任务2.1

鸢尾花数据可视化

1. 参考任务1.2有关鸢尾花数据的相关介绍，了解数据各列的含义。

2. 对每种类型的鸢尾花的四个属性进行可视化绘图，观察各属性在不同类型间的相关性及区分度。

注：

1. 数据可以选择使用任务1.2步骤3清理之后的数据，也可使用sklearn包自带的载入数据方式，参考代码如下：

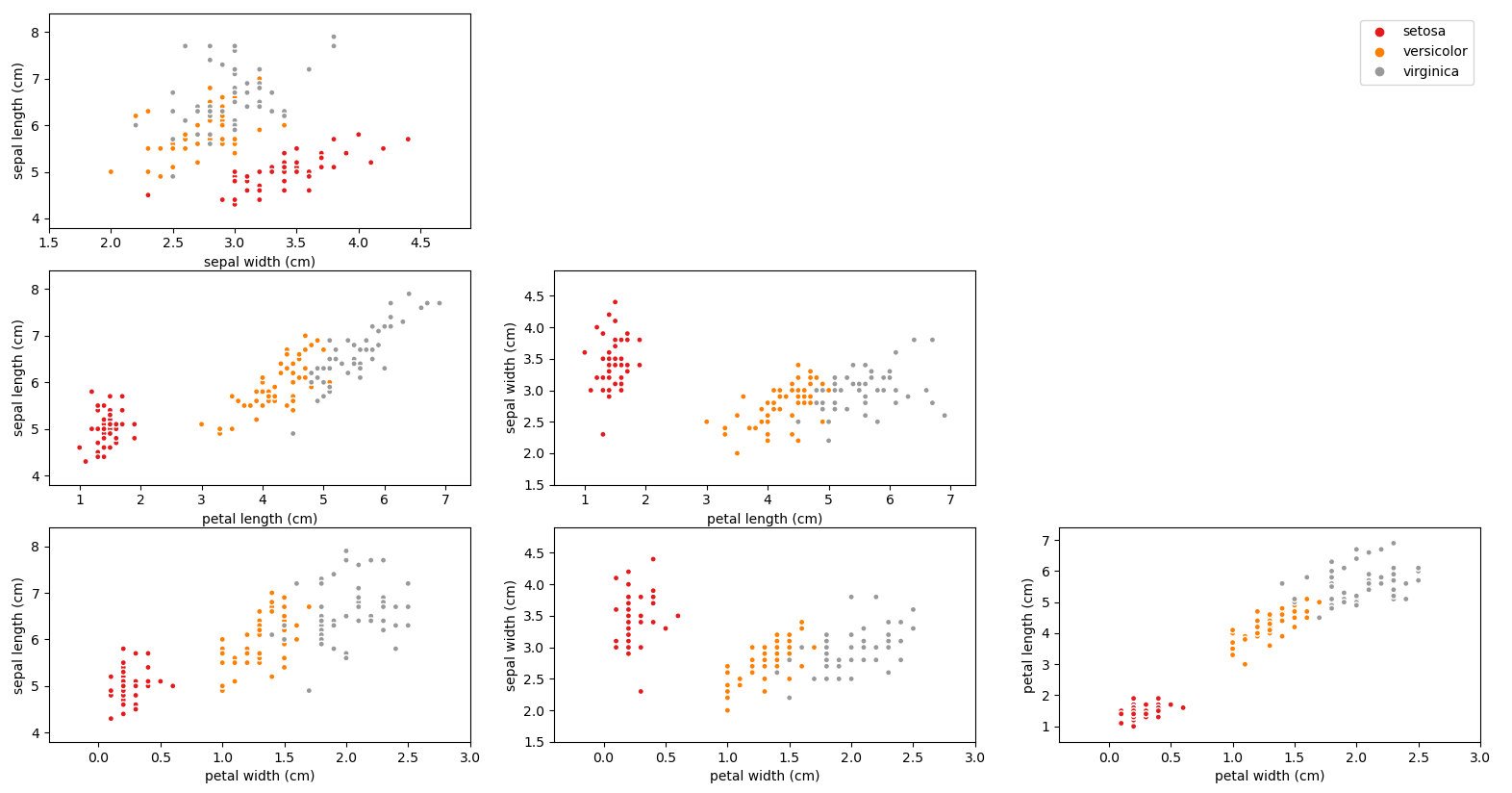
from sklearn import datasets

iris = datasets.load\_iris()

2. 画图可以使用matplotlib包。

3. 可依次选取每项特征绘制，例如，以萼片长度为x轴、萼片宽度为y轴作散点图进行观察。完成后继续观察下一组特征。

4. 结果可参见下图示例。

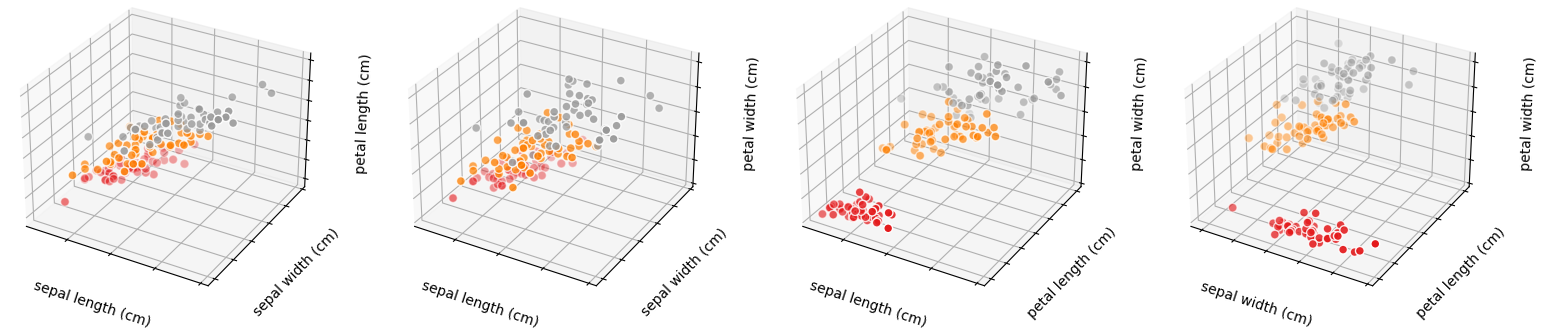


任务2.2

鸢尾花数据的3D可视化

使用如任务2.1中的可视化方法只能观察两两特征间的关系。然而，真实数据中，仅用两组数据往往不足以完成分类问题。例如在上例中，萼片长度和叶片宽度这组数据，尚可满足区分山鸢尾和其它两类鸢尾花的需求，但注意观察图中黄色和灰色部分具有大量重叠数据，明显无法对杂色鸢尾与维吉尼亚鸢尾进行区分。其它散点图中也可发现杂色鸢尾与维吉尼亚鸢尾具有不少重叠。为更好的区分三类鸢尾花，我们可以尝试添加新的特征进行可视化绘图。与任务2.1类似，请依次选取3个特征绘制3D散点图，观察是否足以区分三类花型。

注：绘制3D图可尝试使用mpl\_toolkits.mplot3d库中的Axes3D，示例结果如下：



任务2.3（optional）

使用PCA（主成分分析）对数据进行降维

在任务2.3中尝试选取不同的3项特征重复实验，我们依然无法去除重叠的数据，需要考虑引入全部4种特征进行分类。然而，4项特征全部引入后我们将难以进行数据可视化（无法绘制四维视图），因此我们考虑使用PCA，利用正交变换将以线性相关变量表示的观测数据转化为少数几个以线性无关变量表示的数据，对鸢尾花的四维数据进行降维，然后绘制三维视图。

PCA算法的具体流程请自行在网上搜索研习，注意，其分解后的主成分就是其协方差矩阵对应的特征向量，按照对应的特征值大小进行排序，最大的特征值就是第一主成分，其次是第二主成分，以此类推。

注：PCA算法可考虑使用sklearn.decomposition库。因该算法相对比较简单，对于希望能够更深入理解该算法的同学也欢迎尝试自行实现。