鸢尾花数据集，数据集包含3类共150调数据，每类含50个数据，每条记录含4个特征：花萼长度、花萼宽度、花瓣长度、花瓣宽度

过这4个 特征预测鸢尾花卉属于（iris-setosa, iris-versicolour, iris-virginica）中的哪一品种

In [1]:

```
!head -5 data/Iris.csv
```

```
0,Sepal.Length,Sepal.Width,Petal.Length,Petal.Width,Species
1,5.1,3.5,1.4,0.2,setosa
2,4.9,3,1.4,0.2,setosa
3,4.7,3.2,1.3,0.2,setosa
4,4.6,3.1,1.5,0.2,setosa
```

In [2]:

```
##创建环境

import findspark
findspark.init()
################################################
from pyspark.sql import SparkSession
from pyspark.sql.types import *

spark = SparkSession.builder \
        .master("local[*]") \
        .appName("KNNrdd") \
        .getOrCreate()

sc = spark.sparkContext
```

In [3]:

```
#获得距离
import math
def getDis(array1, array2):
    return math.sqrt(sum([pow(x-y, 2) for x, y in zip(array1, array2)]))
#    return math.sqrt(array1,zip(array2).map(lambda x:pow(x[0]-x[1])).sum)
```

In [4]:

```
K = 15
data = sc.textFile("data/Iris.csv").map(lambda x:x.split(",")).filter(lambda x:x[1]!= "Sepal.L
```

In [5]:

```
data.collect()
```

```
 ('setosa', [27.0, 5.0, 3.4, 1.6, 0.4]),
 ('setosa', [28.0, 5.2, 3.5, 1.5, 0.2]),
 ('setosa', [29.0, 5.2, 3.4, 1.4, 0.2]),
 ('setosa', [30.0, 4.7, 3.2, 1.6, 0.2]),
 ('setosa', [31.0, 4.8, 3.1, 1.6, 0.2]),
 ('setosa', [32.0, 5.4, 3.4, 1.5, 0.4]),
 ('setosa', [33.0, 5.2, 4.1, 1.5, 0.1]),
 ('setosa', [34.0, 5.5, 4.2, 1.4, 0.2]),
 ('setosa', [35.0, 4.9, 3.1, 1.5, 0.2]),
 ('setosa', [36.0, 5.0, 3.2, 1.2, 0.2]),
 ('setosa', [37.0, 5.5, 3.5, 1.3, 0.2]),
 ('setosa', [38.0, 4.9, 3.6, 1.4, 0.1]),
 ('setosa', [39.0, 4.4, 3.0, 1.3, 0.2]),
 ('setosa', [40.0, 5.1, 3.4, 1.5, 0.2]),
 ('setosa', [41.0, 5.0, 3.5, 1.3, 0.3]),
 ('setosa', [42.0, 4.5, 2.3, 1.3, 0.3]),
 ('setosa', [43.0, 4.4, 3.2, 1.3, 0.2]),
 ('setosa', [44.0, 5.0, 3.5, 1.6, 0.6]),
 ('setosa', [45.0, 5.1, 3.8, 1.9, 0.4]),
 ('setosa', [46.0, 4.8, 3.0, 1.4, 0.3]),
```

In [6]:

```
#准备样本数据和测试数据
splits=data.randomSplit([0.7,0.3])
sampleData=splits[0]
testData=splits[1]
```

In [7]:

```
sampleDataBrc = sc.broadcast(sampleData.collect())
```

In [8]:

```
#获取距离最近的K个样本
def getPoint(elem,K):
    print("begin:elem")
#     dis = sampleDataBrc.value.map(lambda x:(getDis(elem[1],x[1]),x[0]))
    dis = [(getDis(elem[1], x[1]), x[0]) for x in sampleDataBrc.value]
#     dis = list(elem.map(lambda x:(getDis(elem[1],x[1]),x[0])))
    #获取距离最近的k个样本
    minDis = sorted(dis,key=lambda x:x[0],reverse=False)[:K]
    minDisLabel = [x[0] for x in minDis]
    print(minDis)
    #取出这k个样本的label并且获取出现最多的label即为测试数据的label
    labels = sorted(minDis,key=lambda x:minDisLabel.count(x[0]),reverse = True)
    return "{},{}=>{}".format(elem[0], elem[1], labels[0][1])
```

In [9]:

```
testData = testData.map(lambda x: getPoint(x, K))
```

In [10]:

```
testData.collect()
```

Out[10]:

```
['setosa,[2.0, 4.9, 3.0, 1.4, 0.2]=>setosa',
 'setosa,[9.0, 4.4, 2.9, 1.4, 0.2]=>setosa',
 'setosa,[10.0, 4.9, 3.1, 1.5, 0.1]=>setosa',
 'setosa,[13.0, 4.8, 3.0, 1.4, 0.1]=>setosa',
 'setosa,[17.0, 5.4, 3.9, 1.3, 0.4]=>setosa',
 'setosa,[20.0, 5.1, 3.8, 1.5, 0.3]=>setosa',
 'setosa,[21.0, 5.4, 3.4, 1.7, 0.2]=>setosa',
 'setosa,[26.0, 5.0, 3.0, 1.6, 0.2]=>setosa',
 'setosa,[27.0, 5.0, 3.4, 1.6, 0.4]=>setosa',
 'setosa,[30.0, 4.7, 3.2, 1.6, 0.2]=>setosa',
 'setosa,[31.0, 4.8, 3.1, 1.6, 0.2]=>setosa',
 'setosa,[33.0, 5.2, 4.1, 1.5, 0.1]=>setosa',
 'setosa,[36.0, 5.0, 3.2, 1.2, 0.2]=>setosa',
 'setosa,[39.0, 4.4, 3.0, 1.3, 0.2]=>setosa',
 'setosa,[41.0, 5.0, 3.5, 1.3, 0.3]=>setosa',
 'setosa,[43.0, 4.4, 3.2, 1.3, 0.2]=>setosa',
 'versicolor,[51.0, 7.0, 3.2, 4.7, 1.4]=>versicolor',
 'versicolor,[61.0, 5.0, 2.0, 3.5, 1.0]=>versicolor',
 'versicolor,[62.0, 5.9, 3.0, 4.2, 1.5]=>versicolor',
 'versicolor,[74.0, 6.1, 2.8, 4.7, 1.2]=>versicolor',
 'versicolor,[75.0, 6.4, 2.9, 4.3, 1.3]=>versicolor',
 'versicolor,[77.0, 6.8, 2.8, 4.8, 1.4]=>versicolor',
 'versicolor,[78.0, 6.7, 3.0, 5.0, 1.7]=>versicolor',
 'versicolor,[80.0, 5.7, 2.6, 3.5, 1.0]=>versicolor',
 'versicolor,[82.0, 5.5, 2.4, 3.7, 1.0]=>versicolor',
 'versicolor,[87.0, 6.7, 3.1, 4.7, 1.5]=>versicolor',
 'versicolor,[89.0, 5.6, 3.0, 4.1, 1.3]=>versicolor',
 'versicolor,[94.0, 5.0, 2.3, 3.3, 1.0]=>versicolor',
 'versicolor,[98.0, 6.2, 2.9, 4.3, 1.3]=>versicolor',
 'virginica,[101.0, 6.3, 3.3, 6.0, 2.5]=>virginica',
 'virginica,[108.0, 7.3, 2.9, 6.3, 1.8]=>virginica',
 'virginica,[112.0, 6.4, 2.7, 5.3, 1.9]=>virginica',
 'virginica,[115.0, 5.8, 2.8, 5.1, 2.4]=>virginica',
 'virginica,[116.0, 6.4, 3.2, 5.3, 2.3]=>virginica',
 'virginica,[117.0, 6.5, 3.0, 5.5, 1.8]=>virginica',
 'virginica,[124.0, 6.3, 2.7, 4.9, 1.8]=>virginica',
 'virginica,[130.0, 7.2, 3.0, 5.8, 1.6]=>virginica',
 'virginica,[132.0, 7.9, 3.8, 6.4, 2.0]=>virginica',
 'virginica,[135.0, 6.1, 2.6, 5.6, 1.4]=>virginica',
 'virginica,[136.0, 7.7, 3.0, 6.1, 2.3]=>virginica',
 'virginica,[137.0, 6.3, 3.4, 5.6, 2.4]=>virginica',
 'virginica,[138.0, 6.4, 3.1, 5.5, 1.8]=>virginica',
 'virginica,[142.0, 6.9, 3.1, 5.1, 2.3]=>virginica',
 'virginica,[143.0, 5.8, 2.7, 5.1, 1.9]=>virginica',
 'virginica,[144.0, 6.8, 3.2, 5.9, 2.3]=>virginica',
 'virginica,[149.0, 6.2, 3.4, 5.4, 2.3]=>virginica']
```

In [ ]:

```

```