Book note: Introduction To Bayesian Networks

20201006

1 Ideas

- Is it possible to encapsulate dataset or expert knowledge into prior distribution to conduction ML, like perception tasks, in the bayesian way?
- Autonomous driving diagnose module (Bayesian implementation as expert system)

2 Statement

- The multi-sensor fusion solution performs better than the single-sensor solution. From the aspect of information theory, it can be attributed to the multi-sensor fusion solution has more observations, so that the entropy will be smaller and the uncertainty will get dropped.

2.1 Relationship between joint entropy, conditional entropy and mutual information

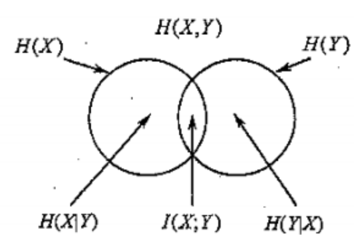

图 1.6 联合熵、条件熵以及互信息之间的关系

1. 朴素贝叶斯模型

. **朴素贝叶斯模型**（naïve Bayes model），又称**朴素贝叶斯分类器**（naïve Bayes classifier），是一个包含一个根节点、多个叶节点的树状贝叶斯网，如图 2.27 所示. 其中叶节点 $A_1,\cdots,A_n$ 是属性变量，描述待分类对象的属性，根节点 C 是类别变量，描述对象的类别. 用朴素贝叶斯模型进行分类就是给定一个数据点，即各属性变量的取值 $A_1=a_1,\cdots,A_n=a_n$，计算后验分布 $P(C\mid A_1=a_1,\cdots,A_n=a_n)$，然后选择概率最大的那个 C 值作为这个数据点所属的类别. 在医疗诊断中，C 代表一系列疾病，$A_1,\cdots,A_n$ 代表这些疾病可能导致的症状，诊断就是根据症状来确定疾病.

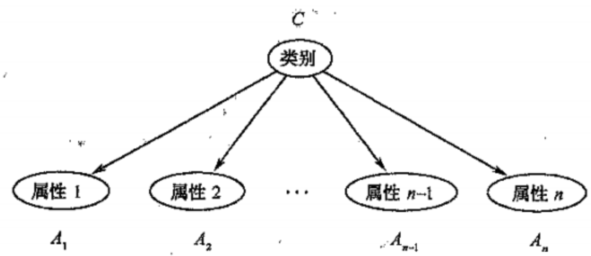

图 2.27　朴素贝叶斯分类器

贝叶斯网推理的 3 个主要问题：后验概率问题、最大后验假设问题（MAP）和最大可能解释问题（MPE），都是 NP-难解的（Cooper，1990；Shimony，1994；Park，2002）. MPE 是典型的组合优化问题，它的判定问题是 NP-完全的（Shimony，1994；Park，2002）；后验概率问题是一个计数问题（counting problem），它是 # P-完全的（Littman et al.，2001；Roth，1996）；MAP 是 NP^{PP}-完全的（Park，2002）.

2.2 Naive bayes model

2.3 MAP is NP-HARD

2.4 An MCMC example: Gibbs sampling

2.5 Why using Beta distribution instead of Gaussian as the prior distribution for MAP?

It is because we want to use a prior distribution with a similar form as the likelihood.

6.1.2 MCMC 抽样

在重要性抽样算法中，不同样本之间相互独立．下面介绍 MCMC 抽样算法，

2.6 Property of MLE

3 Problems

- U-separation
- Latent variable analysis

其中不同样本之间不是相互独立的.

图 6.3 所示的是一个简单的 MCMC 算法，称为**吉布斯抽样**（Gibbs sampling）. 它首先随机生成一个与证据 $\mathbf{E}=\mathbf{e}$ 相一致的样本 $\mathbf{D}_1$ 作为起始样本，此后每一步都从当前样本出发产生下一个样本. 设当前在 $i-1$ 步，为了从 $\mathbf{D}_{i-1}$ 出发得到 $\mathbf{D}_i$，算法首先设 $\mathbf{D}_i=\mathbf{D}_{i-1}$，然后按照某个顺序对非证据变量逐个进行抽样，改变 $\mathbf{D}_i$ 中变量的取值. 设 Z 是下一个待抽样变量，$\mathbf{Y}$ 是 Z 的马尔可夫边界上的变量的集合，$\mathbf{y}_i$ 是 $\mathbf{Y}$ 在 $\mathbf{D}_i$ 中的当前取值. 算法根据分布 $P(Z \mid \mathbf{Y}=\mathbf{y}_i)$ 对 Z 进行抽样，并用抽样结果替代 $\mathbf{D}_i$ 中 Z 的当前取值.

```
GibbsSampling (𝒩, m, E, e, Q, q, ρ)

输入：𝒩——一个贝叶斯网；m——样本量；
      E——证据变量；e——证据变量的取值；
      Q——查询变量；q——查询变量的取值；
      ρ——非证据变量的抽样顺序.
输出：对 P(Q = q | E = e) 的近似；
   m_q ← 0 ;
   随机生成一个与 E = e 一致的样本 D_1 ;
   if (D_1 与 Q = q 一致)
       m_q ← m_q + 1 ;
   end if
   for (i = 2 to m)
       D_i ← D_{i-1} ;
       for (ρ 中的每一个变量 Z)
           设 Y = mb(Z), y_i 是 Y 在 D_i 中的当前取值，从 P(Z | Y = y_i) 抽样；
           用抽样结果替代 D_i 中 Z 的取值；
       end for
       if (D_i 与 Q=q 一致)
           m_q ← m_q + 1;
       end if
   end for
   return m_q/m.
```

图 6.3　吉布斯抽样算法

例 6.4　对如图 6.1 所示的贝叶斯网，用吉布斯抽样算法计算 $P(R=t \mid S=t)$. 首先随机生成一个与证据 $\{S=t\}$ 一致的样本，假设它是 $\mathbf{D}_1=\{C=t, R=t, S=t, W=f\}$. 接下来生成样本 $\mathbf{D}_2$ ：算法从 $\mathbf{D}_2=\mathbf{D}_1=\{C=t, R=t, S=t, W=f\}$ 出发，对非证据变量逐个抽样. 设抽样顺序为 $\langle C, R, W\rangle$. 抽样过程

如下：

(1) 对 C 进行抽样，抽样分布为 $P(C \mid R=t, S=t) \approx (0.444, 0.556)$，假设抽样结果为 $C=f$，于是 $\mathbf{D}_2$ 变为 $\{C=f,\ R=t,\ S=t,\ W=f\}$；

(2) 对 R 进行抽样，此时 C 的取值是 f，因此抽样分布为 $P(R \mid C=f, S=t, W=f) \approx (0.024, 0.976)$，假设抽样结果为 $R=f$，于是 $\mathbf{D}_2$ 变为 $\{C=f,\ R=f,\ S=t,\ W=f\}$；

(3) 对 W 进行抽样，此时 R 的取值是 f，因此抽样分布为 $P(W \mid R=f, S=t)=(0.9, 0.1)$，假设抽样结果为 $W=f$，于是 $\mathbf{D}_2$ 仍是 $\{C=f,\ R=t,\ S=t,\ W=f\}$，这是 $\mathbf{D}_2$ 的最终值. □

设抽样共得到 m 个样本，其中满足 $\mathbf{Q}=\mathbf{q}$ 的有 $m_{\mathbf{q}}$ 个. 那么，可以按下式近似计算后验概率：

$$P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e}) \approx \frac{m_{\mathbf{q}}}{m}.$$

3. 共轭分布族

为加深理解，我

和似然函数 $L(\theta \mid \mathscr{D})$ 的乘积. 在 i. i. d. 假设下，$L(\theta \mid \mathscr{D})$ 是二项似然函数.
面假设先验分布 $p(\theta)$ 来自贝塔分布族. 这是因为贝塔分布族是二项似然函
的**共轭分布族**（conjugate family），即如果先验分布 $p(\theta)$ 是贝塔分布，那么
验分布 $p(\theta \mid \mathscr{D})$ 也是贝塔分布. 这使得贝叶斯估计的计算简单易行. 事实
如果假设 $p(\theta)$ 来自另一分布族，比方说正态分布，那么贝叶斯估计计算起
就要困难得多.

另外，共轭分布族的使用也使得我们可以清楚地了解到贝叶斯估计是怎样
先验知识与观测数据结合到一起的. 假设 $p(\theta)$ 为贝塔分布 $B[\alpha_h, \alpha_t]$ 实际上就
做如下假设：先验知识相当于一组包含 α_h 个头朝上和 α_t 个尾朝上的样本的虚
数据（imaginary data). 贝叶斯估计把这些虚拟数据和实际观测所得到的数据
到一起，得到一组由 $m_h + \alpha_h$ 个头朝上和 $m_t + \alpha_t$ 个尾朝上的样本所组成的数据
是，$p(\theta \mid \mathscr{D})$ 是 $B[m_h + \alpha_h, m_t + \alpha_t]$.

7.5.2 最大似然估计的性质

下面讨论最大似然估计的性质. 设 $\mathcal{N}=(\mathcal{G},\boldsymbol{\theta}_{\mathcal{N}})$ 为一贝叶斯网. 用 $P_{\mathcal{N}}(\mathbf{X}$
$\mathcal{N}$所表示的联合概率分布. 在 $\mathcal{N}$中，把参数换成最大似然估计 $\boldsymbol{\theta}^*$，得到另
贝叶斯网 $\mathcal{N}^*=(\mathcal{G},\ \boldsymbol{\theta}^*)$，记其联合概率分布为 $P^*(\mathbf{X})$. 我们要回答如下
问题：

(1) 如果数据 $\mathcal{D}$ 是从 $P_{\mathcal{N}}(\mathbf{X})$ 中抽样而得来的，那么当样本量 m 趋于
时，P^* 会不会收敛？会不会收敛到 $P_{\mathcal{N}}$？

(2) 如果数据 $\mathcal{D}$ 是从另一个分布 $P(\mathbf{X})$ ($P\neq P_{\mathcal{N}}$) 中抽样而得来的，那
样本量 m 趋于无穷时，P^* 会不会收敛？收敛到什么分布？

如果数据 $\mathcal{D}$ 是从贝叶斯网 $\mathcal{N}$的联合分布 $P_{\mathcal{N}}(\mathbf{X})$ 中抽样而得来的，则称
$\mathcal{D}$ 的**原生模型** (generative model)，称 $P_{\mathcal{N}}(\mathbf{X})$ 是 $\mathcal{D}$ 的**原生分布** (generative
tribution). 这个概念在讨论学习算法的性质时将经常用到.