城市

我们初步尝试基于Age，Sex和fare特征，采用朴素贝叶斯分类器实现预测

例：Titanic生存预测

$$p(y=c|\mathbf{x}) \propto p(\mathbf{x}|y=c)p(y=c)$$

- NBC模型为： $p(\mathbf{x}|y=c,\theta)=\prod_{j=1}^{D} p(x_j|y=c,\theta)$
  - 类先验

    $p(survived)=Ber(\theta_1)$
  - 类条件概率

    $p(Sex=Male|survived=1)=Ber(\theta_2), p(Sex=Male|survived=0)=Ber(\theta_3)$

    $p(Age|survived=1)=\mathcal{N}(\theta_4,\theta_5), p(Age|survived=0)=\mathcal{N}(\theta_6,\theta_7)$

    $p(Fare|survived=1)=\mathcal{N}(\theta_8,\theta_9), p(Fare|survived=0)=\mathcal{N}(\theta_{10},\theta_{11})$
  - 预测：

    $p(survived=1|Sex,Age,Fare) \propto p(Sex|survived=1)p(Age|survived=1)p(Fare|survived=1)p(survived=1)$

    $p(survived=0|Sex,Age,Fare) \propto p(Sex|survived=0)p(Age|survived=0)p(Fare|survived=0)p(survived=0)$

也可以将Age和Fare分别量化为几个类别，然后处理过程同Sex

链规则

- 条件独立的另一个重要应用是对序列的概率分布建模
  - 在自然语言处理、计算生物学、时序预测等问题上均有应用

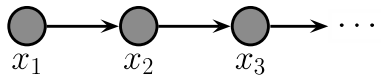
- Recall链规则：给定时间长度为T的序列$X_1,...,X_T$，

$$p(X_1,...,X_T) = p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_2,X_1)...$$
$$= p(X_1)\prod_{t=2}^{T} p(X_t \mid X_{1:t-1})$$

- 即第t时刻的状态X_t只与前t-1个时刻的状态$X_{1:t-1}$相关

Markov链

- 若假设第t时刻的状态X_t只与前一个时刻个时刻的状态X_{t-1}相关，称为一阶Markov假设，得到的联合分布为Markov链（或Markov模型）：



$$p(X_1,...,X_T)=p(X_1)\prod_{t=2}^{T}p(X_t\mid X_{t-1})$$

- 如果进一步假设 $p(X_t\mid X_{t-1})$ 与时间t无关，该链被称为同质的（homogeneous）、稳定的（stationary）或时不变的（time-invariant）.

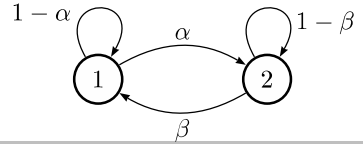
转移矩阵

- 当X_t为离散时，即$X_t \in (1,2,..,K)$则条件分布$p(X_t \mid X_{t-1})$可表示为一个KxK的矩阵，称之为转移矩阵，其中

$$A_{ij} = p(X_t = j \mid X_{t-1} = i)$$

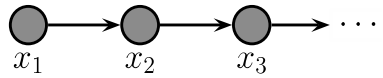
- 表示从状态i转移到状态j的概率。
- $\sum A_{ij} = 1$, 因此亦被称为随机矩阵。
- 矩阵A也可以用 有向图表示：节点表示状态，边表示合法的状态转移，边的权重为转移概率。
- 例：一个两个状态的转移矩阵及对应的图：

$$\mathbf{A} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

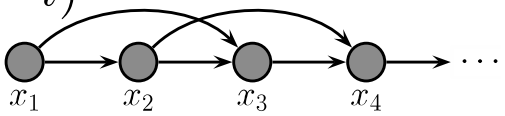


应用：n-gram语言模型

- Markov模型在语言处理模型中应用广泛
- 一元语言模型：只考虑语言中的单个词语的概率 $p(X_t = k)$
- 二元语言模型：考虑相邻两个词语之间的关系
  - 一阶Markov链：$p(X_t = k|X_{t-1} = j)$

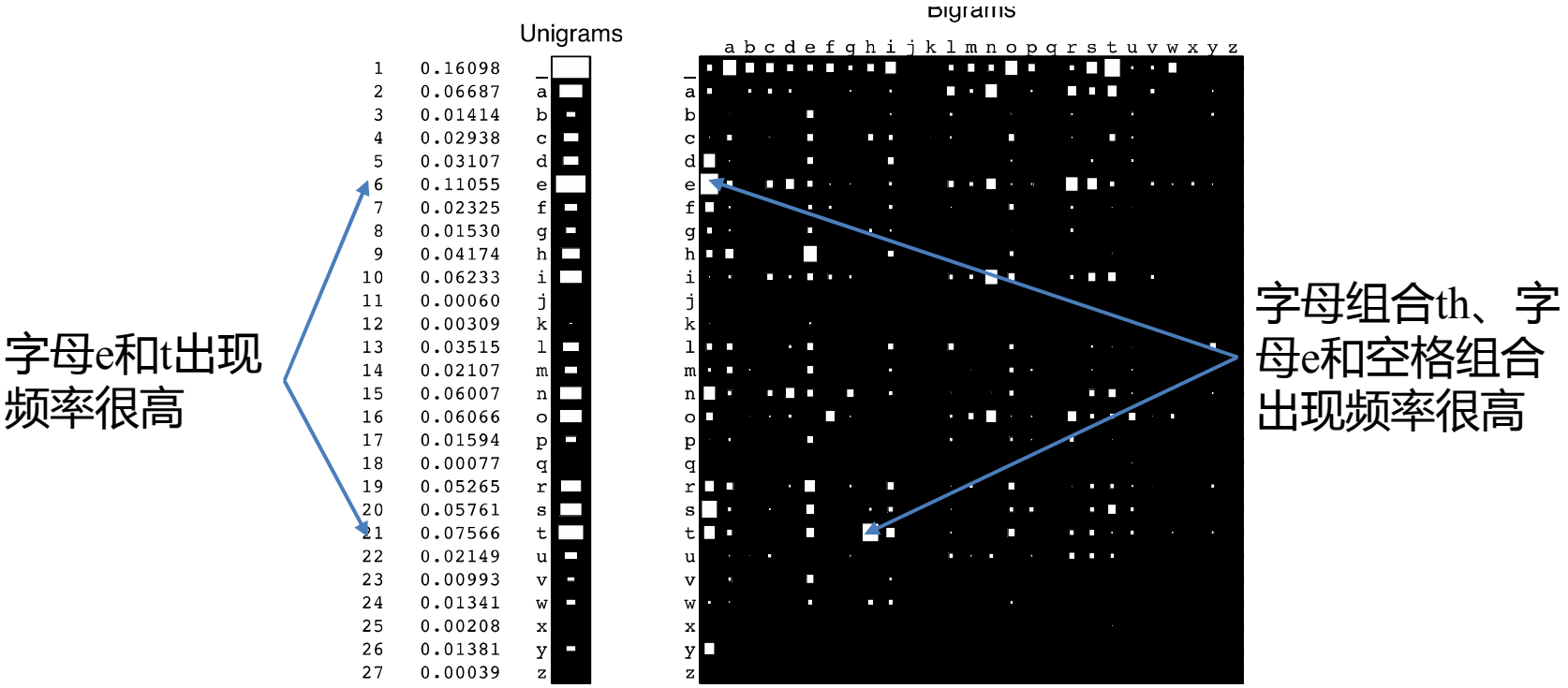


- 三元语言模型：考虑相邻三个词语之间的关系
  - 二阶Markov链：$p(X_t = k|X_{t-1} = j, X_{t-2} = i)$

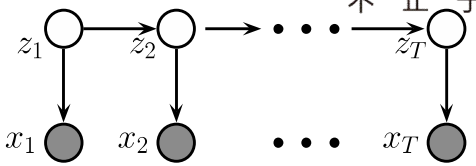


- …
- n元语言模型：考虑相邻n个词语之间的关系
  - n-1阶Markov链：

例：英文字母的n元模型


字母e和t出现频率很高
Unigrams
1 0.16098 _
2 0.06687 a
3 0.01414 b
4 0.02938 c
5 0.03107 d
6 0.11055 e
7 0.02325 f
8 0.01530 g
9 0.04174 h
10 0.06233 i
11 0.00060 j
12 0.00309 k
13 0.03515 l
14 0.02107 m
15 0.06007 n
16 0.06066 o
17 0.01594 p
18 0.00077 q
19 0.05265 r
20 0.05761 s
21 0.07566 t
22 0.02149 u
23 0.00993 v
24 0.01341 w
25 0.00208 x
26 0.01381 y
27 0.00039 z
Bigrams
a b c d e f g h i j k l m n o p q r s t u v w x y z
字母组合th、字母e和空格组合出现频率很高

隐马尔科夫模型（Hidden Markov Model, HMM）



- 如果系统的状态不可见，只能观测到由隐含状态驱动的观测变量，则可用隐马尔可夫模型表示联合概率为：

$$p(\mathbf{z}_{1:T}, \mathbf{x}_{1:T}) = p(\mathbf{z}_{1:T})p(\mathbf{x}_{1:T}|\mathbf{z}_{1:T}) = \left[p(z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})\right]\left[\prod_{t=1}^{T} p(\mathbf{x}_t|z_t)\right]$$

- 其中z_t表示第t时刻的隐含状态，$p(z_t|z_{t-1})$表示转移模型，
- $p(\mathbf{x}_t|z_t)$表示观测模型
- 通常z_t隐含状态是我们更感兴趣的内容，如在语音识别中，z_t是词语，观测x_t为语音波形。

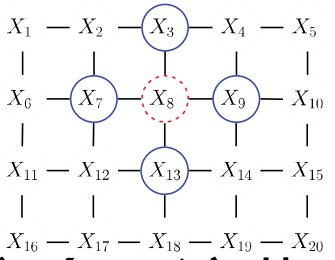
例：词性标注

- 任务：给定一个句子，给出每个单词的词性

$$p(T_1,...,T_M,W_1,...,W_M)=\left[p(T_1)\prod_{m=2}^{M}p(T_m\mid T_{m-1})\right]\left[\prod_{m=1}^{M}p(W_m\mid T_m)\right]$$

- 隐含变量：单词的词性T_m
- 隐含状态转移矩阵：词性之间的转换概率 $p(T_m\mid T_{m-1})$
  - 这个步骤并不需要考虑具体单词，而只考虑词性之间转移概率，如形容词→名词、名词→动词等概率较大
- 观测模型：某个词性下单词出现的概率 $p(W_m\mid T_m)=\dfrac{p(W_m,T_m)}{p(T_m)}$

Markov随机场（MRF）



- 随机场可以看成是一组随机变量的集合（这些随机变量之间可能有依赖关系）
- Markov随机场：加了Markov性质限制的随机场，可用无向图表示
  - 对Markov随机场中的任何一个随机变量，给定场中其他所有变量下该变量的分布，等同于给定场中该变量的邻居节点下该变量的分布
  - 如上图二维网格中，在给定节点X_8的4个直接邻居的情况下，节点X_8与其他节点无关
  - Markov blanket：一个节点的Markov blanket为一个节点集合，在给定集合中节点的情况下，该节点与图中其他节点条件独立
    - 上图中节点X_8的Markov blanket 是X_8 4个直接邻居：X_3、X_7、X_9和X_{13}

MRF的参数化

- Recall：有向图中节点之间的边有方向，采用链规则表示联合分布（每个节点有条件概率分布）
- 无向图中节点之间的边没有方向，不能用链规则表示联合概率 $p(\mathbf{y})$，而是用图中每个最大团 $\mathcal{C}$ 的势能函数（**potential functions**）或因子 $\psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$ 的乘积表示：

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$$

- 其中 $Z(\boldsymbol{\theta}) \triangleq \sum_{\mathbf{x}} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$ 为partition function，使得联合分布的积分为1。

Gibbs distribution

- 一种势能函数的表示方式与统计物理有关系：能量函数
  - 无向图与统计物理之间颇有渊源
- Gibbs分布：

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(-\sum_{c} E(\mathbf{y}_c|\boldsymbol{\theta}_c))$$

- 其中$E(\mathbf{y}_c) > 0$为团簇C中变量相关的能量函数
- 等价于无向图的势能函数为：$\psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c) = \exp(-E(\mathbf{y}_c|\boldsymbol{\theta}_c))$
  - 能量低的配置（状态）对应的概率更大（在统计物理和统计化学中广泛应用）

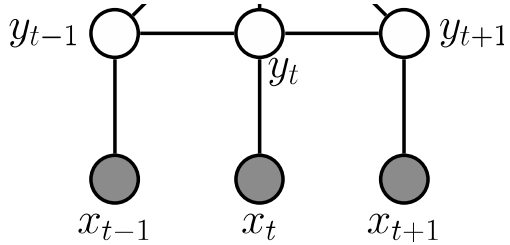
MRF的参数化（cont.）

- 另一种势能函数表示方式是将log势能函数表示为一些函数的线性组合：$\log \psi_c(\mathbf{y}_c) \triangleq \boldsymbol{\phi}_c(\mathbf{y}_c)^T \boldsymbol{\theta}_c$
- 其中组合权重为$\boldsymbol{\theta}$，$\boldsymbol{\phi}_c(\mathbf{y}_c)$为根据变量$\mathbf{y}_c$得到的特征，
- 则log联合概率为

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \sum_c \boldsymbol{\phi}_c(\mathbf{y}_c)^T \boldsymbol{\theta}_c - Z(\boldsymbol{\theta})$$

- 亦被称为最大熵模型或log线性模型。
  - 条件随机场（CRF）、（受限）Boltzmann 机（RBM）可用此形式表示联合概率

条件随机场（CRF）



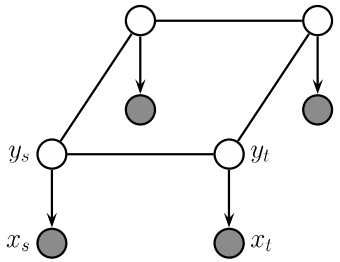
- 如果给定MRF中每个随机变量下面还有观测值，我们要确定的是给定观测条件下MRF的分布，那么该MRF就称为 条件随机场 (Conditional Random Field， CRF)

- CRF的条件分布形式类似于MRF的分布形式，但多了一个观测集合$\mathbf{x}$

$$p(\mathbf{y}|\mathbf{x},\mathbf{w}) = \frac{1}{Z(\mathbf{x},\mathbf{w})}\prod_{t=1}^{T}\psi(y_t|\mathbf{x},\mathbf{w})\prod_{t=1}^{T-1}\psi(y_t,y_{t+1}|\mathbf{x},\mathbf{w})$$

  - CRF是一种判别式图模型

例：CRF

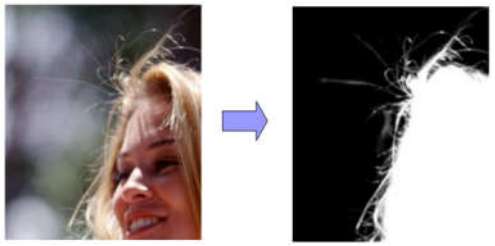


- 一种用于图像分割等任务的模型为

$$p(\mathbf{y},\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{y}|J)\prod_t p(x_t|y_t,\boldsymbol{\theta}) = \left[\frac{1}{Z(J)}\prod_{s\sim t}\psi(y_s,y_t;J)\right]\prod_t p(x_t|y_t,\boldsymbol{\theta})$$

  - 这是一个有向图和无向图结合的例子
- 其中**y**为图像的分割标签，2D无向图网格表示先验$p(\mathbf{y}|J)$，通常我们假设相邻节点倾向于有相同的状态
- $p(x_t|y_t = k,\boldsymbol{\theta})$表示局部证据（ local evidence ），表示观测到像素x_t时，分割标签y_t为k的概率（可用高斯分布或核密度估计表示）

▶例：抠图

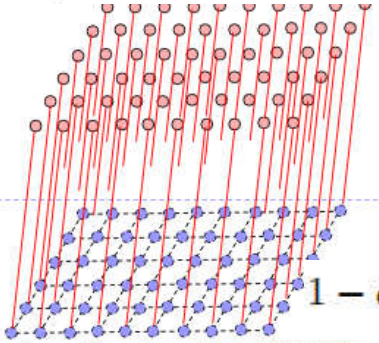


在抠图中，$I(z)=\alpha_z F(z)+(1-\alpha_z)B(z)$

观测变量X为图像中每个像素的颜色C

隐含变量为每个像素的的透明度α、前景颜色F和背景颜色B，以及估计的不确定程度u

$$exp\left(-d_c\left(C_p,\alpha^k F_i^p+(1-\alpha^k)B_j^p\right)^2/2\sigma_d^{k2}\right)$$



Observations Y

Hidden Field X

$$1-exp\left(-(\alpha_1-\alpha_2)^2/\sigma_s^2\right)$$

Markov Random Field (MRF)

Goal: maximize $P(X,Y)$

Assumption: y_i's are independent

Foreground samples F_i^p

Background samples B_i^p

Observed color C_p

$I(z)=\alpha_z F(z)+(1-\alpha_z)B(z)$

Observations y

Hidden Node p

Quantized alpha level α^k

Estimated foreground color $F^*(p)$

Estimated background color $B^*(p)$

Uncertainty $u(p)$

Jue Wang et. An Iterative Optimization Approach for Unified Image Segmentation and Matting， ICCV2005

下节课预告

- 极大似然估计
  - 原理
  - 常见概率分布和机器学习模型的参数估计