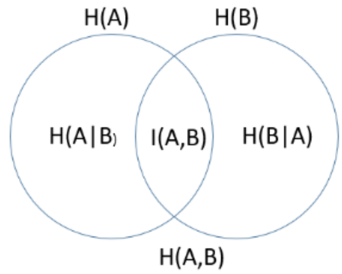
研究生算法课课堂笔记

上课日期：　2016.9.19　　　　　　　　**第(1)节课**

组长学号及姓名：王文军　1601214486

组员学号及姓名：贾　然　1601214458　高成良　1601214429

1. 事件 E，发生概率为$P(E)$，定义 $I(E) = \log\frac{1}{P(E)}$为事件的信息量：
   a) 小概率事件信息量大，大概率事件信息量小。
   b) 最小值为 0 (概率为 1 时)，最大值为无穷(当事件为不可能事件时)。
2. 熵：
   a) 当实验的次数趋于无穷时，平均得到的信息量，刻画随机变量本身的不确定性。
   b) 当随机事件 A 有 k 种可能的结果时，熵的公式为：
   $$H(A) = \sum_{i=1}^{k} p_i \log\frac{1}{p_i}。$$
   c) 最小值为 0，最大值为$log\ k$ (事件 A 出现每种结果的概率都相等时)。
3. 条件熵：
   a) 两个随机事件 A 和 B，已知 B 发生时，A 残留的不确定性，记为$H(A|B)$。
   b) 最小值为 0(当 A 和 B 完全相关时)；最大值为 A 的熵(当 A 和 B 相互独立时)。
4. Mutual information(互信息)：
   a) $I(A,B) = H(A) - H(A|B) = H(B) - H(B|A)$
   $$= \sum_{a,b} P(a,b)\, log\frac{P(a,\ b)}{P(a)P(b)}$$
   b) 如果两个事件独立，互信息为 0；不独立时，互信息大于零。
   c) 区别于熵，熵是平均不确定性的减少(average reduction of uncertainty)。
5. 假设 $f(X) = Y$，则有 $\mathrm{H(X)} \geq H(Y)$：
   a) 直观上，函数的映射消除了一部分不确定性。
   b) 根据6，且有$H(Y|X) = 0$推出$H(\mathrm{X}) = H(Y) + H(X|Y)$。
   c) f为一一映射时等号成立。
6. 熵的 Chain Rule(链式法则)：$H(A) + H(B|A) = H(B) + H(A|B)$
7. $\mathrm{H(A,B)} = \sum_{a,b} P(a,b) \log\frac{1}{P(a,b)}$
8. $\mathrm{H(A|B)} = \sum_{a,b} P(a,b) \log\frac{1}{P(a|b)}$
9. $H(A,B) \leq H(A) + H(B)$，当 A 和 B 相互独立时等号成立：
10. 文氏图 – 只有两个随机变量的时候是正确的，三个随机变量时不一定成立，如下所示：



11. 两个随机变量，条件熵是否对称，即$H(A|B)$与 $H(B|A)$是否相等：
    a) 不对称，无法判断大小关系。
    b) 举例：事件 A 为从 1024 个数字中随机选取一个，事件 B 为判断随机选择的数字是奇数还是偶数，则$H(A|B) = 9$，$H(B|A) = 0$。
12. Mutual Information 是否满足三角不等式：
    a) 不满足，即$I(A,B) + I(B,C)$不一定大于$I(A,C)$。
    b) 例如：事件 B 为早餐是否吃包子，事件 A 为算法课模拟考试成绩，事件 C 为期末考试成绩。