Logistic与SVM分类

一分为二

预备知识

课前

- Coursera吴恩达《机器学习》WEEK 7

https://www.coursera.org/learn/machine-learning

- 《数据挖掘导论》5.5节

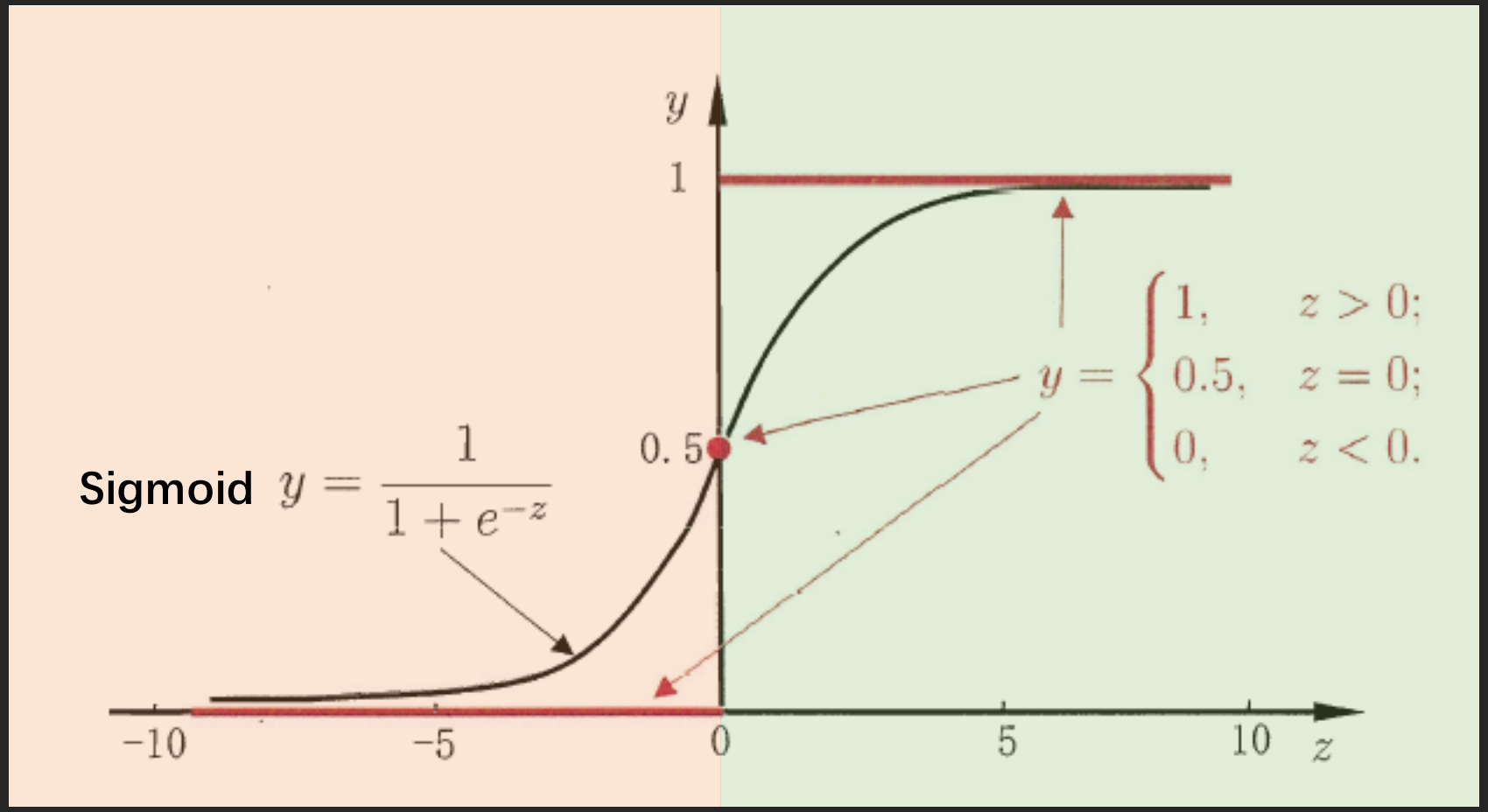
课后

- 学堂在线《数据挖掘：理论与算法》WEEK 5

http://www.xuetangx.com/courses/course-v1:TsinghuaX+80240372X+2016_T2/about

从线性拟合到Logistic分类

$$z = \theta \cdot x$$



Logistic分类的本质

基于统计的分类

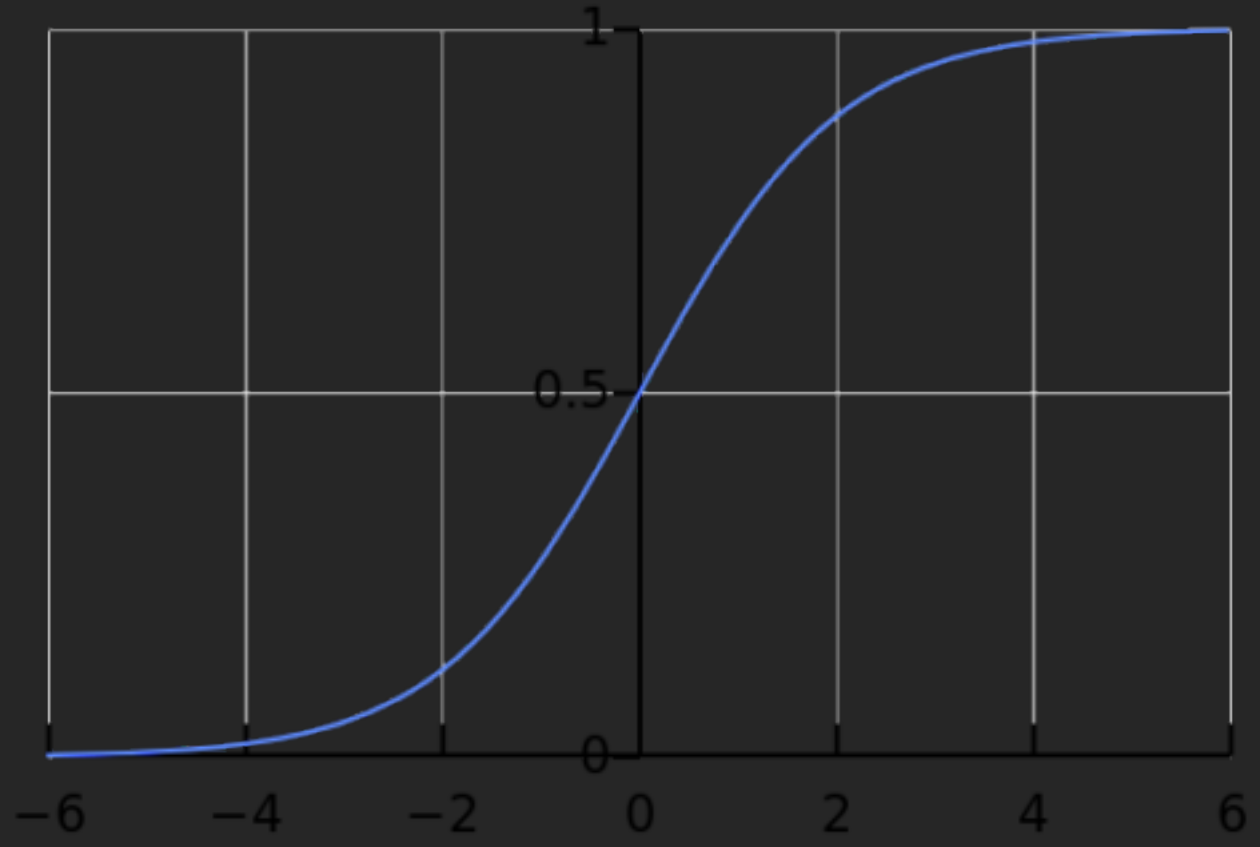
- 结果是一个概率

比如：属于1的概率

以0.5作为阈值

- 单调可导函数

避免无法求导的尴尬



数学模型

Cross Entropy成本函数：

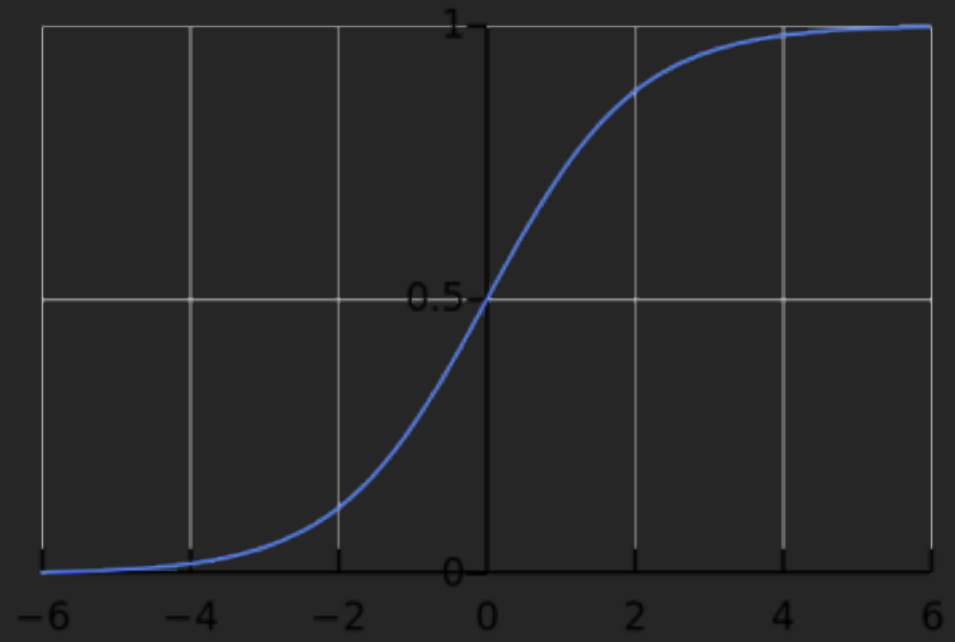
$$J(\theta) = -\sum \delta(y_i = 1) * \log[g(\theta \cdot x_i)] + \delta(y_i = 0) * \{1 - \log[g(\theta \cdot x_i)]\}$$

$g(\cdot)$ – Logistic函数

目标：

$$\min \{J(\theta)\}$$

为什么成本函数与线性拟合不同？



Sigmoid函数的**两侧几乎是平的**

→ 最小二乘函数的**梯度**在大部分情况下**很小**

→ **收敛地很慢**

支持向量机分类的本质

基于几何的分类

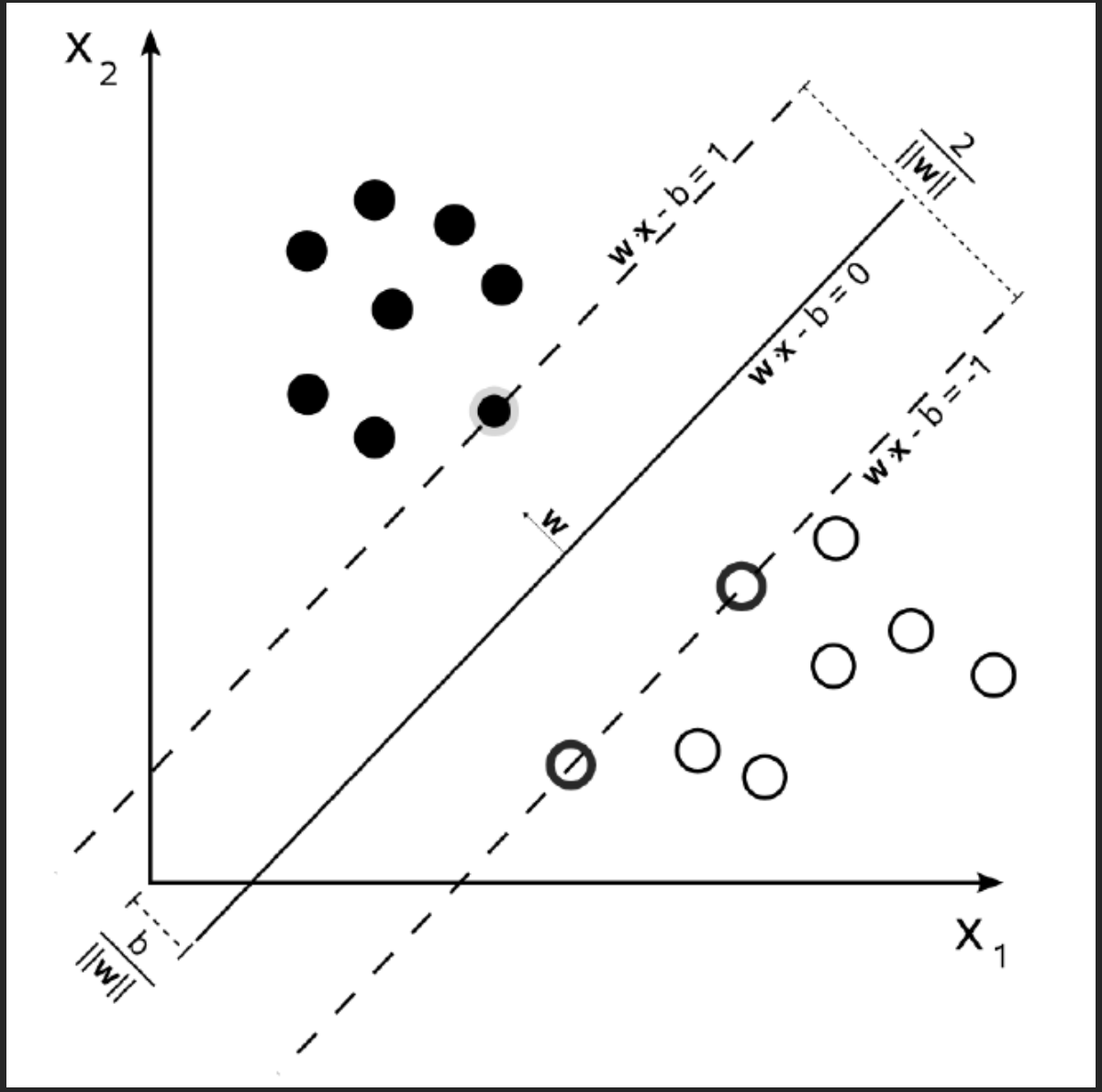
- 最大化分类平面与两个类型之间的距离

优化的过程 – 最大化分类间距

- 模型的最优解只与边界上的数据点有关

边界上的数据点 – 支持向量

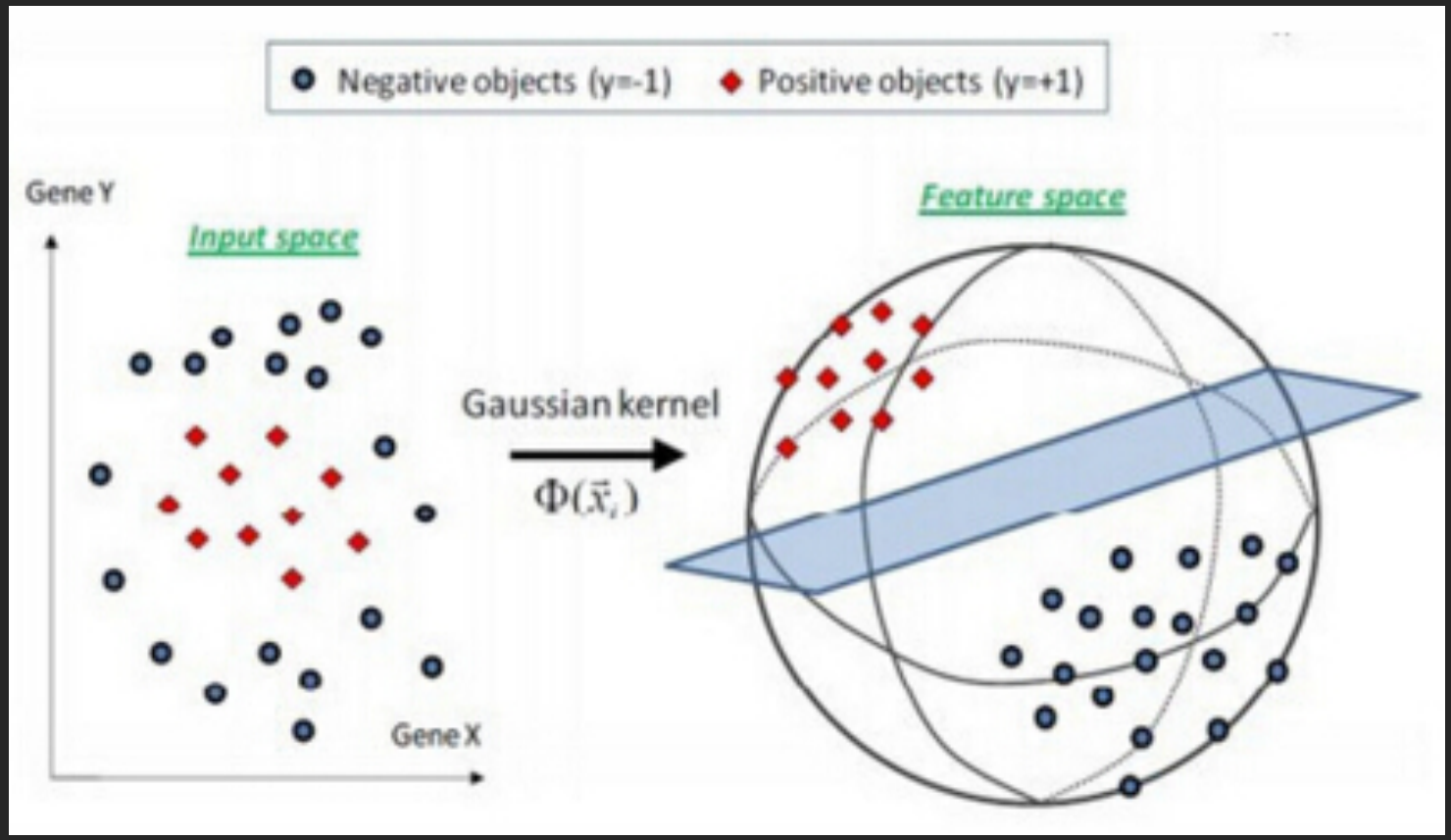
分类



支持向量机的理论推导

略

核函数的本质



核函数是一种特征提取的手段；核函数不仅仅用于支持向量机

样本不均衡问题

- 构造更多的样本，使得样本数均衡

减少多数类别的样本数；增加少数类别的样本数

- 改变分类器的边界

比如：Logistic – 0.5 → m^+/m^-

- 直接研究少数样本的统计特性

即从研究p(y | x)到研究p(x)

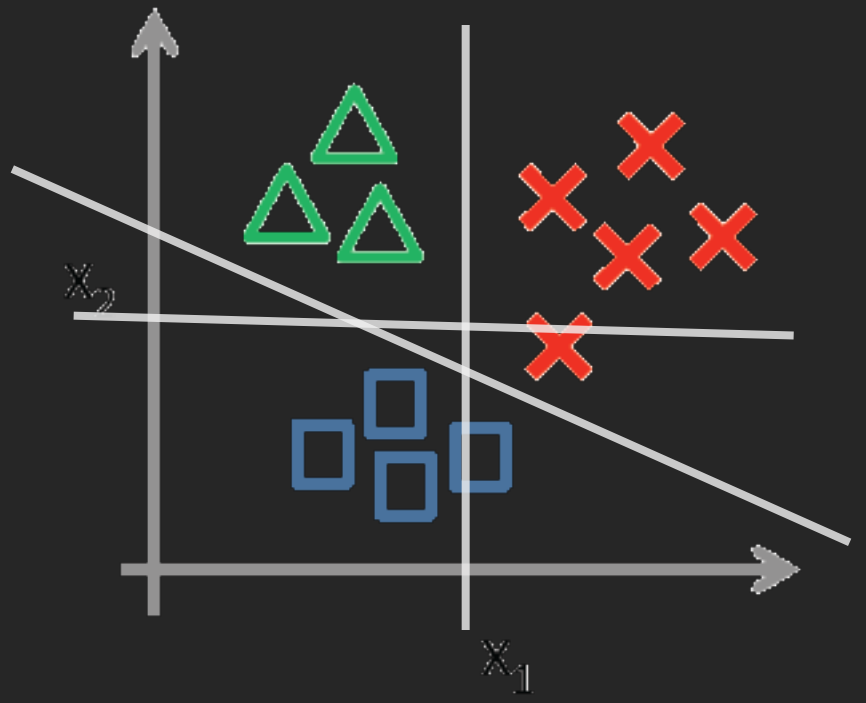
多分类问题

- 通用方法：one vs. all

构造k(k-1)/2个二分类器

- 构造多分类的成本函数

比如：Softmax分类器



Softmax分类器

Cross Entropy成本函数：

$$J(\theta) = -\sum_i \sum_c \delta(y_i = c) * \log[p(y = c \mid \theta, x_i)]$$

目标：

$$\min \{J(\theta)\}$$

类别之间明显互斥 - Softmax

类别之间有交叉 - one vs. all

Next Class – 分类问题（真枪实弹）

课前

- 学习Python的基本语法

https://docs.python.org/2/tutorial/index.html

- 了解Python的Scikit-Learn和TensorFlow机器学习框架

http://scikit-learn.org/

https://www.tensorflow.org/

Next Next Class – 神经网络

课前

- Coursera吴恩达《机器学习》WEEK 4、5

https://www.coursera.org/learn/machine-learning

- 《数据挖掘导论》5.4节

课后

- 学堂在线《数据挖掘：理论与算法》WEEK 4

http://www.xuetangx.com/courses/course-v1:TsinghuaX+80240372X+2016_T2/about