研究生算法课课堂笔记

上课日期： 2016 年 9 月 19 日 第(2)节课

组长学号及姓名：王皓 1601214482

组员学号及姓名：王义中 1601214483 曾有为 1601214494

- 信息熵 PPT 17 页推导过程中有一步没写：

$$\sum_{a,b} P(b) \cdot P(a|b) \cdot log \frac{1}{P(b)}$$

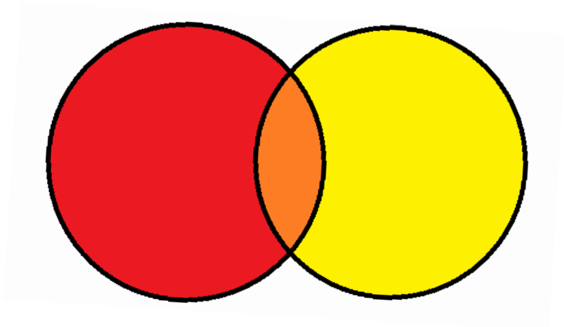
$$= \sum_{b} P(b) \cdot log \frac{1}{P(b)} \cdot \sum_{a} P(a|b) = \sum_{b} P(b) \cdot log \frac{1}{P(b)}$$

- $H(X) \geq H(X|Y)$，但是当已知某一特定 Y=y 时，可能存在$H(X) < H(X|y)$.

  如果以 ++++++---- 表示 6 个正样本和 4 个负样本，可以用 X 来表示样本本身的分布，可以通过分类决策变量 Y 将这 10 个样本分为两类，一类为 ++-- ，另一类为 ++++-- ，所以有$H(X) < 1$，$H(X|y = 0) = 1$.

- $H(A, B) = H(A) + H(B|A) = H(B) + H(A|B)$.

  以 Venn 图表示，红色圆代表$H(A)$，黄色圆代表$H(B)$，含重叠部分。红色区域（不含重叠部分）代表$H(A|B)$，黄色区域同理。橙色部分代表$I(A; B)$。



- 三个随机变量的熵之间的关系比两个变量要复杂，参见讲义第 19 页。Info-theory 的讲义第 19 和 20 页介绍的 Three source 和 Markov source 不作要求。

- 条件熵有什么意义？

  答案：以广告的点击率预测为例。$X = (x_1, x_2, x_3, \ldots, x_n)$．$x_i$是各种特征，包括用户的特征和广告本身的特征。Y 是用户是否点击广告。我们要在已知 X 的基础上预测 Y。如果 Y 是 X 的确定函数，$H(Y|X) \approx 0$。如果已知 X 的情况下只能得到 Y 的一个概率分布，那么$H(Y|X) > 0$。

- 当有算法 A 和算法 B 时，怎么判断哪一个更好？

  1) 根据 misclassification rate。

但是简单地根据错误率来判断并不能很好地反映真实的好坏程度。比如，当判断事件 Y 时，如果算法 A 给出 Y=1 的概率为 0.51，而算法 B 给出 Y=1 的概率为 0.95，他们都判断了 Y=1，但是正确的情况是 Y=0。此时，算法 A 和 B 都犯了 1 次错误，但显然算法 A 要相对好一些。

2) 根据 surprise 的程度。

每次 surprise 的程度可以用$log\frac{1}{q_i}$来表示，q_i表示第 i 次预测的概率。

整体的 surprise 程度可以用下面式子来衡量：

$$E_p\left[log\frac{1}{q_i}\right] = \sum_{i=1}^{n} p_i \cdot log\frac{1}{q_i}$$

其中，p_i是真实的概率分布。

可以利用 Jensen’s inequality 证明，当 $q_i = p_i$ 时，也就是预测的概率与真实分布相同时，整体 surprise 程度最小。

证明在讲课中略去，Intuition 是：熵可以代表最短编码的长度。假设我有真实的概率分布，我可以把概率大的用短编码，把概率小的用长编码，这样整体编码就会变短。但是如果我用的是错误的概率，那么就不能达到最短的编码长度。