Chapter 6：Probabilistic Graphical Model

2018.12.25 Tuesday. By KuKuXia@github.com

ML/DL/RL深度交流讨论群: 622545311，加群时请备注: 姓名-方向-公司OR高校

PGM

- 用观测节点表示观测到的数据，用隐含节点表示潜在的知识，用边来描述知识与数据的相互关系，最好基于这样的关系图获得一个概念分布
- 主要类别
  - 有向图 — 贝叶斯网络(Bayesian Network)
  - 无向图 — 马尔科夫网络(Markov Network)

生成式模型与判别式模型

- 假设可以观测到的变量集合为X,需要预测的变量集合为Y，其他变量集合为Z

$$P(Y|X)=\frac{P(X,Y)}{P(X)}=\frac{\sum_Z P(X,Y,Z)}{\sum_{Y,Z} P(X,Y,Z)}$$

- 对联合概率分布P(X,Y,Z)进行建模，在给定观测集合X的条件下，通过计算边缘分布来得到对变量集合Y的推断
- 生成式模型
  - 朴素贝叶斯
  - 常见模型
    - 贝叶斯网络
    - pLSA
    - LDA
    - 隐马尔科夫模型
- 对序列数据进行建模
- 直接对条件概率分布P(Y,Z|X)进行建模，然后消掉无关变量Z就可以得到对变量集合Y的预测

$$P(Y|X)=\sum_Z P(Y,Z|X)$$

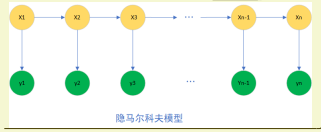
- 判别式模型
  - 最大熵模型
  - 常见模型
    - 条件随机场

马尔科夫模型

- 马尔科夫过程
  - 满足无后效性的随机过程

$$P(x_n|x_1,x_2,...,x_{n-1})=P(x_n|x_{n-1})$$

  - 假设一个随机过程中，t_n时刻的状态x_n的条件分布，仅仅与其前一个状态x_(n-1)有关，则称其为马尔科夫过程
  - 时间和状态取值都是离散的马尔科夫过程也称为马尔科夫链
- 在简单的马尔科夫模型中，所有状态对于观测者都是可见的，因此在马尔科夫模型中仅仅包括状态间的转移概率
- 隐马尔科夫模型
  - 是对隐状态序列和观测状态序列的联合概率P(x,y)进行建模的生成模型
  - 隐马尔科夫模型是对含有未知参数隐状态的马尔科夫链进行建模的生成模型
  - 隐状态x_i对于观测者而言是不可见的，能观测到的只有每个隐状态x_i对应的输出y_i，而观测状态y_i的概率分布仅仅取决于对应的隐状态x_i
  - 假设观测状态间相互独立
  - 参数
    - 隐状态间的转移概率
    - 隐状态到观测状态的输出概率
    - 隐状态x的取值空间
    - 观测状态y的取值空间
    - 初始状态的概率分布
  - 基本问题
    - 概率计算问题
    - 预测问题
    - 学习问题



隐马尔科夫模型

- 对比
- 对比
- 改进
- 改进
- 去掉了隐马尔科夫模型中观测状态相互独立的假设，考虑了整个观测序列，有更强的表达能力
- 直接对标注后的后验概率P(y|x)进行建模的判别式模型

$$P(x_{1...n}|y_{1...n})=\prod_{i=1}^{n} P(x_i|x_{i-1},y_{1...n})$$

  - 最大熵隐马尔科夫模型MEMM: Maximum Entropy Markov Model
  - 建模公式
  - P(x_i|x_(i-1),y_(1...n)会在局部进行归一化，会导致标注偏置问题
  - 由于局部归一化的影响，隐状态会转向于转移到那些后续状态可能更少的状态上，以提高整体的后验概率，这就是标注偏置问题
  - 在MEMM的基础上，进行了全局归一化，解决了局部归一化带来的标注偏置问题

$$P(x_{1...n}|y_{1...n})=\frac{1}{Z(y_{1...n})}\prod_{i=1}^{n}\exp(F(x_i,x_{i-1},y_{1...n}))$$

  - 条件随机场CRF：Conditional Random Field
  - Z(y_(1...n)是在全局范围进行归一化

概率图表示

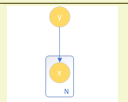
- 朴素贝叶斯模型
  - 通过预测指定样本属于特定类别的概率P(y_i|x)来预测该样本的所属类别

$$P(y_i|x)=\frac{P(x|y_i)P(y_i)}{P(x)}$$

  - 贝叶斯公式

$$P(y_i|x)\propto P(x|y_i)P(y_i)=P(x_1|y_i)P(x_2|y_i)\cdots P(x_n|y_i)P(y_i)$$

  - 后验概率P(x_j|y_i)的取值决定了分类的结果，并且任意特征x_j都由y_i的取值影响



朴素贝叶斯模型的概率图模型

  - 上图为盘式记法，表示变量y同时对x_1,x_2,...,x_n这N个变量产生影响。
- 最大熵模型
  - 信息是指人们对事物的理解不确定性的降低或消除，而熵就是不确定性的度量，熵越大，不确定性就越大
  - 最大熵原理是概率模型学习的一个准则，指导思想是在满足约束条件的模型集合中选取熵最大的模型，即不确定性最大的模型。就是学习到合适的分布P(y|x)，使得条件熵H(P)的取值最大

$$H(P)=-\sum_x P(x)\log P(x)$$

  - 假设离散随机变量x的分布是P(x)，则关于分布P的熵的定义如图，当x服从均匀分布时，对应的熵最大，也就是不确定性最高

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

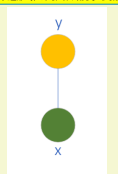
  - 给定离散随机变量x和y上的条件概率分布P(y|x)，定义在条件概率分布上的条件熵

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

$$s.t.,\ E_{\tilde{P}}(f_i)=E_P(f_i),\forall i=1,2,...,M,$$

$$\sum_y P(y|x)=1$$

  - 最大熵模型的学习等价于约束最优化问题
  - $P_w(y|x)=\frac{1}{Z}\exp(\sum_{i=1}^{n} w_i f_i(x,y))$
  - 最大熵模型的表达形式，即最终归结为学习最佳的参数w，使得P_w(y|x)最大化



最大熵模型的概率图模型

  - 概率图角度，P_w(y|x)的表达形式非常类似于势函数为指数函数的马尔科夫网络，其中变量x和y构成了最大团

主题模型

- 基于词袋模型或N-gram模型的文本表示模型有一个明显缺陷，就是无法识别出两个不同的词或词组具有相同的主题
- 主题模型是一种能够将具有相同主题的词或词组映射到同一个维度上去的模型
- 常见主题模型
  - pLSA
    - 一个用生成模型来建模文章的生成过程，频率派思想，将每篇文章对应的主题分布p(z_k|d_m)和每个主题对应的词分布p(w_n|z_k)看成确定的未知参数，并可以求解出来
    - 利用似然函数表示整个语料库中的文本生成概率，但是由于参数中包含有隐含变量，因此无法用最大似然法直接求解，可以利用最大期望算法来解决
  - LDA
    - 贝叶斯学派思想，认为待估计的参数(主题分布和词分布)不是一个固定的常数，而是服从一定分布的随机变量，这个分布符合一定的先验概率分布，即狄利克雷分布(Dirichlet)，并且在观察到样本信息之后，可以对先验分布进行修正，从而得到后验分布
    - 主题个数K是一个预先指定的超参数
      - 确定K的第一种方法：可以利用训练集、验证集进行训练，然后用测试集对超参数进行选择
      - 评估指标 — 困惑度Perplexity
      - 分层狄利克雷过程HDP: Hierarchical Dirichlet Process
        - 是一种非参数主题模型HDP-LDA
        - 不需要预先指定主题的个数，模型可以随着文档数目的变化而自动对主题个数进行调整
        - 缺点是使得模型更加复杂，训练速度更加缓慢。实际中经常采用第一种方法确定主题个数K
- 推荐系统冷启动问题
  - 是指在没有大量用户数据的情况下如何给用户进行个性化推荐，目的是最优化点击率、转化率或者用户体验(用户停留时间、留存率等)
  - 分类
    - 用户冷启动 — 指对一个之前没有行为或行为极少的新用户进行推荐
    - 物品冷启动 — 指为一个新上市的商品或电影(这时没有与之相关的评分或用户行为数据)，寻找具有潜在兴趣的用户
    - 系统冷启动 — 指如何为一个新开发的网站设计个性化推荐系统
  - 解决方法
    - 一般解决方法是基于内容的推荐