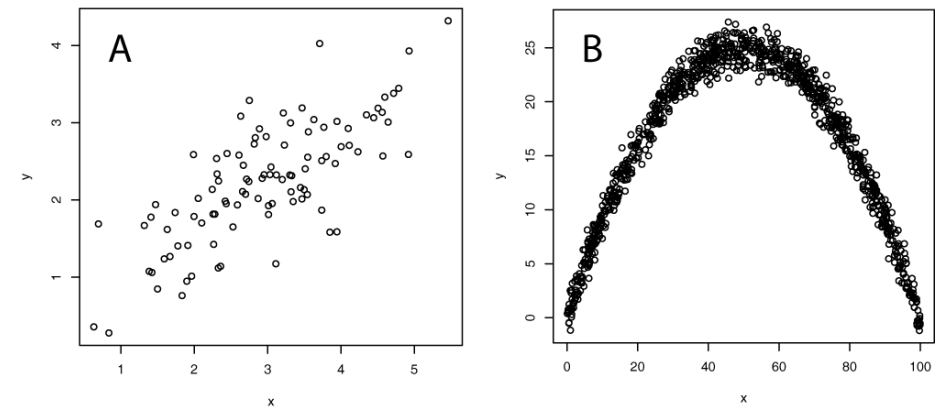
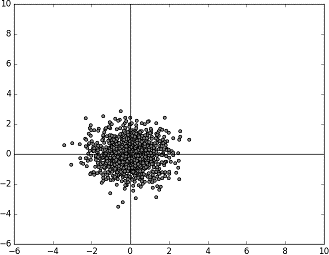
姓名： 学号：

1. 我们知道，对二元属性对象，度量其邻近度的方法有SMC、Jaccard系数、汉明距离等。现有以下两个二元向量：。针对以下两种任务，选择合适的邻近度，简要说明选择的原因，并计算相应的邻近度。
2. 假设为两个不同物种的基因，根据共同含有的基因个数比较两个物种的邻近度。
3. 假设为相同物种的两个个体的基因，根据基因比较两个个体间差异。

解：1）比较两个物种共同含有的基因，我们不关注两个物种都不含有的基因，因此两个向量所代表的属性为非对称属性。应使用Jaccard系数比较邻近度：

1. 为相同物种。设想以下情况，如果两个个体都不含有这个物种可能含有的某个基因，则我们认为这两个个体是有相似性的。因此两个向量所代表的属性为对称属性。应使用SMC比较邻近度：

对于上述情况，计算汉明距离也是合理的，因为汉明距离实际表示了。此处汉明距离为3。

1. 分别针对以下三图中对象在二维平面的分布情况，你认为使用主成分分析（PCA）算法进行降维是否合适？若合适，在图上画出主成分维度的大致方向。若不合适，请简要说明原因。

**C**

解：A适合使用PCA降维，因为数据呈多维高斯分布，且两属性间存在线性相关性，因此使用PCA降维可保留原始数据最多的信息。主成分方向如图。

B不适合使用PCA降维，因为属性间存在高阶相关性（非线性相关），因此舍弃掉的次要成分可能包含反映原始数据情况的重要信息。

注：如果将B图中的数据映射到高维空间，则是可以使用PCA降维的。在这种情况下，需要利用核函数（kernel function）去求高维空间上协方差的特征值/向量。该方法被称为Kernel PCA (KPCA)。

C不适合使用PCA降维。因为属性间不存在线性相关性，舍弃任何次要成分都可能造成大量信息的丢失。