内部讨论的部分总结和延伸

Some Thoughts from the Previous Internal Seminar

王斐, Michael

SenseTime Edu, wangfei1@sensetime.com

Math, Economics, Philosophy (UCD, Nottingham, CUHK)

https://github.com/Michael-yunfei/MDLforBeginners

2020 年 7 月 26 日

本节内容

1. 缘起
2. 初中教材从线性回归引入到分类的疑难
3. 线性分类的切入方法再思考
4. 详解梯度下降法和其动画演示

故事的开端

这个视频录制的想法是从上周内训后的讨论中酝酿的

- 初中教材从线性分类引入线性回归 - 逻辑存在一定问题
- 内部讨论的很多问题并非 '三言两语' 就可以说明白:
  - 想要展开把这个问题说透
  - 暂时回不去就录个视频来说明
  - 即使回去了，也不可能因为这个问题召集大家开个会
  - 个人认为教育团队通过视频的方式进行研讨还是很高效的
- 问题永远比答案更重要 (从与祁老师的讨论中得出的感想):
  - 线性分类如何切入？
  - 为什么要用梯度下降法 (Gradient Descent)
  - 答案全在网上，问题却在心中, 需要得是 '时代在召唤'
  - 磨课:'如切如磋，如琢如磨'!

问题回顾

我这边没有初中教材，所以我凭记忆把当时我们讨论的问题再回顾下。坦白讲，初中和高中的教材在严谨度上和表达存在一定问题, 通过这个视频也特别解释下:

- 当我在 ‘谈论我们初高中的教材问题时，我在谈论些什么’
- 线性回归的基本公式为

$$\begin{aligned} y &= X\beta + \varepsilon \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} \quad i = 1, 2, \cdots m \\ \text{结果} &= \text{参数} \cdot \text{属性} \end{aligned}$$

- 可视化: 表格 + 图形

问题回顾: 数据阵

我们再过一遍下面这张表格

数据阵(dataframe)

属性(n)=7　　　　标注结果

样本 m=15

样本	交易日期	房屋年龄	地铁距离(米)	便利店数量	维度	经度	房屋价格(万台币/坪)	房屋价格(万元/平方米)
1	2012.917	32	84.87882	10	24.98298	121.54024	37.9	2.720
2	2012.917	19.5	306.5947	9	24.98034	121.53951	42.2	3.028
3	2013.583	13.3	561.9845	5	24.98746	121.54391	47.3	3.394
4	2013.500	13.3	561.9845	5	24.98746	121.54391	54.8	3.932
5	2012.833	5	390.5684	5	24.97937	121.54245	43.1	3.093
6	2012.667	7.1	2175.03	3	24.96305	121.51254	32.1	2.303
7	2012.667	34.5	623.4731	7	24.97933	121.53642	40.3	2.892
8	2013.417	20.3	287.6025	6	24.98042	121.54228	46.7	[illegible]
9	2013.500	31.7	5512.038	1	24.95095	121.48458	18.8	[illegible]
10	2013.417	17.9	1783.18	3	24.96731	121.51486	22.1	[illegible]
11	2013.083	34.8	405.2134	1	24.97349	121.53372	41.4	2.971
12	2013.333	6.3	90.45606	9	24.97433	121.5431	58.1	4.169
13	2012.917	13	492.2313	5	24.96515	121.53737	39.3	2.820
14	2012.667	20.4	2469.645	4	24.96108	121.51046	23.8	1.708
15	2013.500	13.2	1164.838	4	24.99156	121.53406	34.3	2.461

X　　　　Y

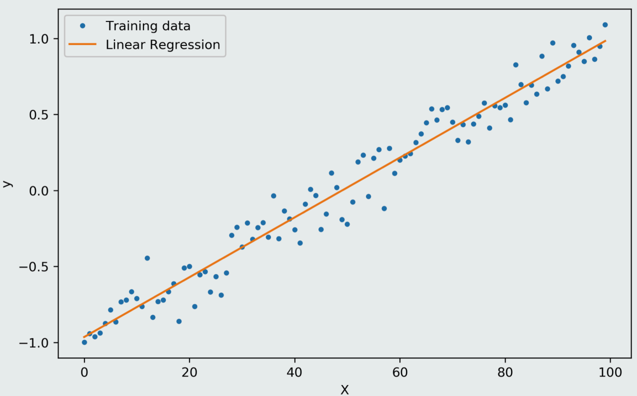
X = m x n = 15 x 7 的矩阵　　　　**Y=15 x 1 向量**

$$\text{房屋价格} = \beta_0 + \beta_1\text{房屋年龄} + \beta_2\text{地铁数量} + \cdots + \beta_7\text{经度}$$

问题回顾: 线性回归详解

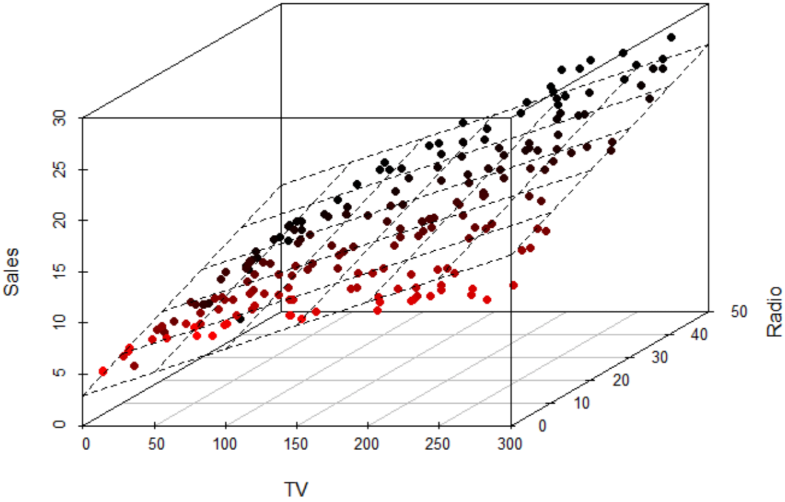
线性回归:

- 等式左边: 结果 y, 假设连续
- 等式右边: 属性 $\beta(w)$ 和参数变量 $X(m \times n)$, m 个样本，n 个属性
- 可视化: 一个结果轴 和多个属性轴(n)
- 可视化: 最高到三维



问题回顾: 线性回归详解

三维图片: 一个结果 (Y) 轴，两个属性轴 (x_1, x_2)



从线性回归引入到分类

从线性回归切入到分类有两个角度:

- 继续沿用线性回归思路:
  - 逻辑回归 (Logistic Regression)
  - 线性判别分析 (Linear Discriminant Analysis, LDA)
- 抛弃线性回归思路，通过向量点积引入 (试讲时会特别讲解)

教材的问题有:

- 沿用了线性回归思路，但是并没有介绍逻辑回归 (logistic regression) 或者线性判别分析 (LDA)
- 想要介绍分界面，但是没有把背后的数学点积问题讲清楚

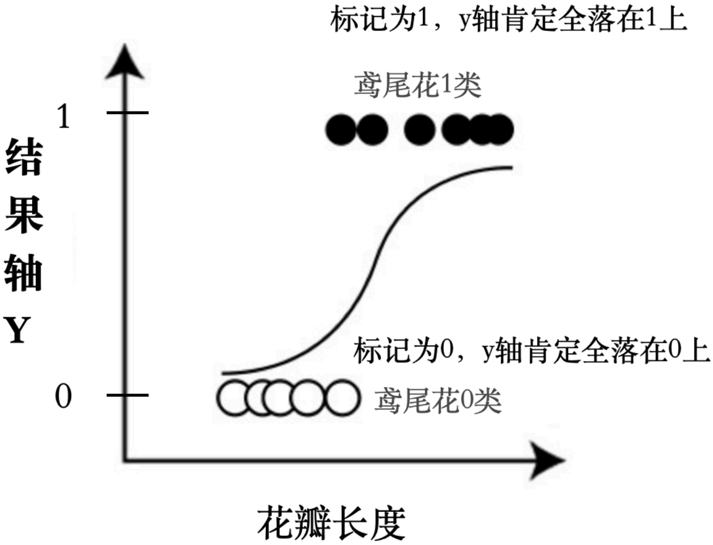
什么意思？请看下图！

从线性回归引入到分类

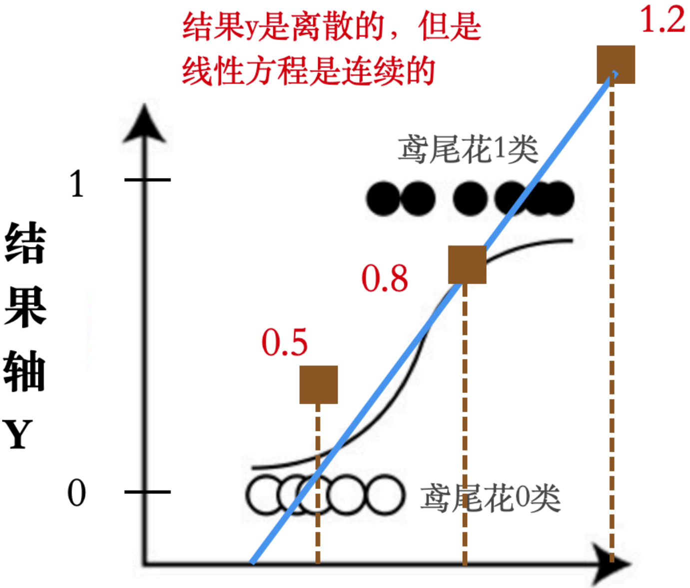
再看一下表格

属种	结果 (y)	花瓣长度	花瓣宽度
山鸢尾	0	1.4	0.2
山鸢尾	0	1.7	0.4
变色鸢尾	1	3.9	1.4
变色鸢尾	1	4.9	1.5
维吉尼亚鸢尾	2	6.9	2.3
维吉尼亚鸢尾	2	6.1	1.9

从线性回归引入到分类



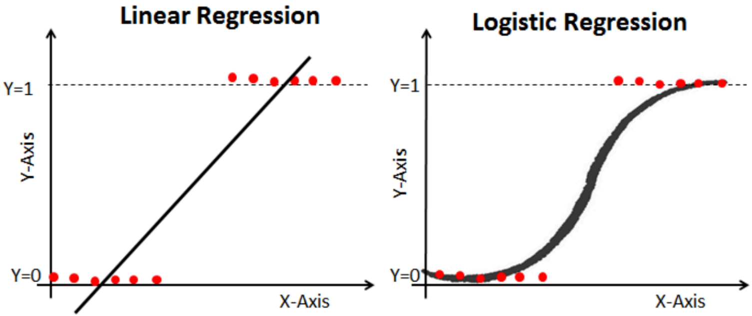
从线性回归引入到分类



从线性回归引入到分类

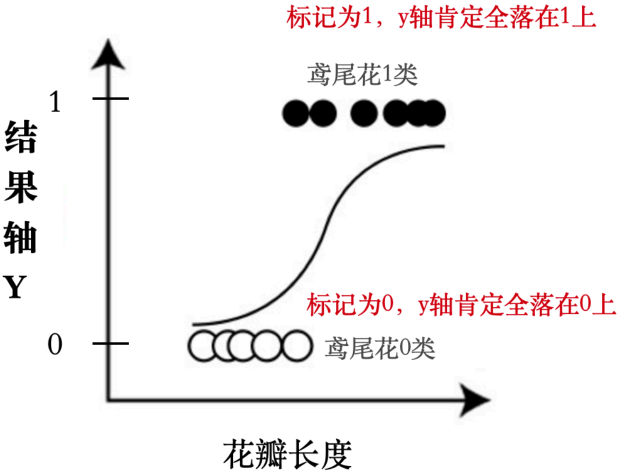
引入逻辑回归 (logistic regression), sigmoid function:

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{X\beta+\varepsilon}}$$



从线性回归引入到分类

我就不明白，怎么教材上分类的时候，Y 轴上会出现一条斜线 呢？

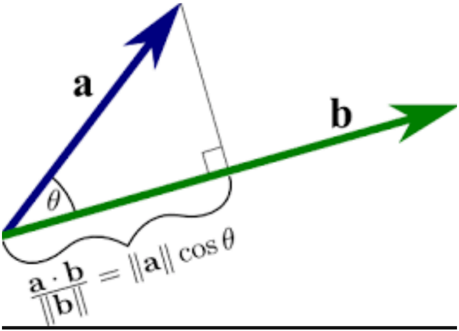


线性分类的切入

线性分类的切入:

- 延续线性回归思路，需要转换 Y 轴，借助 logistic regression(sigmoid function)
- 不延续线性回顾思路，完全代数性质 (点积)

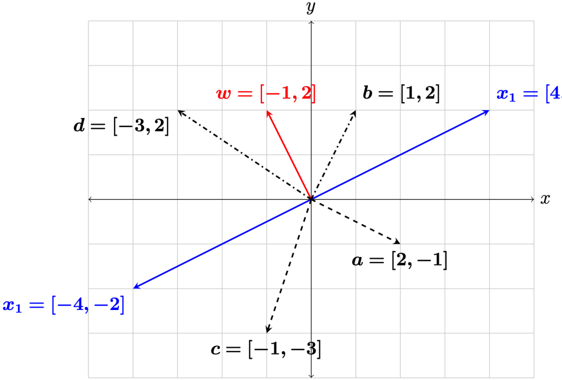
$$a \cdot b = ||a|| ||b|| \cos\theta$$



线性分类的切入

Vladimir Vapnik 还活着，所以可以了解到最初他是如何创立支持向量机 (Support vector machine) 的。

与 w 点积	<0	$=0$	>0
	a	x_1	b
	c	x_2	d



Reference: Patrick Winston (Youtube, 点击率 1 百万)

高中教材

高中教材 36 页，下列公式的错误:

$$\gamma = \min \gamma^i \not\Longrightarrow \min \frac{2}{\gamma}$$

应该为

$$\gamma = \max \gamma^i \Rightarrow \min \frac{2}{\gamma}$$

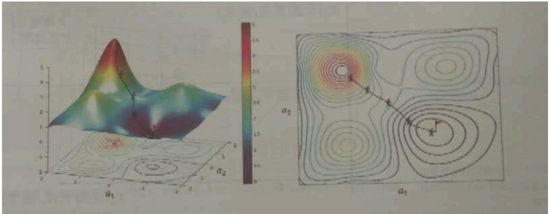


图 2-12 优化过程示意图

梯度下降法: 为什么?

天不生仲尼 (牛顿)，亘古如长夜；求极限值 (最大值，最小值)

- 笨办法，遍历 $[1, 5, 8, 9]$;
- 设置方程，求导 (一阶条件，二阶条件);
- 但是计算机不会求导 (需要人求导后，告诉它公式)
- 而且计算机不会解方程:
  - 极限值位置，遍历？
  - 有规律的找到极限值, 随机找一个点，然后迭代逼近优化值

最小化损失函数 (损失越小，预测越准确):

$$L = \min_{w,b} f(w, b)$$

我们关心什么？

- <u>是什么 w 值</u>，让 L 最小

梯度下降法原理

机器学习中，梯度下降法 (Gradient Descent) 的公式可能是最为常见的公式之一了:

$$w_{t+1} = w_t - \alpha f'(x)$$

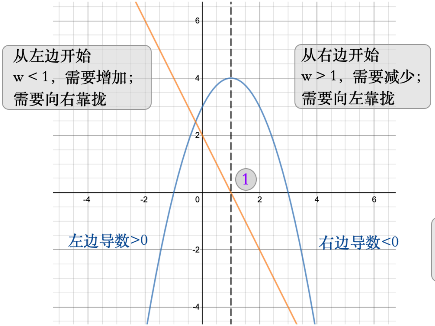
- α learning rate(更新参数/速率)
- $f'(x)$ 为损失函数 $f(w, b)$ 的导数
- 算法：随机取一个 w，然后逼近优化值。

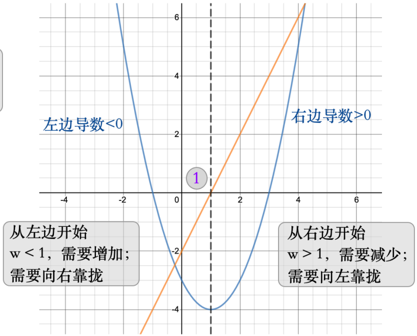
我们就通过下面的例子和动画来理解下梯度下降法的原理。虽然具体的损失函数可能不同，但是原理是一致的，我们用一个最简单的例子来说明:

$$f(x) = x^2 - 2x - 3$$
$$f'(x) = 2x - 2$$

梯度下降法原理

当方程处于极限值时，导数等于 0: 极限值左右两边，导数正负值相反。





$w_{t+1} = w_t + \boxed{\alpha f'(w)}$

正值

w 增加，向右靠拢

$w_{t+1} = w_t + \boxed{\alpha f'(w)}$

负值

w 减少，向左靠拢

$w_{t+1} = w_t - \boxed{\alpha f'(w)}$

负值

w 增加，向右靠拢

$w_{t+1} = w_t - \boxed{\alpha f'(w)}$

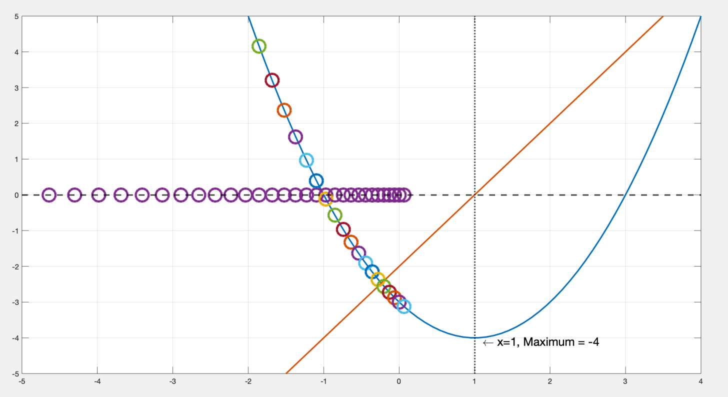
正值

w 减少，向左靠拢

梯度下降法原理

$$w_{t+1} = w_t - \alpha f'(w)$$

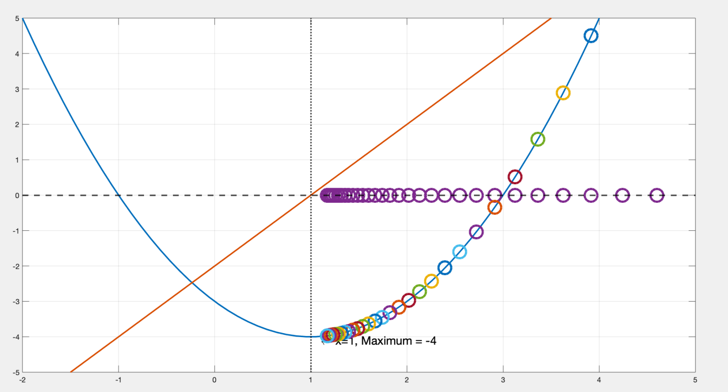
w 增加，向右靠拢



梯度下降法原理

$$w_{t+1} = w_t - \alpha f'(w)$$

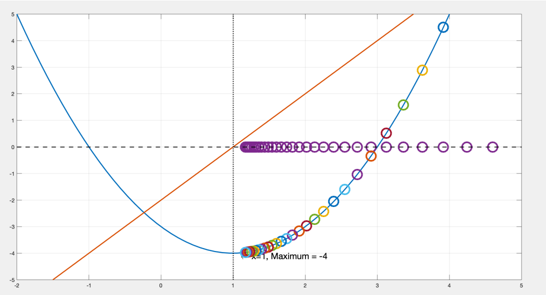
w 减少，向左靠拢



梯度下降法原理: 动画

下面我们看一下梯度下降法的动画！请注意:

- 计算机是看不到我们看到的图形的
- 计算机只认识公式:$w_{t+1} = w_t - \alpha f'(w)$
- 计算机也搞不懂你为什么让它计算这些东西，更不清楚为什么要迭代 30 次或者 50 次，也不明白 α 选大选小的影响
- 两个大脑论



结语

因为平时大家都很忙，各自有自己的项目和工作，有些想法不好集中沟通的，我觉得大家都可以做成视频:

- 超越时空来共享想法
- 有助于系统梳理思路
- 而且非常有助于提供反馈
- 答案全在网上，问题却在心中, 需要得是 ‘时代在召唤’
- 问题比答案更重要!
- How many roads must a man walk down before you call him a man?
- The answer, my friend, is blowing in the wind