法律声明

- 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

回归实践

邹博

主要内容

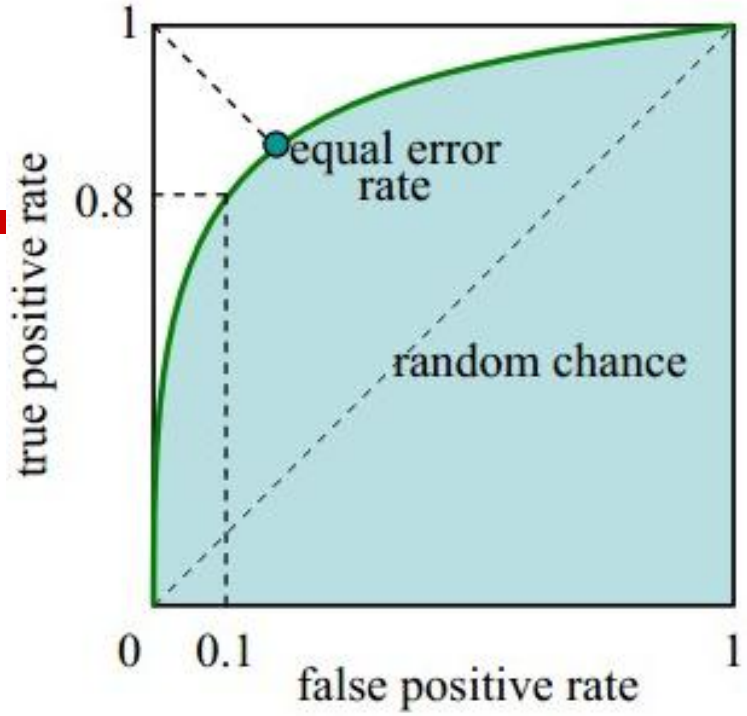
- AUC
  - 分类器指标
- 代码实践
  - 调参与交叉验证
- 该部分PPT中仅列举模型效果截图，详细内容请参考该PPT的配套代码。

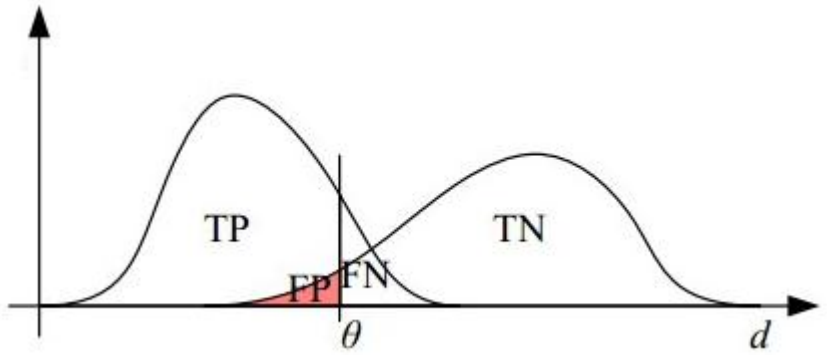
AUC

预测值 / 实际值	Positive	Negtive
正	TP	FN
负	FP	TN

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$



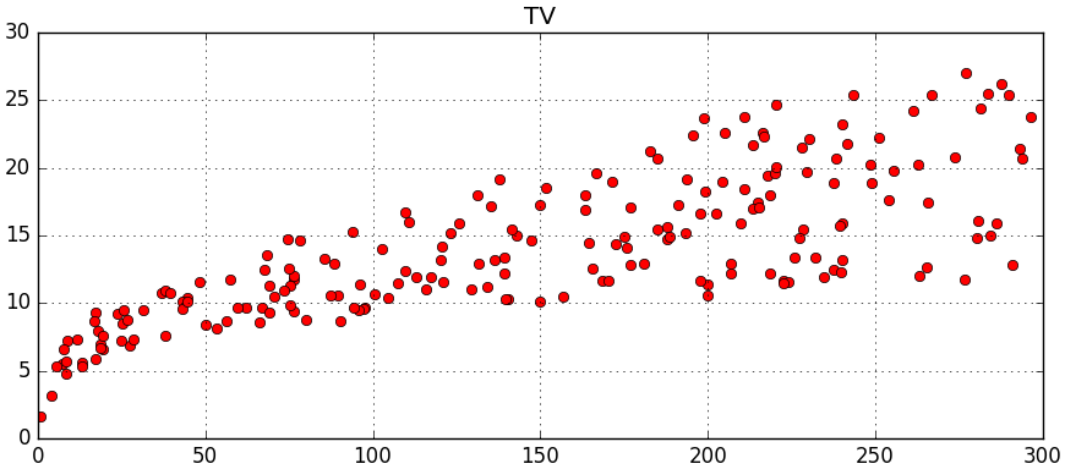


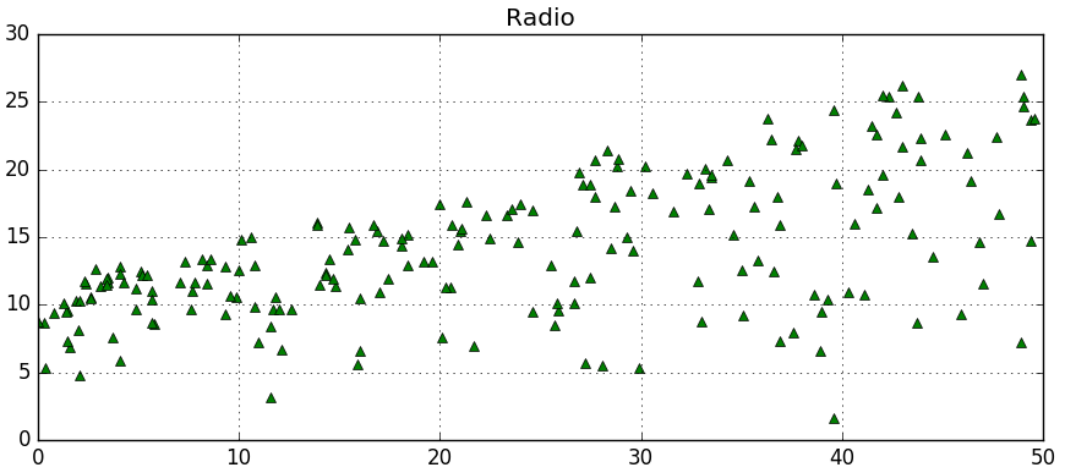
Receiver Operating Characteristic

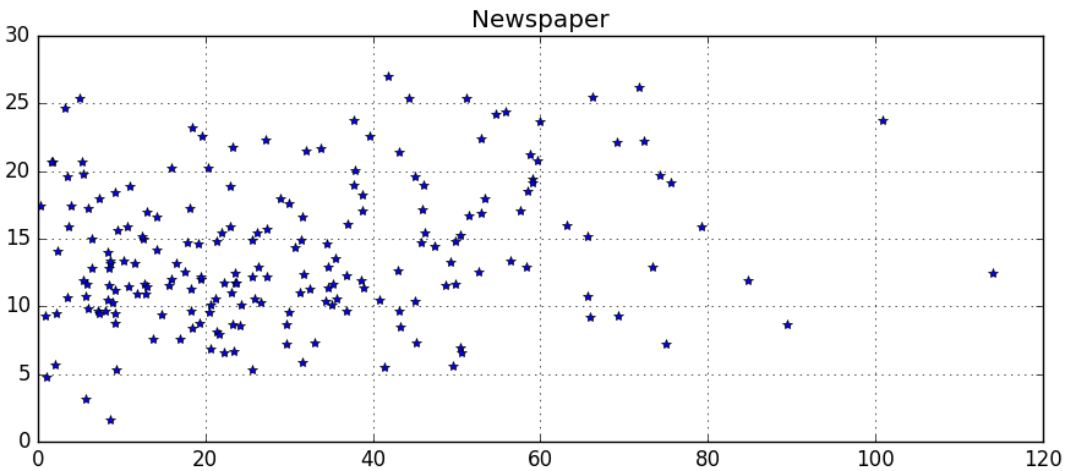
Area Under Curve

数据显示

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6
12	214.7	24	4	17.4
13	23.8	35.1	65.9	9.2
14	97.5	7.6	7.2	9.7
15	204.1	32.9	46	19
16	195.4	47.7	52.9	22.4
17	67.8	36.6	114	12.5
18	281.4	39.6	55.8	24.4
19	69.2	20.5	18.3	11.3
20	147.3	23.9	19.1	14.6
21	218.4	27.7	53.4	18
22	237.4	5.1	23.5	12.5
23	13.2	15.9	49.6	5.6
24	228.3	16.9	26.2	15.5
25	62.3	12.6	18.3	9.7
26	262.9	3.5	19.5	12
27	142.9	29.3	12.6	15
28	240.1	16.7	22.9	15.9
29	248.8	27.1	22.9	18.9
30	70.6	16	40.8	10.5
31	292.9	28.3	43.2	21.4
32	112.9	17.4	38.6	11.9
33	97.2	1.5	30	9.6
34	265.6	20	0.3	17.4
35	95.7	1.4	7.4	9.5
36	290.7	4.1	8.5	12.8
37	266.9	43.8	5	25.4
38	74.7	49.4	45.7	14.7
39	43.1	26.7	35.1	10.1
40	228	37.7	32	21.5

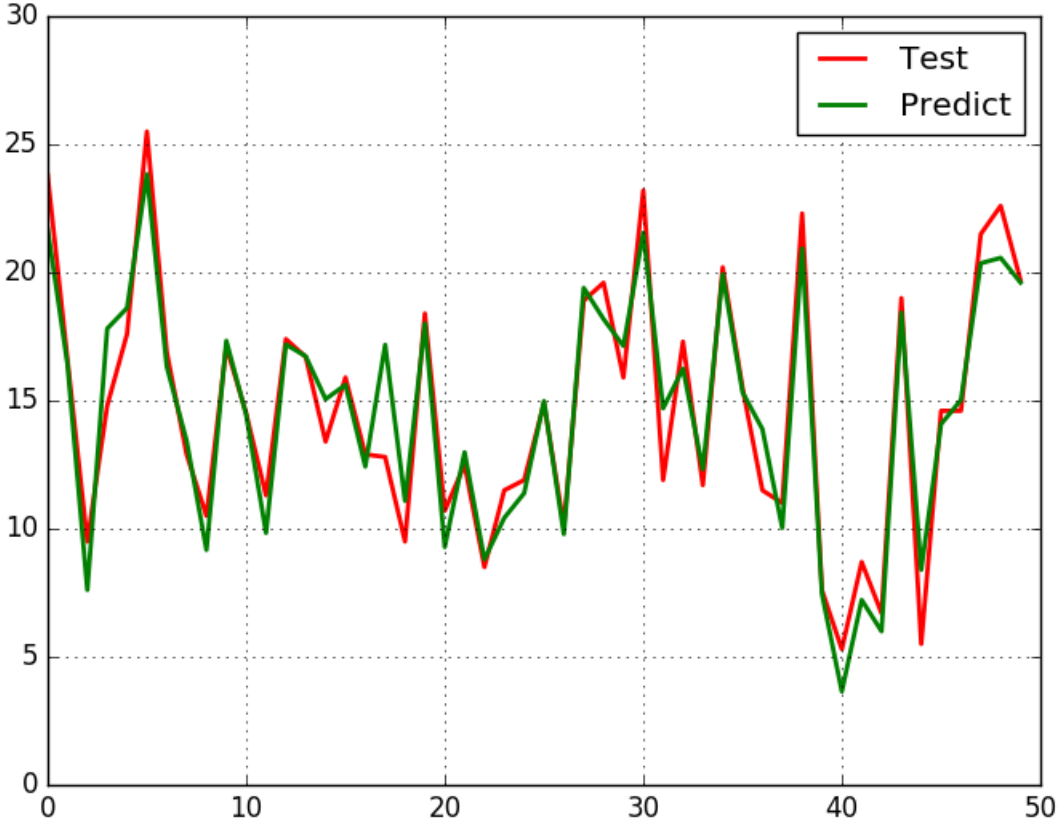






拟合与预测

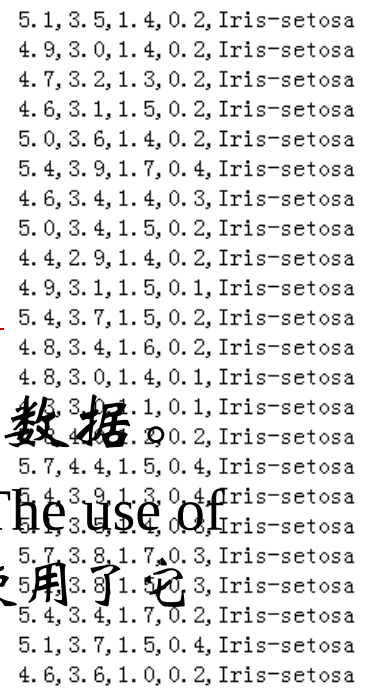
- y=2.877+0.046*TV+0.179*Radio+0.0035*Newspaper



小结

- 本模型虽然简单，但它涵盖了机器学习的相当部分的内容。
  - 使用75%的训练集和25%的测试集
  - 分析模型后，使用最为简单的方法：直接删除；得到了更好的预测结果。
- 奥卡姆剃刀
  - 如果能够用简单模型解决问题，则不使用更为复杂的模型。因为复杂模型往往增加不确定性，造成过多人力和物力成本，且容易过拟合。

鸢尾花数据集



- 鸢尾花数据集或许是最有名的模式识别测试数据。
  - 早在1936年，模式识别的先驱Fisher就在论文“The use of multiple measurements in taxonomic problems”中使用了它(直至今日该论文仍然被频繁引用)。
- 该数据集包括3个鸢尾花类别，每个类别有50个样本。其中一个类别是与另外两类线性可分的，而另外两类不能线性可分。
  - 由于Fisher的最原始数据集存在两个错误(35号和38号样本)，实验中我们使用的是修正过的数据。
- 下载链接：http://archive.ics.uci.edu/ml/datasets/Iris

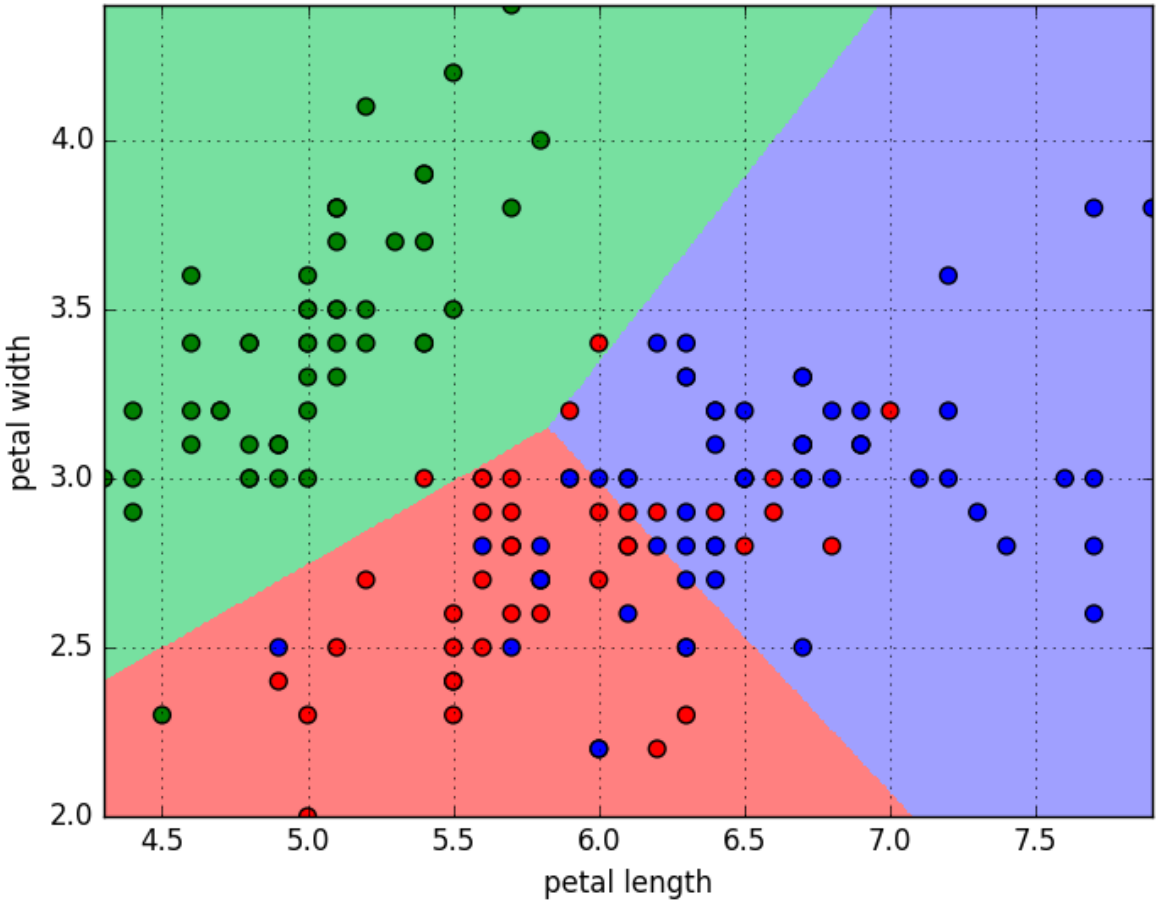
数据描述

- 该数据集共150行，每行1个样本。每个样本有5个字段，分别是
  - 花萼长度(单位cm)
  - 花萼宽度(单位：cm)
  - 花瓣长度(单位：cm)
  - 花瓣宽度(单位：cm)
  - 类别(共3类)
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

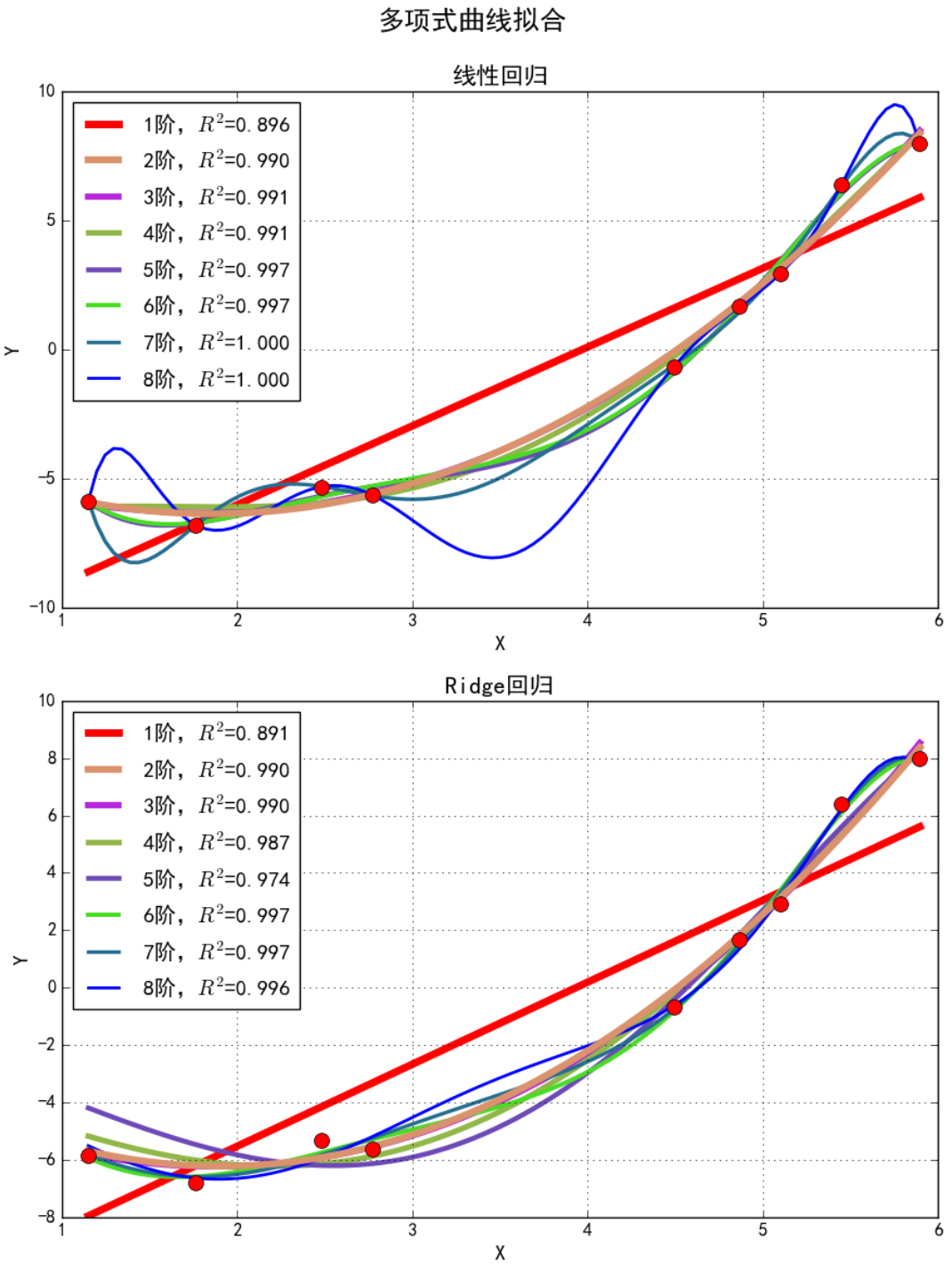
```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
```

鸢尾花数据集的分类

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
```



超参与过拟合


多项式曲线拟合
线性回归
1阶，R^2=0.896
2阶，R^2=0.990
3阶，R^2=0.991
4阶，R^2=0.991
5阶，R^2=0.997
6阶，R^2=0.997
7阶，R^2=1.000
8阶，R^2=1.000
Y
X
Ridge回归
1阶，R^2=0.891
2阶，R^2=0.990
3阶，R^2=0.990
4阶，R^2=0.987
5阶，R^2=0.974
6阶，R^2=0.997
7阶，R^2=0.997
8阶，R^2=0.996
Y
X

作业

- 推导Softmax回归的梯度公式。
- 参考给出的Logistic回归或线性回归代码，使用其他数据集做分类或预测实验。

我们在这里

- http://wenda.ChinaHadoop.cn
  - 视频/课程/社区
- 微博
  - @ChinaHadoop
  - @邹博_机器学习
- 微信公众号
  - 小象
  - 大数据分析挖掘

感谢大家！

恳请大家批评指正！