研究生算法课课堂笔记

上课日期：2016 年 9 月 26 日　　第(1)节课

组长学号及姓名：张晓德 1601214529
组员学号及姓名：孙鹏晖 1601214522
组员学号及姓名：曾　立 1601214526

一、 内容概要

本节课内容主要包括以下几点：

1、 上节课复习：条件熵（Conditional Entropy）与互信息（Mutual Information）

2、 交叉熵（Cross Entropy）

3、 KL-Distance (KL-divergence)

二、 详细内容

1、 上节课复习：条件熵（Conditional Entropy）与互信息（Mutual Information）

回顾熵的概念，事件 A 的熵可以表示为

$$\mathrm{H(A)} = \sum_{i=1}^{k} p_i log \frac{1}{p_i}$$

但是我们不能根据条件概率公式将条件熵写为

$$\mathrm{H(A|B)} = \sum_{a,b} p(a|b) log \frac{1}{p(a|b)}$$

条件熵应该用联合概率进行加权，正确的表达为

$$\mathrm{H(A|B)} = \sum_{a,b} p(a，\ b) log \frac{1}{p(a|b)}$$

由于在给定 a 和 b 时，$p(a, b)$是固定的，而一般情况下$p(a|b)$和$p(b|a)$不相等，所以条件熵不具有对称性。

互信息（Mutual Information）可以理解为已知一个随机变量而导致另一个随机变量不确定性的减少，即：

$$\mathrm{I(A;B)} = \mathrm{H(A)} - \mathrm{H(A|B)}$$

$$= \sum_{a} p(a) \log \frac{1}{p(a)} - \sum_{a,\mathrm{b}} p(a,b) \log \frac{1}{p(a|b)}$$

$$= \sum_{a,b} p(a,b) \log \frac{1}{p(a)} - \sum_{a,\mathrm{b}} p(a,b) \log \frac{1}{p(a|b)}$$

$$= \sum_{a,b} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

从上式可以看出，互信息具有对称性，但是注意互信息不满足三角不等式。

推导时用到下面的等式：

$$\sum_{a,b} p(a,b) \log \frac{1}{p(a)}$$
$$= \sum_{a} \sum_{b} p(a)p(b|a) log \frac{1}{p(a)}$$
$$= \sum_{a} p(a) log \frac{1}{p(a)} \sum_{b} p(b|a)$$

注意到$\sum_b p(b|a) = 1$，所以互信息的表达式正确。

2、 交叉熵（Cross Entropy）

$$\mathrm{CH(A,B)} = \sum p(a) \frac{1}{p(b)}$$

注意：如果 A 和 B 是两个不相关的随机变量，计算它们的交叉熵是没有意义的。交叉熵用来描述同一个随机变量的真实概率分布和估计概率分布之间的关系，而不是不同的随机变量之间的关系。相比之下，前面提到的互信息和条件熵都是在研究不同的随机变量之间的关系。

现在考虑对某个随机变量 X，其真实的概率分布为 p，但我们认为其概率分布为 q，即 p 和 q 是同一个随机变量的两种不同的概率分布。定义交叉熵如下

$$\mathrm{CH(p,q)} = \sum_{i=1}^{k} p_i log \frac{1}{q_i}$$

可以看出，交叉熵不具有对称性，反例如 p=(1/2, 1/2)和 q=(1, 0)。

评价两种分类方法的好坏，如果单纯地考虑正确率，有时候并不能全面地反映出算法的好坏。可以考虑用交叉熵进行评价。假设真实的概率分布为 p，算法给出的概率分布为 q。在 test set 上进行 n 次实验，令 loss function 或者叫 cost function 为

$$f = \frac{1}{\mathrm{n}} (\log \frac{1}{c_1} + \log \frac{1}{c_2} + \cdots + \log \frac{1}{c_n})$$

其中$c_1 \ldots c_n$(可以相同)取自$q_1 \ldots q_{\mathrm{k}}$，其中 k 为实验中所有可能的情况数（对分类任务来说，就是类别的总数）。当 n 趋于无穷时，上式趋于

$$\lim_{n \to \infty} f = \sum_{i=1}^{k} p_i log \frac{1}{q_i} = Ep\left(log \frac{1}{q_i}\right) = CH(p,q)$$

即极限情况下损失函数趋于交叉熵。我们希望损失函数的取值越小越好，这样说明算法的预测性能越好（least surprised）。考虑下面几种情况对损失函数的影响：

$p_i = 0, q_i \neq 0$	损失函数中不出现第 i 种情况，这时q_i的取值对损失函数没有直接影响，但是会间接影响其它情况的概率（因为加起来要等于 1）
$p_i = 0, q_i = 0$	预测的概率与真实的概率相等，这是一种理想的情况
$p_i \neq 0, q_i = 0$	$log\frac{1}{q_i}$会趋于无穷大，导致损失函数的值非常大

从上述分析可以得出，如果算法给出的概率为 0，则可能会因为对个别样本的误判导致对算法的整体评价非常不好，所以一般将概率设定在区间（ε，$1-\varepsilon$）.

课上老师举了 logistic regression 的损失函数的例子，比如二分类问题中正例的概率为 p，目标值是 *y*，其损失函数为

$$f = -\mathrm{ylogp} - (1-\mathrm{y})\log(1-p),\ y = \{0,1\}$$

y = 1 表示真实的情况为正例，此时损失函数为$f = -logp$，如果算法给出的预测 p=0，则损失函数值为无穷大；*y*=0 表示真实的情况为反例，此时损失函数为$f = -\log(1-p)$，如果算法给出的预测 *p*=1，则损失函数值也为无穷大。

3、 KL-Distance

考虑对某个随机变量 X，其真实的分布为 p，但我们认为其分布为 q，定义其 KL-Distance 如下：

$$\mathrm{KL(p||q)} = \mathrm{CH(p,q)} - \mathrm{H(X)}$$

$$= \sum_{i=1}^{k} p_i log\frac{1}{q_i} - \sum_{i=1}^{k} p_i log\frac{1}{p_i}$$

$$= \sum_{i=1}^{k} p_i \log\frac{p_i}{q_i}$$

$\mathrm{I(A,B)}$与$\mathrm{KL-distance}$具有如下关系:

$$\mathrm{I(A;B)} = KL\big(p(a,b)\,||\,p(a)p(b)\big)$$

$$= \sum_{a,b} p(a,b)log\frac{1}{p(a)p(b)} - \sum_{a,b} p(a,b)log\frac{1}{p(a,b)}$$

可以看出上面两式相减后的结果正是 I(A;B)的定义，被减式是 CH(p(a,b), p(a)p(b))，减式则是 H(A,B)。

KL-Distance 并不是真正的距离，因为它不满足距离定义的三个条件：非负性，对称性和三角不等式。KL-Distance 只满足非负性，不具有对称性和三角不等式的性质。

很容易举出反例来说明 KL 距离不具有对称性，如分布 *p* 为(1/2, 1/2)，分布 *q* 为(1/4, 3/4)，计算可知 KL(*p*||*q*)与 KL(*q*||*p*)不等。

上课时讨论了 KL-Distance 不满足三角不等式的性质，但没有给出反例。我们小组课后进行了讨论，结果如下：

假设有三个概率分布 *p*，*q* 和 *r*，三角不等式是指KL(*p*||*q*) + KL(q||r) ≥ KL(p||r)。我们需要构造反例来说明存在这样的三个概率分布，使得：

$$
\begin{aligned}
&KL(p \,||\, q) + KL(q \,||\, r) - KL(p \,||\, r) \\
&= \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i} + \sum_{i=1}^{k} q_i \log \frac{q_i}{r_i} + \sum p_i \log \frac{r_i}{p_i} \\
&= \sum_{i=1}^{k} p_i \log (\frac{p_i}{q_i} \times \frac{r_i}{p_i}) - \sum_{i=1}^{k} q_i \log \frac{r_i}{q_i} \\
&= \sum_{i=1}^{k} (p_i - q_i) \log \frac{r_i}{q_i} < 0
\end{aligned}
$$

令 *p* 为(0.6, 0.4)，*q* 为(0.4, 0.6)，*r* 为(*x*, *y*)，则上式可化为

$$0.2 \log \frac{x}{0.4} - 0.2 \log \frac{y}{0.6} < 0，其中 x + y = 1。$$

当 *x* 和 *y* 分别取 0.4 和 0.6 时，左式结果为 0；当 *x* 和 *y* 分别取 0.3 和 0.7 时，左式是一个负数减去一个正数，结果必然小于 0（所以三角不等式不一定成立）；当 *x* 和 *y* 分别取 0.7 和 0.3 时，左式是一个正数减去一个负数，结果必然大于 0。这个反例的精髓在于：当 q 和 r 的概率分布与 p 的概率分布相比越偏越远时，以 q 做为中介计算出来的 KL 距离不如直接计算 p 和 r 的 KL 距离大。

三、 总结

概念	适用对象	公式	对称性	三角不等式	最值
Conditional Entropy	不同的随机变量之	$H(A\|B) = \sum_{a,b} p(a, b) log \frac{1}{p(a\|b)}$	否	-	[0, H(A)]

Mutual Information	间	$I(A;B)=\sum_{a,b} p(a,b)\log\frac{p(a,b)}{p(a)p(b)}$	是	否	[0, min{H(A), H(B)}]
Cross Entropy	同一随机变量的不同分布	$CH(p,q)=\sum_{i=1}^{k} p_i log\frac{1}{q_i}$	否	否	[H(X), ∞)
KL-Distance		$KL(p\|\|q)=CH(p,q)-H(X)$	否	否	[0, ∞)

四、 附录

（该部分内容为自主探讨，敬请批评指正）

1. 交叉熵不满足三角不等式，即存在分布 p、q、r，使得 CH(p,q)+CH(q,r)-CH(p,r)<0，注意 p、q、r 都是对同一随机事件 X 的描述。

我们构造这样一种分布，使得 p 和 q 差异很小，q 和 r 差异很小，但 p 和 r 差异却相对较大。这样的话 CH(p,q)和 CH(q,r)都较小，但 CH(p,r)可以很大，按这个思路我们选择分布 p 为(1/4+t, 1/4-t, 1/4, 1/4)，分布 q 为(1/4, 1/4, 1/4, 1/4)，分布 r 为(1/4-t, 1/4+t, 1/4, 1/4)。

计算出 CH(p||q)=2，CH(q||r)= $1-\frac{1}{4}\log(\frac{1}{16}-t^2)$，而 CH(p||r)=CH(q||r)+ $t\log\frac{\frac{1}{4}+t}{\frac{1}{4}-t}$

要构造反例，只需要使得$t\log(1+\frac{8}{\frac{1}{t}-4})$>2，其中$0<t<\frac{1}{4}$。注意到左式是一个连续函数，t=0 时左式为 0，t=1/4 时左式为正无穷大，所以满足条件的 t 一定是存在的。

事实上我们选择$t=\frac{1023}{4096}$即可使左式大于 2，在这个概率分布下交叉熵不满足三角不等式。

（以此也可说明 KL 距离不满足三角不等式。注意虽然三个概率分布描述同一个事件，但

在选择一个分布为真实的时候，H(X)也就被改变为那个分布对应的熵，所以不能直接忽略KL距离中的H(X)部分）

2. 对最值的讨论

H(X)的最小值是0，当且仅当$p_i = 1$。它的最大值是 log k(k表示可能的情况数)，当且仅当每种可能性的概率都相等。

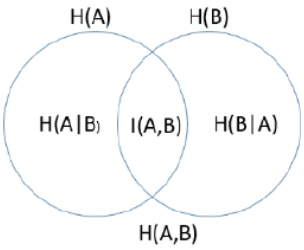
H(A|B)的最小值是0，当B事件的结果能决定A事件的结果时或者A的熵本来就是0时。它的最大值是H(A)，当且仅当A和B是独立的。

CH(p,q)的最小值是H(X)，当且仅当q分布和p分布完全相同时。它的最大值是正无穷，比如当$p_i \neq 0$而$q_i = 0$时，整体趋向无穷大。

由CH的最值可以推导得到KL(p||q)的取值区间是[0, ∞)，在给定事件X的情况下，H(X)是用真实概率分布p算出来的，变化的只是估计概率分布q。

I(A;B)=H(A)-H(A|B)=H(B)-H(B|A)，其最小值是0，当且仅当A和B完全无关。值得说明的是最大值：当A和B等价时，取得H(A)=H(B)；当A能决定B时，H(B|A)=0<H(A|B)，所以H(A)>H(B)，此时取得H(B)；当B能决定A时，H(A|B)=0<H(B|A)，所以H(A)<H(B)，此时取得H(A)；否则，H(A|B)>0，H(B|A)>0，I(A;B)<H(A)且I(A;B)<H(B)。综上所述，I(A;B)的最大值是min{H(A), H(B)}。

(这一结论可以从文氏图直观地得到，当H(A)和H(B)完全分离时，I(A;B)=0取最小值；当有一个被另一个全部包含时，I(A;B)取得最大值min{H(A), H(B)}.)



（此图摘自上次同学笔记）

3. H(B)=0时的推论(当且仅当某个$p_i = 1$时取得)

此时H(A|B)=H(A,B)-H(B)=H(A,B)，$H(A,B) \geq H(A)$，而$H(A \mid B) \leq H(A)$，所以H(A|B)=H(A)。$H(B \mid C) \leq H(B) = 0$，所以H(B|C)=0。又$H(A \mid C) \leq H(A)$，

所以在H(B)=0时有$H(A \mid B) + H(B \mid C) \geq H(A \mid C)$。