作业 2 随机向量的分布

作业目的: 体会条件独立

1、现需要设计一个根据一个人是否是学生 S（布尔变量）和其体重 W（连续变量）判断该人的性别 G（布尔变量）。假设在给定 G 的情况下 S 和 W 独立，且假设概率分布 $p(W \mid G = female)$ 和 $p(W \mid G = male)$ 为高斯分布且二者的方差相等。

（a）可以用朴素贝叶斯分类器实现吗？

（b）如果可以用朴素贝叶斯分类器的话，需要估计从训练数据中估计哪些分布的哪些参数。

2、体会条件独立带来模型参数的减少

考虑一个C个类别的产生式分类器，其中类条件概率密度为$p(\mathbf{x}|y)$ ，假设类先验$p(y)$为均匀分布。假设D维特征均为二值变量，即$x_j \in \{0, 1\}$。假设在给定类别的条件下，各个特征独立（朴素贝叶斯假设），我们可以记

$$p(\mathbf{x} \mid y = c, \theta) = \prod_{j=1}^{D} Ber(x_j \mid \theta_{jc}),$$

模型共需要DC个参数。

(a) 考虑一个不同的“全”模型，即所有变量都相关。则条件概率$p(\mathbf{x}|y = c)$应该是什么样子？表示$p(\mathbf{x}|y = c)$需要多少个参数？

(b) 当样本数目N较小时，条件独立模型和全模型哪个模型的性能会更好？

(c) 当样本数目N较大时，上述两个模型哪个模型的性能更好？