马飞宇的第一次信息论作业

mfy

2024 年 9 月 26 日

1 数据读入

并没有什么好说的，用 pandas 的读 read_excel 函数就好了，值得注意的是为了进行矩阵运算或者向量运算，我们将 dataframe 转化为了 np 中的矩阵。

在下面，为同时适用于两组数据，我们做出如下说明:

- 认为矩阵是 m*n 的，即 m 行 n 列。
- 记 x，y 的分布服从向量 p_x, p_y, 即 x 取第一种取值概率为 p_x 的第一个元素，p_x, p_y 均为行向量。

2 data1 公式分析

2.1 H(X)

首先我们观察数据，发现数据是二维离散的随机变量，然后，我们可以得知如仅需计算 X 的熵，那么需要求出 X 的概率分布情况：

$$p(x_j) = \sum_{i=1}^{m} p_{i,j} \tag{1}$$

进一步地，根据上面计算所得的概率，我们可以得到 X 的熵为：

$$H(X) = -\sum_{j=1}^{n} p(x_j) * \log_2 p_(x_j) \tag{2}$$

2.2 H(Y)

类似的，Y 的熵计算方法如下：

$$p(y_i) = \sum_{j=1}^{n} p_{i,j} \tag{3}$$

进一步地，根据上面计算所得的概率，我们可以得到 X 的熵为：

$$H(Y) = -\sum_{i=1}^{m} p(y_i) * \log_2 p_(y_i) \tag{4}$$

2.3 H(X,Y)

我们知道，对离散情形下，想要得到联合熵，有公式：

$$H(X\ Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p_{i,j} * \log_2 p_{i,j} \tag{5}$$

显然在读入数据之后只需按上面的公式进行计算即可得出结论，但需要注意的是，由于数据文件中有零元素，所以如果希望代码不报错，且得到我们真正希望的结果，就需要将 $p_{i,j} == 0$ 的情形考虑进来，并在其为真时不做操作（或者按约定进行加零，但实际不操作就可以，我之所以写了是因为我不太喜欢不操作）

2.4 H(X|Y)

我们知道，条件熵在上述离散情形下有：

$$H(X|Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p_{i,j} * \log_2 p(y|x = x_i) \tag{6}$$

上面出现的唯一新变量就是 $p(y|x)$, 通过概率论中的知识，不难知道：

$$p(y|x=x_i)=\frac{p_{i,j}}{p(x_i)} \tag{7}$$

然后按上述公式进行运算即可，同样需要注意零元素。

2.5 H(Y|X)

我们知道，条件熵在上述离散情形下有：

$$H(Y|X)=-\sum_{j=1}^{n}\sum_{i=1}^{m}p_{i,j}*\log_2 p(x|y=y_j) \tag{8}$$

上面出现的唯一新变量就是 $p(x|y)$, 通过概率论中的知识，不难知道：

$$p(x|y=y_j)=\frac{p_{i,j}}{p(y_j)} \tag{9}$$

然后按上述公式进行运算即可，同样需要注意零元素。

2.6 I(X;Y)

我们知道，互信息的公式为：

$$I(X;Y)=-\sum_{i=1}^{4}\sum_{j=1}^{4}p_{i,j}\log_2\frac{p_{i,j}}{p(x_i)p(y_i)} \tag{10}$$

代码中我们只需要修改循环就好了，也就是说并无特别的操作，尤其在该情况下 $p(x_i),p(y_i)>0$, 并不需要增加一层判断，虽然建议这样做，以在更多情形下适用。

3 data2 公式分析

只是将数据又四行四列变成三行四列，易得。