In [ ]:

```
knn.kneighbors_graph(iris.data[:1]).toarray()
```

In [ ]:

```
knn.radius_neighbors(iris.data[:1])
```

In [ ]:

```
knn.radius_neighbors(iris.data[:1], radius = 0.2)
```

11.5 实战练习

对iris数据进行标化，然后重新拟合俩终最近邻分类方法，观察其分析结果的变化，并思考原因。

对boston数据进行KNN回归，并进行参数调优，找到最优模型。

提示：需要将数据拆分为训练集和验证集进行结果验证。

尝试使用SVM分类方法对logit表单数据进行建模分析。

12 生存分析

12.1 生存分析的基本概念

12.2 计算生存概率

12.2.1 生存曲线的计算

class statsmodels.duration.survfunc.SurvfuncRight(

```
time : 时间变量
status : 生存结局变量，1代表事件发生，0代表截尾
entry = None : 进入时间变量，在该时点之前案例并未暴露在风险中
title = None : 生存分析图表中使用的标题
freq_weights = None
exog = None : 生存状态的影响因素
bw_factor = 1.0 : Band-width multiplier for kernel-based estimation
```

)

statsmodels.duration.survfunc.SurvfuncRight类的方法：

```
plot([ax]) : 生存分析曲线
quantile(p) : 指定生存概率所对应的生存时间
quantile_ci(p[, alpha, method]) : 对应生存时间的可信区间
simultaneous_cb([alpha, method, transform]) : 生存函数的置信带
summary() : 模型分析结果的汇总
```

分析实例

KM表单是慢性活动性肝炎的临床试验数据（Altman and Bland，1998），列出了44名慢性活动性肝炎患者的生存时间（月）。这些患者被随机分配至Prednisolone新药组或对照组，每组22名。之后对这些患者进行随访，记录他们死亡发生的时间直至研究结束。

months：患者的生存时间（月）
status：是否删失，没有删失0，失访1，研究结束时仍存活2
group：组别，Prednisolone新药组1，对照组2

In [ ]:

```
dfsuv = pd.read_excel('DmData.xlsx', sheet_name = 'KM')

dfsuv.head()
```

In [ ]:

```
from sklearn.preprocessing import binarize

dfsuv['event'] = 1- binarize(pd.DataFrame(dfsuv.status))
dfsuv.head(10)
```

In [ ]:

```
suv = sm.SurvfuncRight(dfsuv.month, dfsuv.event)
```

In [ ]:

```
# 计算中位生存时间
suv.quantile(0.5)
```

In [ ]:

```
# 计算中位生存时间的可信区间
suv.quantile_ci(0.5)
```

In [ ]:

```
suv.summary()
```

In [ ]:

```
suvplt = suv.plot()
```

In [ ]:

```
# 绘制分组生存曲线图
ax = plt.axes()

df1 = dfsuv.query('group == 1')
sm.SurvfuncRight(df1.month, df1.event).plot(ax)

df2 = dfsuv.query('group == 2')
sm.SurvfuncRight(df2.month, df2.event).plot(ax)

ax
```

12.2.2 生存曲线的比较

class statsmodels.duration.survdiff(

time : 时间变量
status : 生存结局变量，1代表事件发生，0代表截尾
group : 希望进行比较的分组变量

weight_type = 'fh' : 具体使用的检验方法
'fh' : Fleming-Harrington，检验中所有时间点等权重，即标准的log-rank检验
'gb' : Gehan-Breslow，按照该时间点暴露在风险中的个案数对时间点加权
'tw' : Tarone-Ware，按照该时间点暴露在风险中个案数的平方根对时间点加权

)

In [ ]:

```
kmtest = sm.duration.survdiff(time = dfsuv.month,
                              status = dfsuv.event, group = dfsuv.group)
kmtest
```

In [ ]:

```
sm.duration.survdiff(dfsuv.month, dfsuv.event, dfsuv.group,
                     weight_type = 'gb')
```

In [ ]:

```
sm.duration.survdiff(dfsuv.month, dfsuv.event, dfsuv.group,
                     weight_type = 'tw')
```

12.3 Cox比例风险模型

class statsmodels.duration.hazard_regression.PHReg(

```
    endog : 生存时间变量
    exog : 自变量矩阵
    status = None : 生存结局变量，1代表事件发生，0代表截尾
    entry = None
    strata = None : 分层变量
    offset = None : 模型偏移量
    ties = 'breslow' : 打结数据的处理方式，'breslow' or 'efron'
    missing = 'drop'
```

)

12.3.1 基本模型输出

In [ ]:

```
phres = sm.PHReg(dfsuv.month, dfsuv.group, dfsuv.event).fit()
```

In [ ]:

```
phres.summary()
```

In [ ]:

```
phres.baseline_cumulative_hazard
```

In [ ]:

```
phres.t_test("group = 0") # phres.wald_test("group = 0")
```

In [ ]:

```
phres.f_test("group = 0")
```

12.3.2 复杂分析实例

某研究者欲研究肺癌四种亚型的生存时间有无差别，收集了一些肺癌病例的数据，共有以下变量：

```
Sur_time：生存时间（单位：天）
Status：指示生存状态的变量。0，失访；1，死亡
Cell：癌细胞病理类型。1，腺癌；2，大细胞癌；3，小细胞癌；4，鳞癌
Health：病人入院时的身体健康指数，取值在0~100之间
Diagtime：从诊断为肺癌到开始治疗的时间间隔（月）
Age：病人的年龄
Sex：性别。1，男；2，女
```

试比较各种病理类型肺癌病人的生存曲线是否相同，数据见表单lung_ca。

In [ ]:

```
dflung = pd.read_excel('dmdata.xlsx', sheet_name = 'lung_ca')
dflung.head()
```

In [ ]:

```
from statsmodels.formula.api import phreg

phres = phreg("sur_time ~ C(type) + health + diagtime + age + sex",
                dflung, status = dflung.status, ties = "breslow").fit()
```

In [ ]:

```
phres.summary()
```

In [ ]:

```
phres.llf * 2
```

In [ ]:

```
phres2 = phreg("sur_time ~ health + diagtime + age + sex",
                dflung, status = dflung.status, ties = "breslow").fit()
phres2.llf * 2
```

In [ ]:

```
import scipy.stats as ss

1- ss.chi2.cdf(2 * (phres.llf - phres2.llf), 3)
```

12.4 生存分析中的分层因素

In [ ]:

```
phres3 = phreg("sur_time ~ health + diagtime + age + sex",
                dflung, status = dflung.status, strata = dflung.type,
              ties = "breslow").fit()
```

In [ ]:

```
phres3.summary()
```

In [ ]:

```
phres3.llf * 2
```

In [ ]:

```
phres3.baseline_cumulative_hazard
```

12.5 实战练习

试分别计算km表单数据分两个实验组的25%、75%生存概率对应的生存时间及其可信区间。将计算出的生存时间和生存分析表相对照，理解生存时间的含义。

使用似然比检验方法对肝癌研究数据进行变量的筛选，以得到最终模型。