线性模型

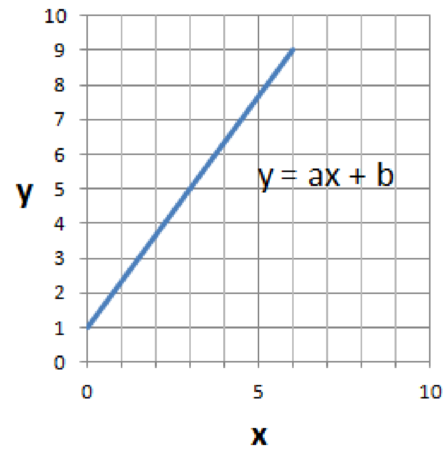
主讲人：屠恩美

《机器学习与知识发现》

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

线性关系与线性模型

- 线性(近似线性)关系是生活中最常见的关系，包括正比与反比关系
  - 物体运动路程和运动速度的关系
  - 物体的重量与其密度间的关系
  - 商品的售价与数量
  - GPA与学习的时间（近似）
  - ……



- 线性模型是研究的**最久远**、**最透彻**也是应用范围**最广泛**的一种模型
  - 概念、算法简单，易于理解和实现
  - 物理含义直观明确，可解释性好
  - 是许多复杂非线性模型的基础

线性代数回顾

- 向量 $\mathbf{x}=\left(x_1,x_2,\ldots,x_d\right),\mathbf{y}=\left(y_1,y_2,\ldots,y_d\right)$
  - 1-范数: $\|\mathbf{x}\|_1=\sum_{i=1}^{d}|x_i|$ ， 2-范数: $\|\mathbf{x}\|_2=\sqrt{\sum_{i=1}^{d}x_i^2}$
  - 欧氏距离： $d(\mathbf{x},\mathbf{y})=\sqrt{\sum_{i=1}^{d}(x_i-y_i)^2}=\|\mathbf{x}-\mathbf{y}\|_2$
  - 内积： $\mathbf{x}^T\mathbf{y}=\sum_{i=1}^{d}x_iy_i$
- 矩阵 $\mathbf{A}=\left(\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_d\right)$ ，则 $\mathbf{Ax}=\sum_{i=1}^{d}x_i\mathbf{a}_i$, $\mathbf{x}^T\mathbf{Ay}=\sum_{i=1}^{d}\sum_{j=d}^{d}x_iy_ja_{ij}$
- 迹 $\mathrm{tr}\mathbf{A}=\sum_{i=1}^{d}a_{ii}$ ， 特征分解 $\mathbf{Au}=\lambda\mathbf{u}$

推荐参考： http://www.cs.cmu.edu/~zkolter/course/15-884/linalg-review.pdf

第一部分

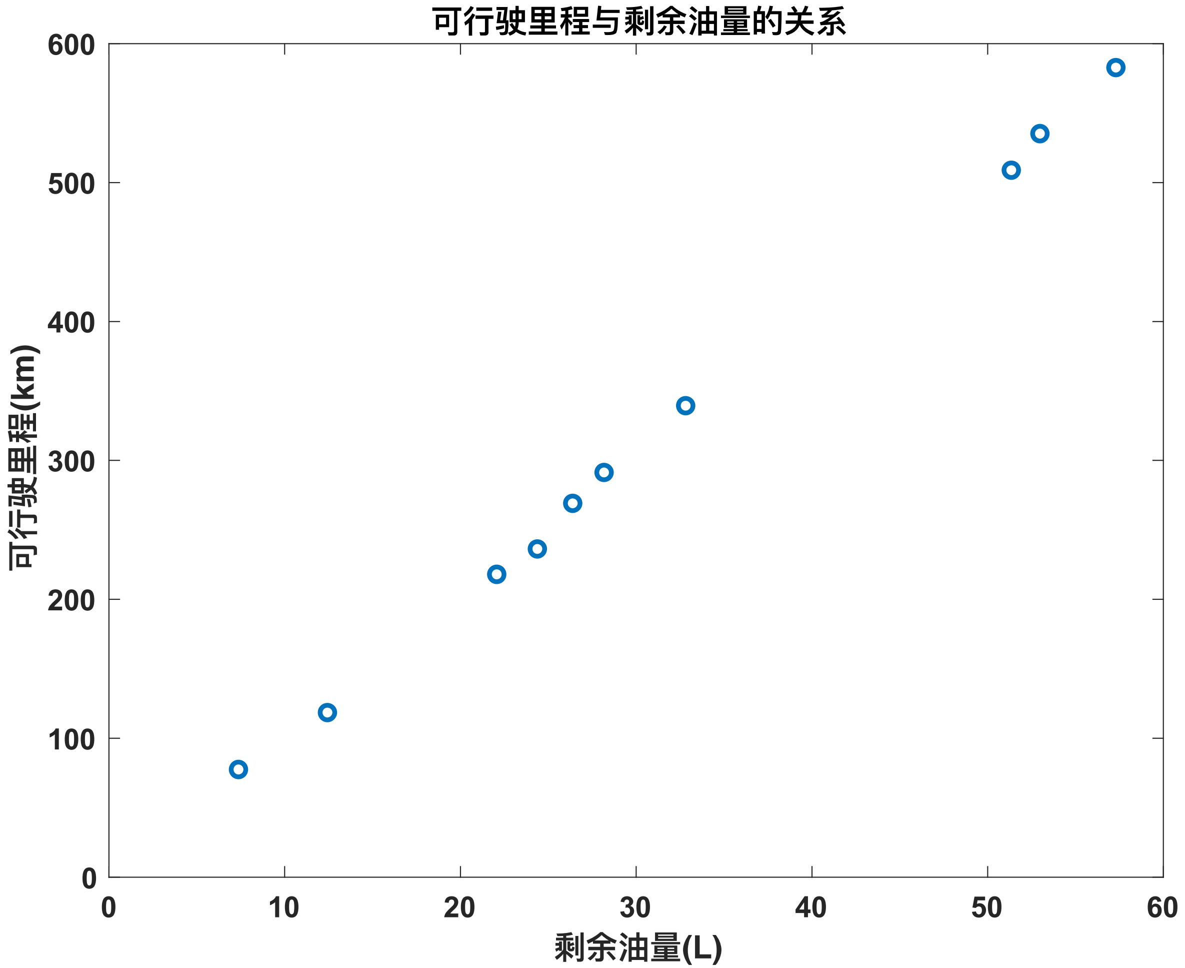
线性回归模型

线性关系例子

- 开车往往需要知道剩余油量还能开多远
- 最好能够根据任意剩余油量预测还能行驶多少公里。

线性关系例子

剩余油量L	可行驶距离km
24.6	235.5
53.6	534.6
32.8	338.4
22.2	217.2
12.8	117.2
26.6	268.6
57.8	581.8
7.6	76.9
28.8	290.6
51.8	508.5



代数的角度

从代数的角度解决问题:

- Step 1: 假设方程关系 $y = ax + b$ ，其中a, b是未知数，待求解
- Step 2: 测量两组剩余油量-可行驶里程之间对应关系的数据

剩余油量	可行驶里程
32	318
7.4	73

- Step 3: 带入模型求解未知数

$$\begin{cases} 73 = 7.4a + b \\ 318 = 32a + b \end{cases} \Rightarrow \begin{cases} a = 9.95 \\ b = -0.69 \end{cases}$$

机器学习角度

如何从机器学习的角度去求解?

- Step 1: 假设**模型**关系 $f(x)=wx+b$，其中w, b 是待**学习参数**
- Step 2: 测量一组剩余油量-可行驶里程对应数值

$$\left(x_1,y_1\right),\left(x_2,y_2\right),\ldots,\left(x_n,y_n\right)$$

- Step 3: 均方误差损失函数 (因为不存在某个w 满足所有测量，因此尽可能接近)

$$J(w,b)=\sum_{i=1}^{n}\left(f(x_i)-y_i\right)^2$$

- Step 4: 最小化目标函数，求得最优参数值

$$\left(w^*,b^*\right)=\arg\min J(w,b)$$

即求解最优的$\left(w^*,b^*\right)$使得$J(w^*,b^*)=\min\sum_{i=1}^{n}\left(f(x_i)-y_i\right)^2$ 如何求解?

模型求解

- 目标函数

$$J(w^*,b^*)=\min\sum_{i=1}^{n}\left(f(x_i)-y_i\right)^2=\min\sum_{i=1}^{n}\left(wx_i+b-y_i\right)^2$$

- 回忆高数中函数极小值点的必要条件：**一阶导数等于0**

$$\begin{cases}\dfrac{\partial J}{\partial w}=0\\[2ex]\dfrac{\partial J}{\partial b}=0\end{cases}\longrightarrow\quad(w^*,b^*)$$

- 计算一阶导数

$$\begin{cases}\dfrac{\partial J}{\partial w}=2\sum\limits_{i=1}^{n}\left(wx_i+b-y_i\right)x_i\\[2ex]\dfrac{\partial J}{\partial b}=2\sum\limits_{i=1}^{n}\left(wx_i+b-y_i\right)\end{cases}$$

模型求解

- 令一阶导数等于0

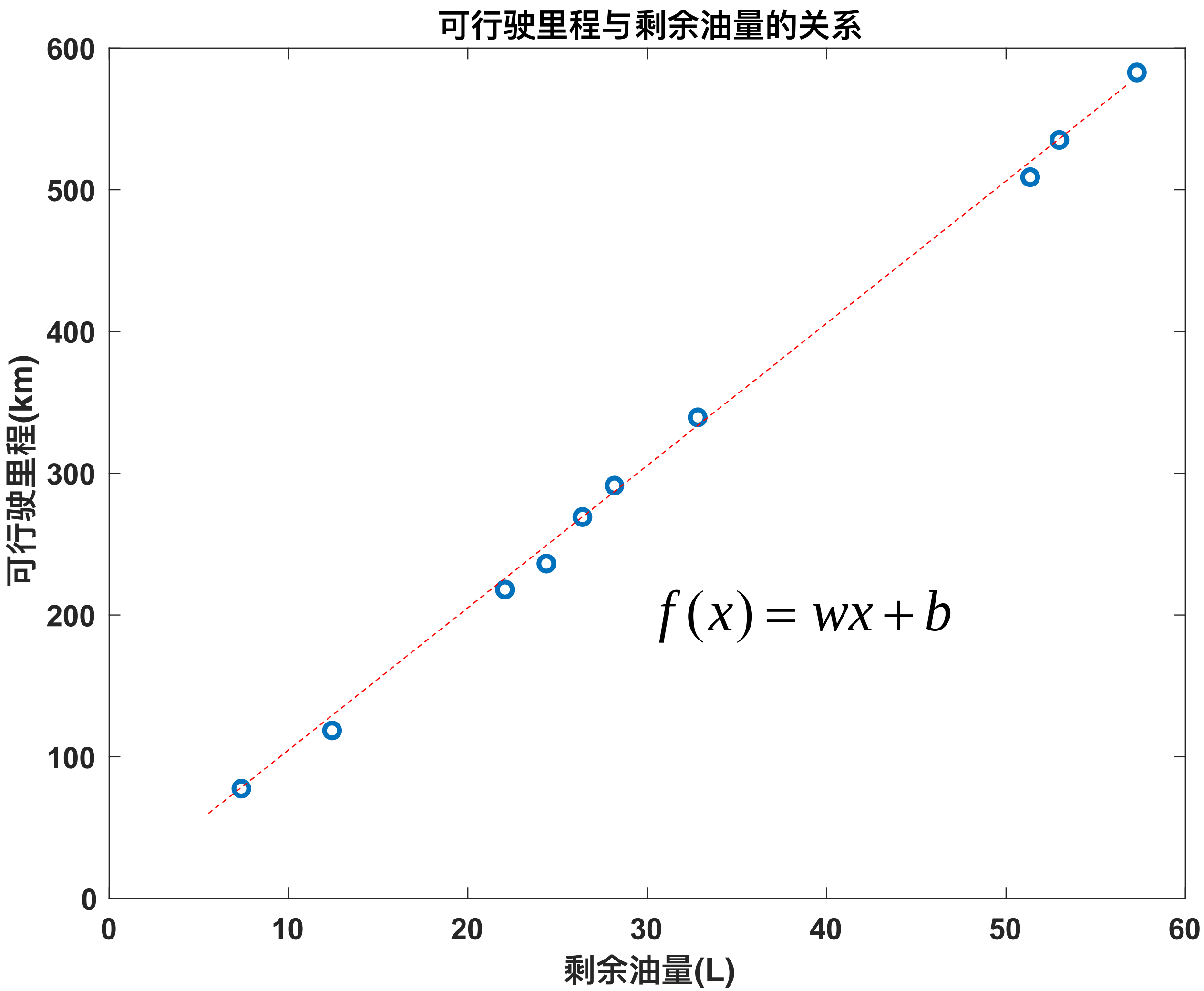
$$\begin{cases} 2\sum_{i=1}^{n}\left(wx_i+b-y_i\right)x_i=0 \\ 2\sum_{i=1}^{n}\left(wx_i+b-y_i\right)=0 \end{cases}$$

- 可算出模型参数

$$\begin{cases} w^*=\dfrac{\sum_{i=1}^{n}y_i\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2} \\ b^*=\dfrac{1}{n}\sum_{i=1}^{n}\left(y_i-wx_i\right) \end{cases}$$

$$(\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i)$$

模型结果


可行驶里程与剩余油量的关系
可行驶里程(km)
剩余油量(L)
0
100
200
300
400
500
600
10
20
30
40
50
60
$f(x) = wx + b$

最小二乘法

- 基于**均方误差**最小化的模型求解方法称为最小二乘法(Least Square Method, LSM)

$$\begin{cases} J(\theta)=\sum_{i=1}^{n} e_i^2 & \text{目标函数} \\ e_i = y_i - f(x_i \mid \theta), \quad i=1..n & \text{误差项} \end{cases}$$

- 通常情况：方程个数大于模型参数个数；含义：最小化所有误差之和
- 求解：参数梯度

$$\frac{\partial J}{\partial \theta} = 2\sum_{i=1}^{n} e_i \frac{\partial e_i}{\partial \theta} = -2\sum_{i=1}^{n} e_i \frac{\partial f(x_i \mid \theta)}{\partial \theta}$$

- $f(x \mid \theta)$ 是关于 θ 的线性函数，可直接解析求解最优参数 θ
- $f(x \mid \theta)$ 是关于 θ 的非线性函数，采用分步迭代算法。每步先采用Taylor展式局部线性逼近 $f(x \mid \theta)$，再解析求解

多元线性回归问题

- 现实应用中往往是**多个因素**主导，会更复杂
  - 剩余油量
  - 车子载重
  - 车子年限
  - ……
- 因此，输入往往是**多变量**

  输入变量（自变量）： $\mathbf{x}$ = (剩余油量，车子载重，车子年限，……)

  输出变量（因变量）： y = 可行驶里程

- 如何构建模型，根据多个输入预测可行驶的里程?

多元线性回归模型

- 构建模型：输出的目标值（标签）是样本属性的线性组合

$$f(\mathbf{x}) = \underbrace{\mathbf{w}}_{\text{权值向量}}{}^T \mathbf{x} + \underbrace{b}_{\text{偏移量}}$$

此时式中 $\mathbf{w} \in \mathbb{R}^d, \mathbf{x} \in \mathbb{R}^d, b \in \mathbb{R}$

- 为了表达的简洁和公式推导的方便，通常

$$f(\hat{\mathbf{x}}) = \hat{\mathbf{w}}^T \hat{\mathbf{x}}$$

式中 $\hat{\mathbf{w}} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$， $\hat{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$

多元线性回归模型

- 如果把样本数据采用矩阵的形式记为

$$\mathbf{X} = \begin{bmatrix} \hat{\mathbf{x}}_1^T \\ \hat{\mathbf{x}}_2^T \\ \vdots \\ \hat{\mathbf{x}}_n^T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T, & 1 \\ \mathbf{x}_2^T, & 1 \\ \vdots & \\ \mathbf{x}_n^T, & 1 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{f} = \begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix}$$

则一组测量数据方程组可整体写为

$$\begin{cases} f(\hat{\mathbf{x}}_1) = \hat{\mathbf{w}}^T \hat{\mathbf{x}}_1 \\ f(\hat{\mathbf{x}}_2) = \hat{\mathbf{w}}^T \hat{\mathbf{x}}_2 \\ \vdots \\ f(\hat{\mathbf{x}}_n) = \hat{\mathbf{w}}^T \hat{\mathbf{x}}_n \end{cases} \longrightarrow \quad \mathbf{f} = \mathbf{X}\hat{\mathbf{w}}$$

多元线性回归模型

类似一元回归情况，最小二乘法求解：

- 构建**均方误差**损失函数

$$J(\hat{\mathbf{w}}) = \sum_{i=1}^{n}\left(f(\mathbf{x}_i) - y_i\right)^2 = (\mathbf{f} - \mathbf{y})^T(\mathbf{f} - \mathbf{y})$$

$$= (\mathbf{X}\hat{\mathbf{w}} - \mathbf{y})^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y})$$

- 对参数求一阶导数并等于0

$$\frac{\partial J}{\partial \hat{\mathbf{w}}} = 2\mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) = 0 \quad \longrightarrow \quad \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}} - \mathbf{X}^T\mathbf{y} = 0$$

- 根据 $\mathbf{X}$ 的不同情况（方程个数与变量个数大小关系），解也有所不同

多元线性回归模型

- 如果 $\mathbf{X}$ 是列满秩（方程个数>变量个数），则 $\mathbf{X}^T\mathbf{X}$ 可逆，此时有唯一解

$$\hat{\mathbf{w}}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- 如果 $\mathbf{X}$ 非列满秩（方程个数<变量个数），则 $\mathbf{X}^T\mathbf{X}$ 不可逆，此时有（无穷）多个解（*）

$$\hat{\mathbf{w}}^* = \mathbf{A}^{\dagger}\mathbf{X}^T\mathbf{y} + \left(\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A}\right)\mathbf{u}\ , \quad \mathbf{A} = \mathbf{X}^T\mathbf{X},\ \mathbf{u} \text{ is a free vector}$$

- 求得 $\hat{\mathbf{w}}^*$ 后，可以利用模型预测任何给定 $\mathbf{x}$ 对应的函数值

$$f(\hat{\mathbf{x}}) = \left(\hat{\mathbf{w}}^*\right)^T\hat{\mathbf{x}} = \left(\mathbf{w}^*\right)^T\mathbf{x} + b^*$$

多元回归例子

测量次数	剩余油量(L)	车子载重(kg)	行驶里程(km)
1	50	75	500
2	50	225	440
3	50	150	460
4	10	75	100
5	10	150	80
6	10	225	60

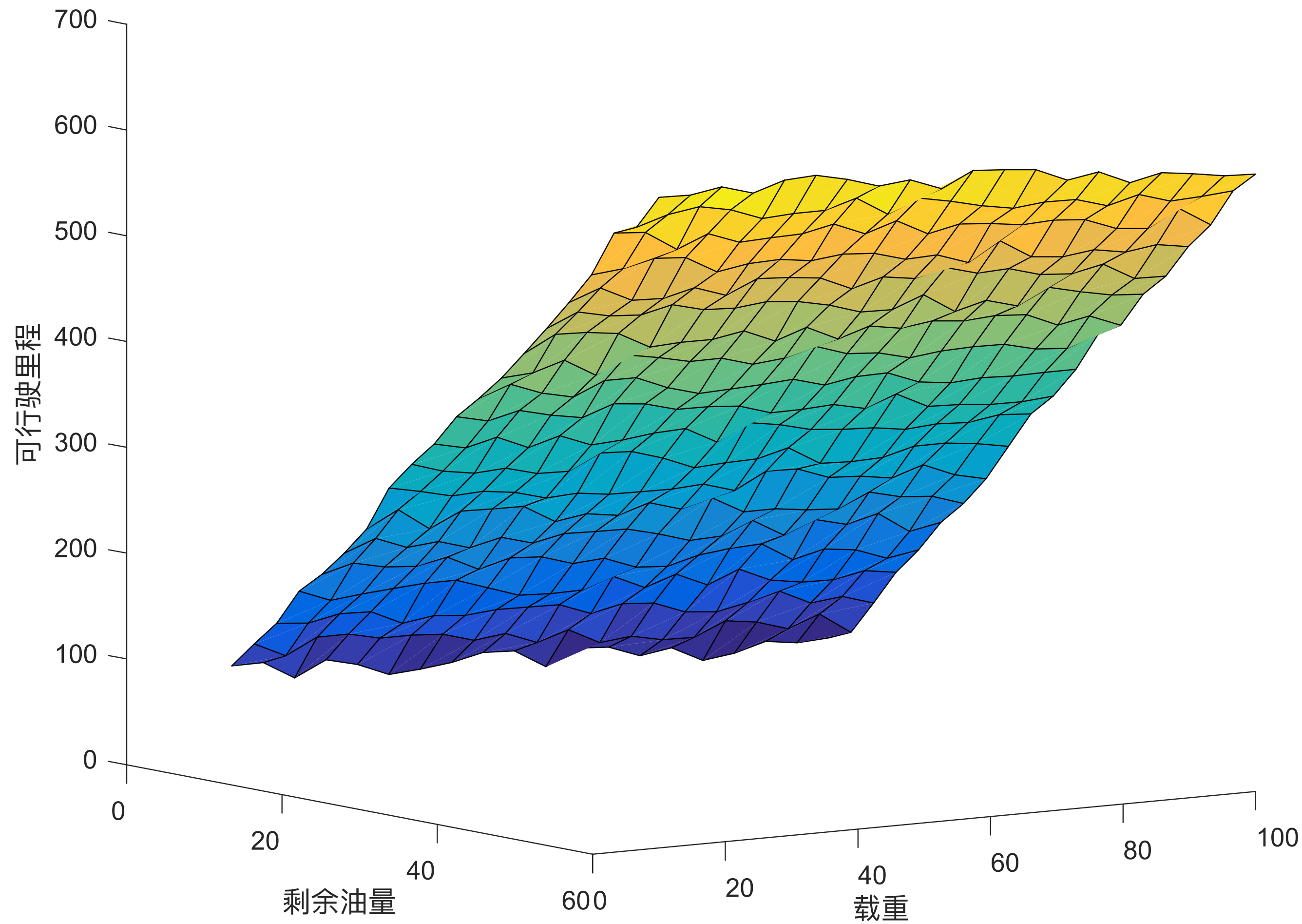
$$\begin{cases} \mathbf{w}^* = \begin{bmatrix} 9.66 \\ -0.33 \end{bmatrix} \\ b^* = 33.3 \end{cases}$$

思考: 1) 为什么$\mathbf{w}^*$的第二个元素是负的?

2) $\mathbf{w}^*$元素的(绝对值)大小关系有什么含义?

2) b的值表示什么意义?

多元回归模型


可行驶里程
700
600
500
400
300
200
100
0
0
20
40
60
剩余油量
0
20
40
60
80
100
载重

岭回归（Ridge Regression）

- Q: X 是行满秩矩阵时的多个解，最终该选那个?
- 领回归：在线性回归基础上引入正则化项

$$J(\hat{\mathbf{w}}) = \left(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\right)^T \left(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\right) + \lambda \left\|\mathbf{w}\right\|^2$$

式中 λ >0 正则化系数

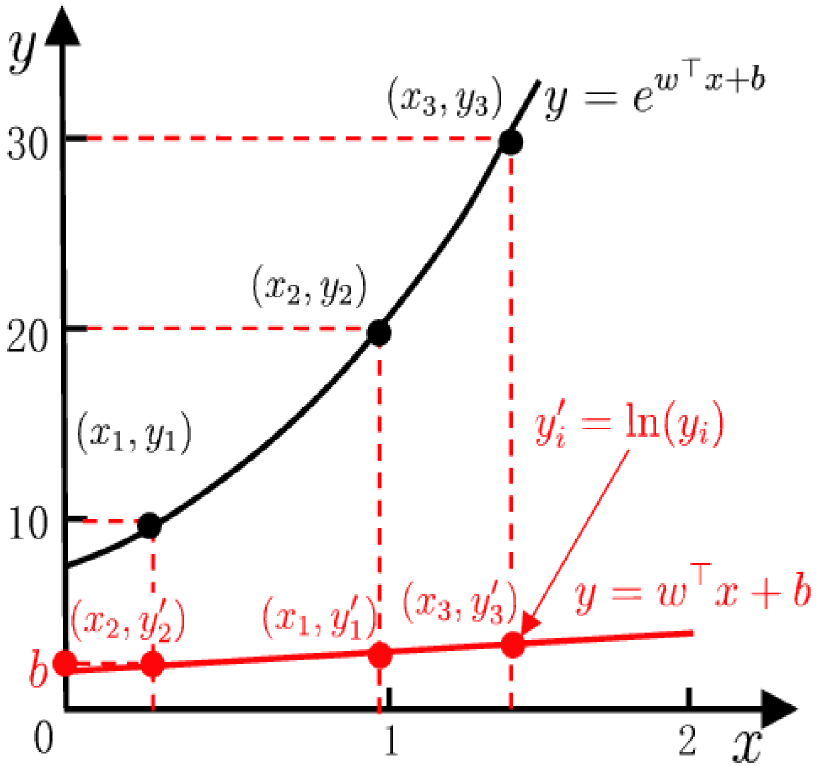
$$\frac{\partial J}{\partial \hat{\mathbf{w}}} = 2\mathbf{X}^T\left(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\right) + 2\lambda\hat{\mathbf{w}} = 0$$

$$\hat{\mathbf{w}}^* = \underbrace{\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}}_{\text{总是可逆，防止模型退化}} \mathbf{X}^T\mathbf{y}$$

线性模型拓展

- 线性模型稍加改动就可以处理非线性拟合，例如

$$y = e^{\mathbf{w}^T\mathbf{x}+b}$$



$$\ln y = \mathbf{w}^T\mathbf{x} + b$$

Log变换

$$\tilde{y} = \mathbf{w}^T\mathbf{x} + b$$

广义线性模型

- 更一般地，对于任意单调可逆函数 g

$$y = g(\mathbf{w}^T \mathbf{x} + b)$$

- 只要令

$$\tilde{y} = g^{-1}(y) = \mathbf{w}^T \mathbf{x} + b$$

就可以使用标准的线性拟合算法（最小二乘或最大似然）进行求解。

- 这类模型统称为广义线性模型 (GLM, generalized linear model)

第二部分

线性分类模型

二分类问题

- 二分类是生活中最常见的分类问题:
  - 是或者不是，知道或者不知道，有或者没有
  - 明天下雨 还不下雨
  - 机器发生故障还是没有故障
  - 患了某种疾病还是没有患病
  - ……
- 通常把两类分别叫负类和正类，用0-1表示，因此二分类也叫0-1分类

样本特征: x: x_1, x_2, x_4, ..., x_n

样本标签: y: 1, 0, 0, ..., 1

线性分类器

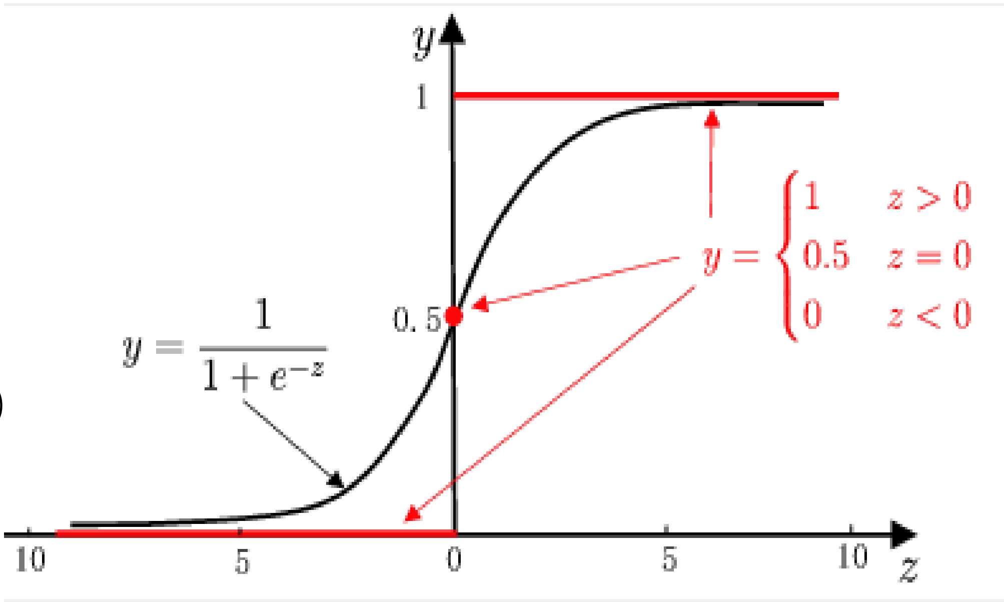
- 如何把线性拟合 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$ 的输出转换为0-1分类?
  - 阶跃函数

$$y = \begin{cases} 0, & z < 0; \\ 0.5, & z = 0; \\ 1, & z > 0, \end{cases}$$

问题：不连续(不能做GLM中g)

  - Sigmoid函数

$$y = \frac{1}{1 + e^{-z}}$$



单调可微、任意阶可导； $y \in [0,1]$ 可以看作是属于正类的**概率**

对数几率回归分类器

- 考虑到 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$

$$y = \frac{1}{1+e^{-z}} \Rightarrow y = \frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}$$

- 利用前面讲到的线性模型拓展技巧，可转换为线性模型

$$\ln\frac{y}{1-y} = \mathbf{w}^T\mathbf{x} + b$$

- 如何求模型参数$(\mathbf{w}, b)$？给定如下标记样本，无法代入上式用最小二乘法求解（Q: 为何？），需要其他求解方法

$$\begin{array}{llllll} \mathbf{x}: & \mathbf{x}_1, & \mathbf{x}_2, & \mathbf{x}_4, & \ldots, & \mathbf{x}_n \\ y: & 1, & 0, & 0, & \ldots, & 1 \end{array}$$

对数几率回归分类器

- 如前所述，y 是正类的概率，那么负类的概率是多少？因为

$$\begin{cases} y = \dfrac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}} = \dfrac{e^{\mathbf{w}^T\mathbf{x}+b}}{1+e^{\mathbf{w}^T\mathbf{x}+b}} \\ \text{正类概率}+\text{负类概率}=1 \end{cases}$$

- 所以

$$\text{负类概率}=\ 1-y = \frac{1}{1+e^{\mathbf{w}^T\mathbf{x}+b}}$$

- 记

$$\begin{cases} p(y=1\,|\,\mathbf{x}) = \dfrac{e^{\mathbf{w}^T\mathbf{x}+b}}{1+e^{\mathbf{w}^T\mathbf{x}+b}} & \text{正类概率} \\ p(y=0\,|\,\mathbf{x}) = \dfrac{1}{1+e^{\mathbf{w}^T\mathbf{x}+b}} & \text{负类概率} \end{cases}$$

对数几率回归分类器

- 构建对数似然函数

$$l(\mathbf{w},b)=\sum_{i=1}^{n}\ln p(y_i=j\,|\,\mathbf{x}_i,\mathbf{w},b);\quad j\in\{0,1\}$$

- 因为 $p(y_i=j\,|\,\mathbf{x}_i,\mathbf{w},b)=\underbrace{y_i p(y_i=1\,|\,\mathbf{x}_i,\mathbf{w},b)}_{\text{positive class}}+\underbrace{(1-y_i)p(y_i=0\,|\,\mathbf{x}_i,\mathbf{w},b)}_{\text{negtive class}}$

- 所以 $l(w,b)=\sum_{i=1}^{n}\ln\left(y_i\frac{e^{\mathbf{w}^T\mathbf{x}_i+b}}{1+e^{\mathbf{w}^T\mathbf{x}_i+b}}+(1-y_i)\frac{1}{1+e^{\mathbf{w}^T\mathbf{x}_i+b}}\right)$

$$=\sum_{i=1}^{n}\ln\left(\frac{y_ie^{\mathbf{w}^T\mathbf{x}_i+b}+1-y_i}{1+e^{\mathbf{w}^T\mathbf{x}_i+b}}\right)=\sum_{i=1}^{n}\left[\underbrace{\ln\left(y_ie^{\mathbf{w}^T\mathbf{x}_i+b}+1-y_i\right)}_{\text{分别考虑}y_i=0\text{和}y_i=1\text{情况}}-\ln\left(1+e^{\mathbf{w}^T\mathbf{x}_i+b}\right)\right]$$

$$=\sum_{i=1}^{n}\left[y_i\left(\mathbf{w}^T\mathbf{x}_i+b\right)-\ln\left(1+e^{\mathbf{w}^T\mathbf{x}_i+b}\right)\right]$$

对数几率回归分类器

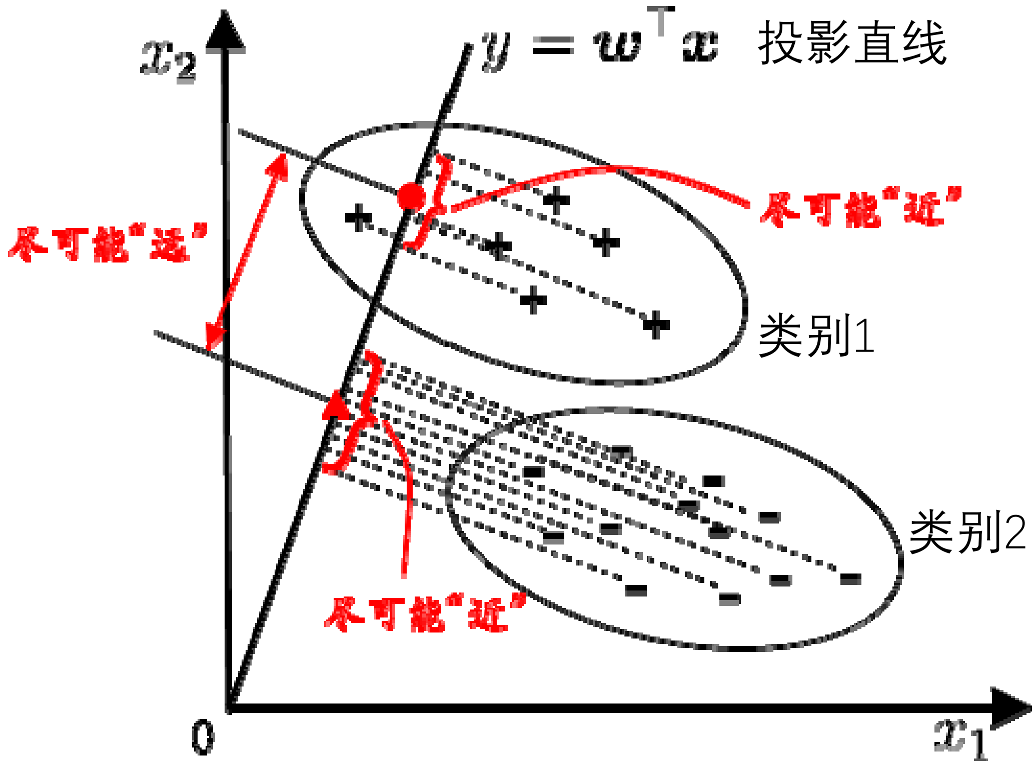
- 最大化 $l(\mathbf{w},b)$ 等价于最小化 $-l(\mathbf{w},b)$
- 是关于 $(\mathbf{w},b)$ 的凸函数，可以使用梯度下降法、牛顿法等优化算法求解
- 如梯度下降

$$\begin{cases} \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda\Delta\mathbf{w} = \mathbf{w}^{(t)} - \lambda \left.\dfrac{\partial l(\mathbf{w},b)}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}^{(t)},b=b^{(t)}} \\ b^{(t+1)} = b^{(t)} - \lambda\Delta b = b^{(t)} - \lambda \left.\dfrac{\partial l(\mathbf{w},b)}{\partial b}\right|_{\mathbf{w}=\mathbf{w}^{(t)},b=b^{(t)}} \end{cases}$$

式中
$$\begin{cases} \dfrac{\partial l(\mathbf{w},b)}{\partial \mathbf{w}} = -\sum_{i=1}^{n}[\mathbf{x}_i y_i - \mathbf{x}_i p(y_i = 1 \mid \mathbf{x}_i, \mathbf{w}, b)] \\ \dfrac{\partial l(\mathbf{w},b)}{\partial b} = -\sum_{i=1}^{n}[y_i - p(y_i = 1 \mid \mathbf{x}_i, \mathbf{w}, b)] \end{cases}$$

线性判别分析

- (Linear Discriminant Analysis, LDA) 是另一种经典的线性分类器，也是一种经典的监督降维算法
- 核心思想：寻找投影到直线，使**同类尽可能近**，**异类尽可能远**



线性判别分析

- 数据集 $D=\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, y\in\{0,1\}$，二分类问题
- 投影前，每类的均值和协方差矩阵

$$\begin{cases} \mathbf{u}_0 = \dfrac{1}{n_0}\sum\limits_{y_i=0}\mathbf{x}_i\ , & \boldsymbol{\Sigma}_0 = \dfrac{1}{n_0-1}\sum\limits_{y_i=0}\mathbf{x}_i\mathbf{x}_i^T \\ \mathbf{u}_1 = \dfrac{1}{n_1}\sum\limits_{y_i=1}\mathbf{x}_i\ , & \boldsymbol{\Sigma}_1 = \dfrac{1}{n_1-1}\sum\limits_{y_i=1}\mathbf{x}_i\mathbf{x}_i^T \end{cases}$$

- 投影后，每类的均值和协方差（投影到**直线**，因此均值协方差都是实数）

$$\begin{cases} \hat{u}_0 = \mathbf{w}^T\mathbf{u}_0\ , & \hat{\Sigma}_0 = \mathbf{w}^T\boldsymbol{\Sigma}_0\mathbf{w} \\ \hat{u}_1 = \mathbf{w}^T\mathbf{u}_1\ , & \hat{\Sigma}_1 = \mathbf{w}^T\boldsymbol{\Sigma}_1\mathbf{w} \end{cases}$$

线性判别分析

- 同类点尽可能近，要求 $\hat{\Sigma}_0+\hat{\Sigma}_1$ 尽可能小（每类的方差小，说明分布集中）
- 异类点尽可能远，要求 $\left|\hat{u}_0-\hat{u}_1\right|$ 尽可能大（类中心距离大，说明分布较远）
- 两者同时考虑，则最大化目标函数如下

$$J=\frac{\left\|\hat{u}_0-\hat{u}_1\right\|^2}{\hat{\Sigma}_0+\hat{\Sigma}_1}=\frac{\left\|\mathbf{w}^T\mathbf{u}_0-\mathbf{w}^T\mathbf{u}_1\right\|^2}{\mathbf{w}^T\boldsymbol{\Sigma}_0\mathbf{w}+\mathbf{w}^T\boldsymbol{\Sigma}_1\mathbf{w}}=\frac{\mathbf{w}^T\left(\mathbf{u}_0-\mathbf{u}_1\right)\left(\mathbf{u}_0-\mathbf{u}_1\right)^T\mathbf{w}}{\mathbf{w}^T\left(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1\right)\mathbf{w}}$$

- 定义类内离散度矩阵和类间离散度矩阵

$$\mathbf{S}_w=\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1 \quad \mathbf{S}_b=\left(\mathbf{u}_0-\mathbf{u}_1\right)\left(\mathbf{u}_0-\mathbf{u}_1\right)^T$$

- 则目标函数简化为 $J=\dfrac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}}$

线性判别分析

- 等价表示（与 $\mathbf{w}$长度无关，即 $J(\alpha\mathbf{w})=J(\mathbf{w}),\quad \forall\alpha\in\mathbb{R}$）

$$\max J=\max\frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \quad\Longleftrightarrow\quad \begin{aligned}&\min\quad -\mathbf{w}^T\mathbf{S}_b\mathbf{w}\\ &\text{s.t.}\quad \mathbf{w}^T\mathbf{S}_w\mathbf{w}=1\end{aligned}$$

长度无关

- 构建拉格朗日乘子，转化为无约束的优化问题

$$L(\mathbf{w},\lambda)=-\mathbf{w}^T\mathbf{S}_b\mathbf{w}+\lambda\left(\mathbf{w}^T\mathbf{S}_w\mathbf{w}-1\right)$$

- 对变量求导

$$\frac{\partial L(\mathbf{w},\lambda)}{\partial\mathbf{w}}=-2\mathbf{S}_b\mathbf{w}+2\lambda\mathbf{S}_w\mathbf{w}$$

- 令导数为0，可得 $\mathbf{S}_b\mathbf{w}=\lambda\mathbf{S}_w\mathbf{w}$，广义特征分解可求。或更直观的方法：

线性判别分析

- 因为 $\mathbf{S}_b = (\mathbf{u}_0 - \mathbf{u}_1)(\mathbf{u}_0 - \mathbf{u}_1)^T$，可知

$$\mathbf{S}_b \mathbf{w} = (\mathbf{u}_0 - \mathbf{u}_1)\underbrace{(\mathbf{u}_0 - \mathbf{u}_1)^T \mathbf{w}}_{\text{实数}} = \alpha(\mathbf{u}_0 - \mathbf{u}_1)$$

- 代入前式 $\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}$ 可得

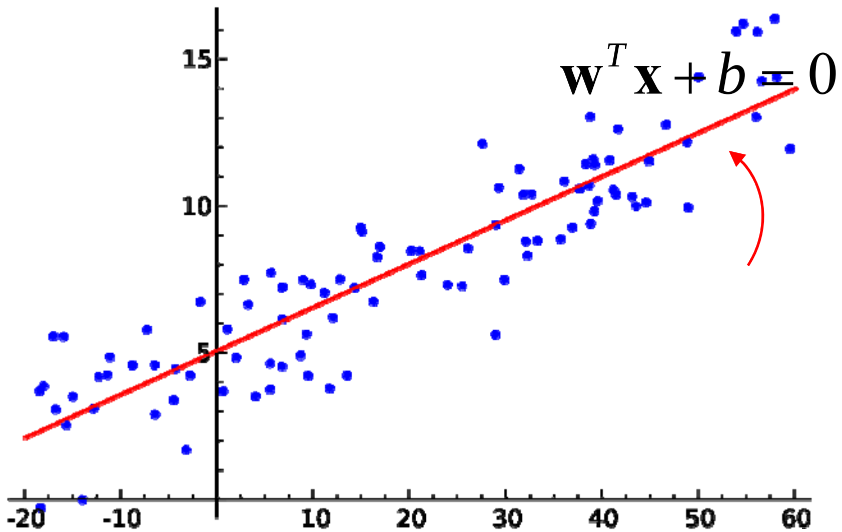
$$\alpha(\mathbf{u}_0 - \mathbf{u}_1) = \lambda \mathbf{S}_w \mathbf{w} \quad \Longrightarrow \quad \mathbf{w} = \alpha\lambda \mathbf{S}_w^{-1}(\mathbf{u}_0 - \mathbf{u}_1)$$

- 因 α, λ 均为实数，且$\mathbf{w}$**长度无关**，因此可把前面系数置为1，得最终解

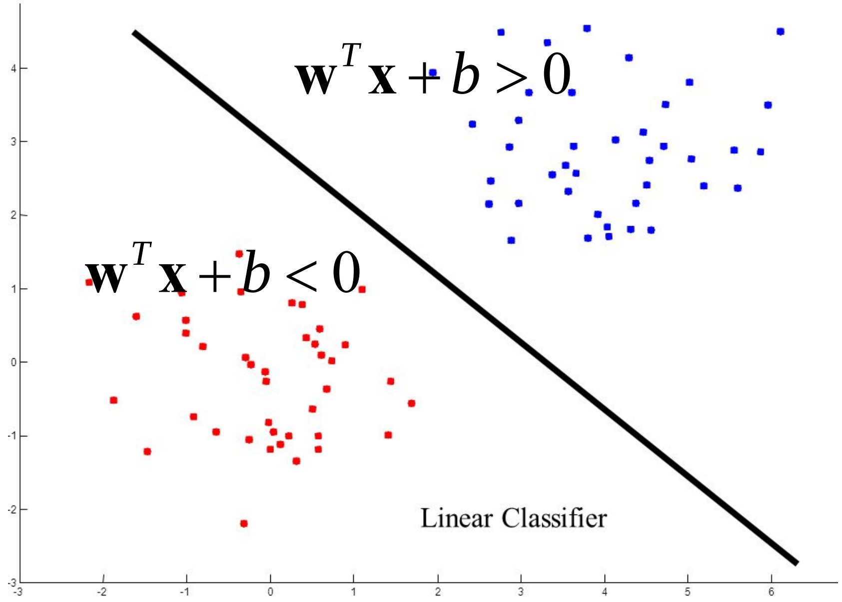
$$\mathbf{w}^* = \mathbf{S}_w^{-1}(\mathbf{u}_0 - \mathbf{u}_1)$$

线性拟合与线性分类关系

- 都是寻找一个线性方程，但是
  - 拟合：让尽可能多的点落到线上（**满足线性等式**）
  - 分类：让不同类的点位于线的两边 (**满足线性不等式**)



线性拟合



线性分类

小结

- 线性回归

$$J(\hat{\mathbf{w}})=\sum_{i=1}^{n}\left(f(\mathbf{x}_i)-y_i\right)^2=(\mathbf{f}-\mathbf{y})^T(\mathbf{f}-\mathbf{y})=\left(\mathbf{X}\hat{\mathbf{w}}-\mathbf{y}\right)^T\left(\mathbf{X}\hat{\mathbf{w}}-\mathbf{y}\right)$$

- 岭回归

$$J(\hat{\mathbf{w}})=\left(\mathbf{X}\hat{\mathbf{w}}-\mathbf{y}\right)^T\left(\mathbf{X}\hat{\mathbf{w}}-\mathbf{y}\right)+\lambda\|\mathbf{w}\|^2$$

- 对数几率回归（分类器）

$$l(\mathbf{w},b)=\sum_{i=1}^{n}\ln p(y_i=j\mid\mathbf{x}_i,\mathbf{w},b);\quad j\in\{0,1\}$$

- 线性判别分析

$$J=\frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}}$$