
线性回归

陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

实例

检验项目	测定结果	单位	参考范围
1. 白细胞计数	6.30	$10^9/L$	4.00 - 10.0
2. 中性粒细胞百分比	62.4	%	50.0 - 70.0
3. 淋巴细胞百分比	33.1	%	20.0 - 40
4. 单核细胞百分比	4.5	%	3.0 - 10.0
5. 中性粒细胞计数	4.10	$10^9/L$	2.00 - 7.0
6. 淋巴细胞计数	2.0	10^9	0.8 - 4.00
7. 单核细胞计数	0.20	$10^9/L$	0.12 - 0.8
8. 红细胞计数	4.33	$10^{12}/L$	4.09 - 5.74
9. 血红蛋白	117	↓ g/L	120 - 172
10. 红细胞压积	34.7	↓ %	38.0 - 50.8
11. 平均红细胞体积	80.0	↓ fL	83.9 - 99.1
12. 平均血红蛋白量	27.1	↓ pg	27.8 - 33.8
13. 平均血红蛋白浓度	338	g/L	320 - 355
14. 红细胞分布宽度CV	13.1	%	0.0 - 14.6
15. 血小板	287.0	$10^9/L$	85.0 - 363.0
16. 血小板平均分布宽度	10.9	↓ fL	12.0 - 22.0

回归分析（Regression Analysis）

定义：确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。

分类：线性回归、逻辑回归、多项式回归、岭回归。

应用：预测或估计（价格、年龄等）。

回归分析 (Regression Analysis)

职业：人类学家、气象学家、地理学家、统计学家、探险家

评价：比 10 个生物学家中的 9 个更懂数学和物理，

比 20 个 数学家中的 19 个更懂生物和医学。

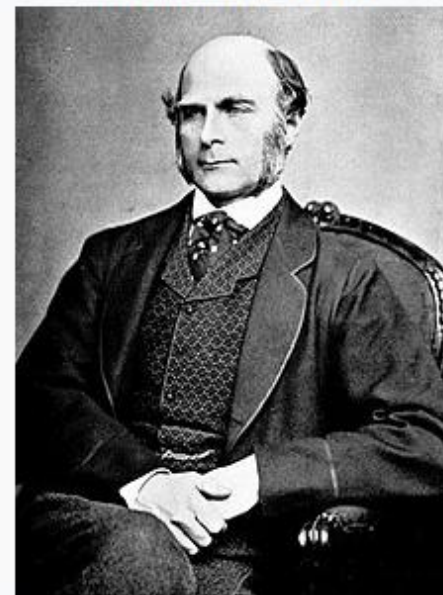
Galton 研究身高的遗传关系时发现了**回归效应**：

总的趋势是父亲的身高增加时，儿子的身高也倾向于增加。

1. 父亲高于平均身高时，他们的儿子身高比他更高的概率要小于比他更矮的概率；
2. 父亲矮于平均身高时，他们的儿子身高比他更矮的概率要小于比他更高的概率。

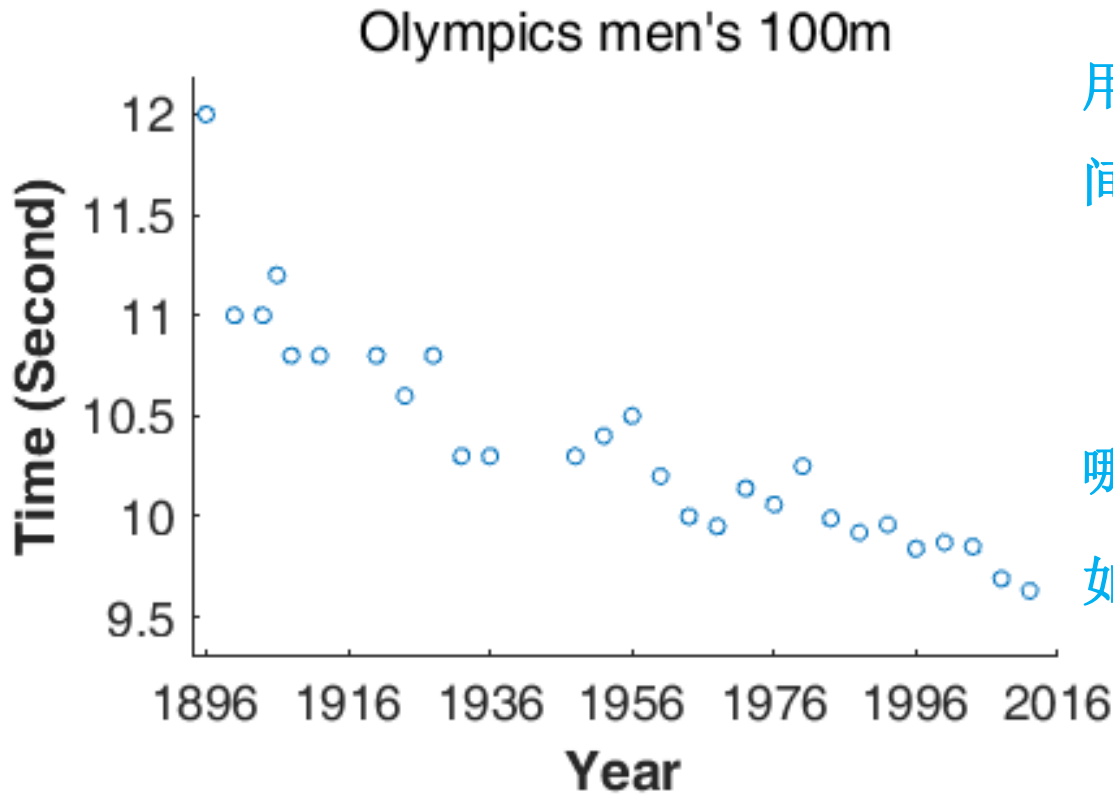
世界上最早研究大数据及其交叉学科的科学之一

Sir Francis Galton



Born	16 February 1822 Birmingham, West Midlands, England
Died	17 January 1911 (aged 88) Haslemere, Surrey, England
Residence	England
Nationality	British
Alma mater	King's College London Trinity College, Cambridge

Olympics men's 100m



用直线来描述年份与时间的关系：

$$y = w_1x + w_0$$

哪条直线是最佳的呢？

如何确定最优的参数 w_1 和 w_0 ？

损失函数

平方损失函数（Squared Loss Function）：

$$\mathcal{L}_i(\mathbf{w}_0, \mathbf{w}_1) = (y_i - (w_0 + w_1 x_i))^2$$

绝对损失函数（Absolute Loss Function）：

$$\mathcal{L}_i(\mathbf{w}_0, \mathbf{w}_1) = |y_i - (w_0 + w_1 x_i)|$$

平均损失函数（Average Loss Function）：

$$\mathcal{L}(\mathbf{w}_0, \mathbf{w}_1) = \frac{1}{n} \sum_i (y_i - (w_0 + w_1 x_i))^2$$

线性回归模型

线性回归(最小二乘)模型 (Linear Regression):

$$\operatorname{argmin}_{w_0, w_1} \mathcal{L}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n [y_i - (w_0 + w_1 x_i)]^2$$

把目标函数 $\mathcal{L}(w_0, w_1)$ 展开, 有

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n [y_i^2 - 2y_i(w_0 + w_1 x_i) + w_0^2 + 2w_0 w_1 x_i + w_1^2 x_i^2]$$

线性回归模型

$\mathcal{L}(w_0, w_1)$ 关于 w_0 求导:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n [y_i^2 - 2y_i(w_0 + w_1 x_i) + w_0^2 + 2w_0 w_1 x_i + w_1^2 x_i^2]$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = -2 \left(\frac{1}{n} \sum_{i=1}^n y_i \right) + 2w_0 + 2w_1 \left(\frac{1}{n} \sum_{i=1}^n x_i \right)$$

最优解: $w_0^* = \left(\frac{1}{n} \sum_{i=1}^n y_i \right) - w_1^* \left(\frac{1}{n} \sum_{i=1}^n x_i \right)$

线性回归模型

$\mathcal{L}(w_0, w_1)$ 关于 w_1 求导:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n [y_i^2 - 2y_i(w_0 + w_1x_i) + w_0^2 + 2w_0w_1x_i + w_1^2x_i^2]$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{2}{n} \sum_{i=1}^n x_i(w_0 - y_i) + 2w_1 \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right)$$

最优解: $w_1^* = -\frac{1}{n} \sum_{i=1}^n x_i(w_0^* - y_i) / \frac{1}{n} \sum_{i=1}^n x_i^2$

线性回归模型

把 w_0^* 代入 w_1^* 的方程，化简后有：

$$w_1^* = \frac{\sum_{i=1}^n y_i (x_i - \frac{1}{n} \sum_{i=1}^n x_i)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2}$$

$$w_0^* = \left(\frac{1}{n} \sum_{i=1}^n y_i \right) - w_1^* \left(\frac{1}{n} \sum_{i=1}^n x_i \right)$$

线性回归模型

$$\text{令 } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i \quad \overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2$$

则

$$w_1^* = \frac{\sum_i y_i \left(x_i - \frac{1}{n} \sum_{i=1}^n x_i \right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2} = \frac{\frac{1}{n} \left(\sum_{i=1}^n y_i \left(x_i - \frac{1}{n} \sum_{i=1}^n x_i \right) \right)}{\frac{1}{n} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right)} = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$w_0^* = \left(\frac{1}{n} \sum_{i=1}^n y_i \right) - w_1^* \left(\frac{1}{n} \sum_{i=1}^n x_i \right) = \bar{y} - w_1^* \bar{x}$$

线性回归模型

对于给定数据，线性回归模型

$$\operatorname{argmin}_{w_0, w_1} \frac{1}{n} \sum_{i=1}^n [y_i - (w_0 + w_1 x_i)]^2$$

的解为：

$$w_1^* = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad w_0^* = \bar{y} - w_1^* \bar{x}$$

人工例子

对下面一组数据进行线性回归分析：

x	1	3	5
y	4.8	11.3	17.2

解：

i	x_i	y_i	$x_i * y_i$	$x_i * x_i$
1	1	4.8	1 x 4.8	1 x 1
2	3	11.3	3 x 11.3	3 x 3
3	5	17.2	5 x 17.2	5 x 5
平均值				

人工例子

对下面一组数据进行线性回归分析：

x	1	3	5
y	4.8	11.3	17.2

解：

i	x_i	y_i	$x_i * y_i$	$x_i * x_i$
1	1	4.8	1 x 4.8	1 x 1
2	3	11.3	3 x 11.3	3 x 3
3	5	17.2	5 x 17.2	5 x 5
平均值	3	11.1	41.57	11.67

人工例子

对下面一组数据进行线性回归分析：

x	1	3	5
y	4.8	11.3	17.2

解：

i	xi	yi	xi * yi	xi * xi
平均值	3	11.1	41.57	11.67

$$w_1^* = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{41.57 - 3 \times 11.1}{11.67 - (3)^2} = 3.1$$

$$w_0^* = \bar{y} - w_1^* \bar{x} = 11.1 - 3.1 \times 3 = 1.8$$

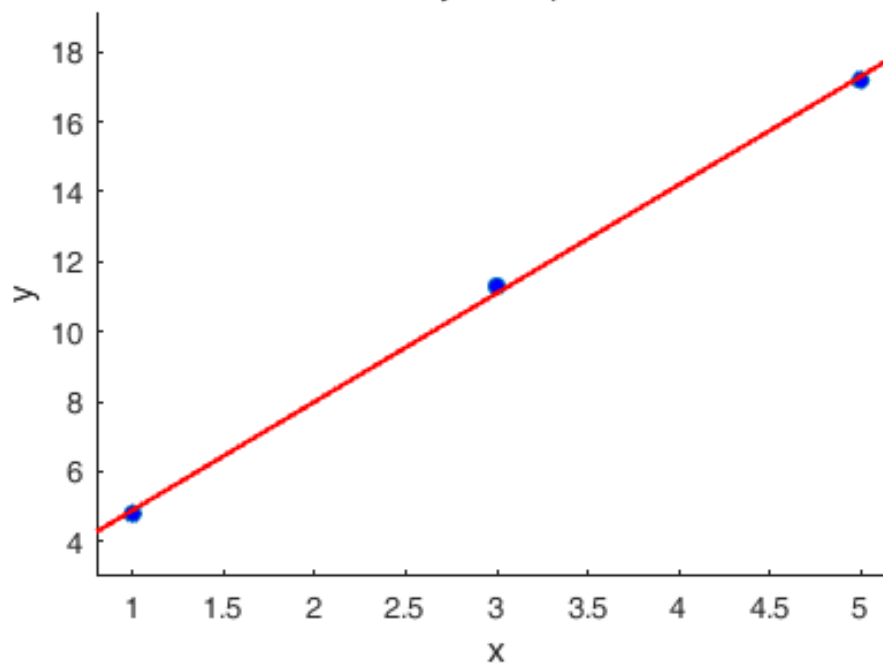
所求直线方程为： $y = 3.1 x + 1.8$

人工例子

数据:

x	1	3	5
y	4.8	11.3	17.2

Toy example



Olympics men's 100m

经过计算得到:

i	xi	yi	xi * yi	xi * xi
平均值	1954.5	10.36	20236.2	3.82×10^6

因此:

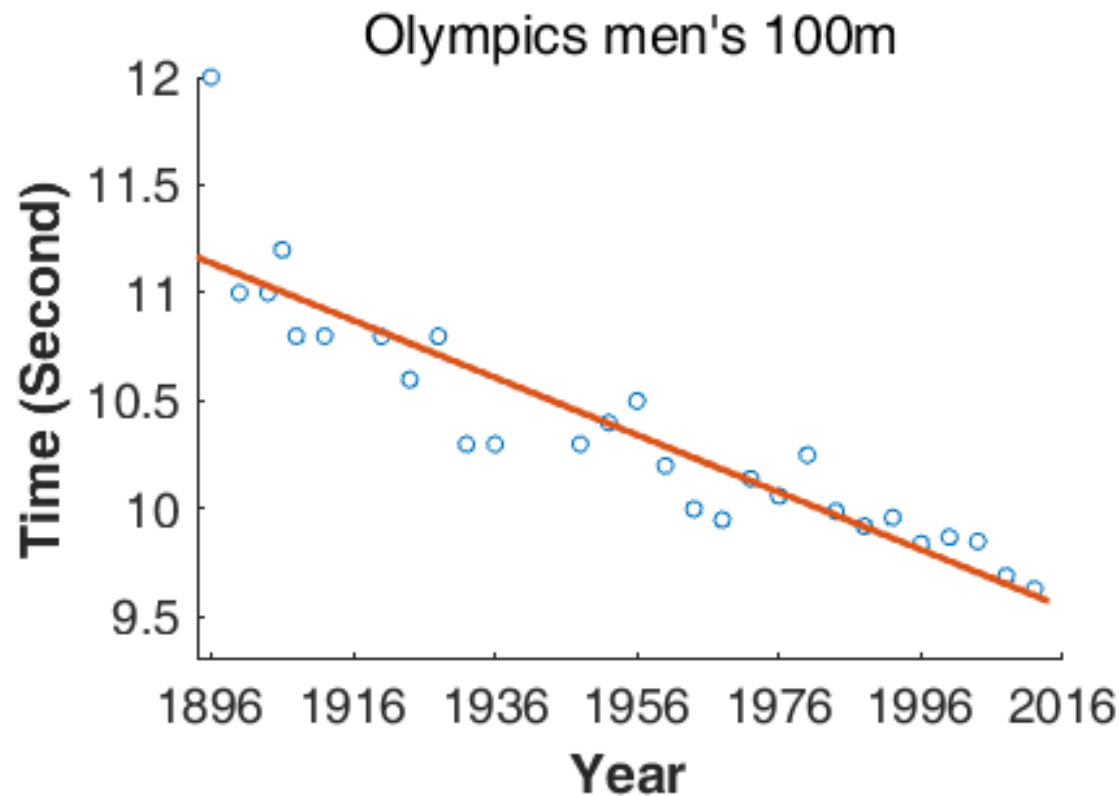
$$w_1^* = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{20236.2 - 1954.5 \times 10.36}{3.82 \times 10^6 - (1954.5)^2} = -0.0133$$

$$w_0^* = \bar{y} - w_1^* \bar{x} = 10.36 + 0.0133 \times 1954.5 = 36.355$$

所求直线方程为: $y = -0.0133 x + 36.355$

Olympics men's 100m

$$y = -0.0133x + 36.355$$



预测 (Prediction)

留一法 (Leave one out) :

把 $x = 2016, x = 2020$ 分别代入直线方程,

$$y = -0.0133x + 36.355$$

得到 $y = 9.54, y = 9.49$ 。

然而, 2016年里约奥运会男子100米最好成绩是9.81s (博尔特)。

说明:

- 线性回归算法学习了一条直线, 或者说两个参数(Paremetrics)
- 我们可以利用线性回归算法对未知数据进行预测(Prediction)
- 线性回归的效果主要取决于数据本身的分布情况(Distribution)

多元线性回归

线性回归：寻求最优参数 w_0, w_1 ，使得

$$y_i \approx w_0 + w_1 x_i = \hat{x}_i \hat{w}$$

其中 $\hat{w} = (w_0, w_1)^T$, $\hat{x}_i = (1, x_i)$.

考虑 d 维空间中的数据点 $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$,

$$y_i \approx w_0 + (w_1, w_2, \dots, w_d) \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = \hat{x}_i \hat{w}$$

其中 $\hat{w} = (w_0, w_1, \dots, w_d)^T$, $\hat{x}_i = (1, x_{i1}, \dots, x_{id})$.

多元线性回归

令 $\mathbf{y} = (y_1, y_2, \dots, y_n)^T \in \mathbb{R}^n$

$$\mathbf{X} = \begin{pmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \vdots \\ \widehat{x}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \in \mathbb{R}^{n \times (d+1)}$$

则

$$\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \widehat{x}_1 \widehat{w} \\ \widehat{x}_2 \widehat{w} \\ \vdots \\ \widehat{x}_n \widehat{w} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix}$$

多元线性回归

因此，单变量线性回归模型

$$\operatorname{argmin}_{w_0, w_1} \frac{1}{n} \sum_{i=1}^n [y_i - (w_0 + w_1 x_i)]^2$$

可以推广到多变量回归模型

$$\operatorname{argmin}_{\hat{\mathbf{w}}} \mathcal{L}(\hat{\mathbf{w}}) = \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2$$

该问题的最优解：

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

人工例子（回顾）：

对下面一组数据进行线性回归分析：

x	1	3	5
y	4.8	11.3	17.2

解： $y = 3.1x + 1.8$

首先构造矩阵

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix}, \quad y = \begin{pmatrix} 4.8 \\ 11.3 \\ 17.2 \end{pmatrix}$$

人工例子（回顾）：

计算矩阵乘积 $X^T X$ 及其逆矩阵 $(X^T X)^{-1}$ ：

$$X^T X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix} = \begin{pmatrix} 3 & 9 \\ 9 & 35 \end{pmatrix}$$

$$(X^T X)^{-1} = \frac{1}{3 \times 35 - 9 \times 9} \begin{pmatrix} 35 & -9 \\ -9 & 3 \end{pmatrix}$$

利用公式 $\hat{w}^* = (X^T X)^{-1} X^T y$ ：

$$\hat{w}^* = \frac{1}{24} \begin{pmatrix} 35 & -9 \\ -9 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 4.8 \\ 11.3 \\ 17.2 \end{pmatrix} = \begin{pmatrix} 1.8 \\ 3.1 \end{pmatrix}$$

多元线性回归

对于给定数据，线性回归模型

$$\operatorname{argmin}_{w_0, w_1} \frac{1}{n} \sum_{i=1}^n [y_i - (w_0 + w_1 x_i)]^2$$

的解为：公式1.

$$w_1^* = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad w_0^* = \bar{y} - w_1^* \bar{x}$$

公式2.

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

概率解释 (Probabilistic interpretation)

假设变量 y_i 和变量 x_i 满足:

$$\varepsilon_i = y_i - \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i$$

其中误差 ε_i 服从高斯分布 $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2)$:

$$p(\varepsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\varepsilon_i)^2}{2\sigma^2}\right)$$

那么

$$p(y_i | \hat{\mathbf{x}}_i; \hat{\mathbf{w}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i)^2}{2\sigma^2}\right)$$

似然函数 (Likelihood function)

概率 $p(\mathbf{y}_i | \hat{\mathbf{x}}_i; \hat{\mathbf{w}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(\mathbf{y}_i - \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i)^2}{2\sigma^2})$ 被看做输出变量 y_i 关于输入变量 \hat{x}_i 和固定参数 $\hat{\mathbf{w}}$ 的函数。

同时，这个量也可以被当做在已知变量 y_i 和变量 \hat{x}_i 的前提下，关于参数 $\hat{\mathbf{w}}$ 的函数，即似然函数 (**Likelihood function**)。

$$\mathcal{L}_i(\hat{\mathbf{w}}) = p(y_i | \hat{x}_i; \hat{\mathbf{w}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - \hat{\mathbf{w}}^T \hat{x}_i)^2}{2\sigma^2})$$

极大似然法 (Maximum Likelihood)

假设所有数据都是独立同分布 (independent and identically distributed , 简称 i.i.d.) 的, 则

$$\mathcal{L}(\hat{\mathbf{w}}) = \prod_{i=1}^n \mathcal{L}_i(\hat{\mathbf{w}}) = \prod_{i=1}^n p(\mathbf{y}_i | \hat{\mathbf{x}}_i; \hat{\mathbf{w}})$$

极大似然法: 令每个样本 $\hat{\mathbf{x}}_i$ 输出为 \mathbf{y}_i 的概率越大越好

$$\operatorname{argmax}_{\hat{\mathbf{w}}} \mathcal{L}(\hat{\mathbf{w}})$$

极大似然法 (Maximum Likelihood)

考虑对数似然函数 (Log-likelihood function) :

$$\begin{aligned}\ell(\hat{\mathbf{w}}) &= \ln \mathcal{L}(\hat{\mathbf{w}}) = \ln \prod_{i=1}^n p(\mathbf{y}_i | \hat{\mathbf{x}}_i; \hat{\mathbf{w}}) \\ &= \sum_{i=1}^n \ln p(\mathbf{y}_i | \hat{\mathbf{x}}_i; \hat{\mathbf{w}}) = \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i)^2\end{aligned}$$

因此, 下面两个优化模型等价:

$$\operatorname{argmax}_{\hat{\mathbf{w}}} \sum_{i=1}^n \ln p(\mathbf{y}_i | \hat{\mathbf{x}}_i; \hat{\mathbf{w}}) \quad \Leftrightarrow \quad \operatorname{argmin}_{\hat{\mathbf{w}}} \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i)^2$$

最小二乘法 vs 极大似然法

对于给定训练集 $\{(x_i, y_i)\}_{i=1}^n$ ，我们希望找到数据 x_i 与其对应的标签 y_i 之间的函数关系 f ，即

$$y_i \approx f(x_i)$$

从概率角度来说，我们希望找到恰当的分布函数 p ，使得在给定 x_i 时 y_i 发生的概率越大越好，即

$$p(y_i|x_i) \approx 1$$

殊途同归

最小二乘法:

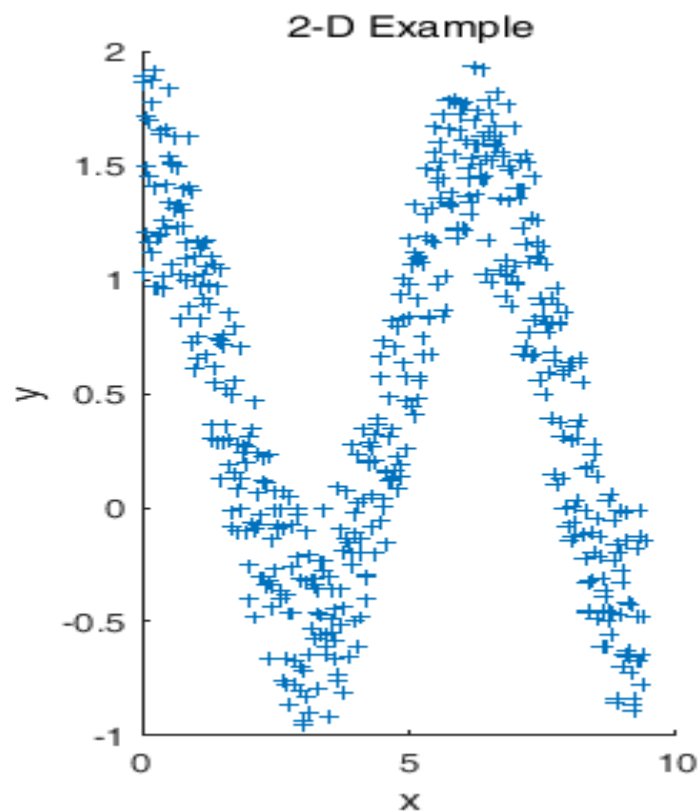
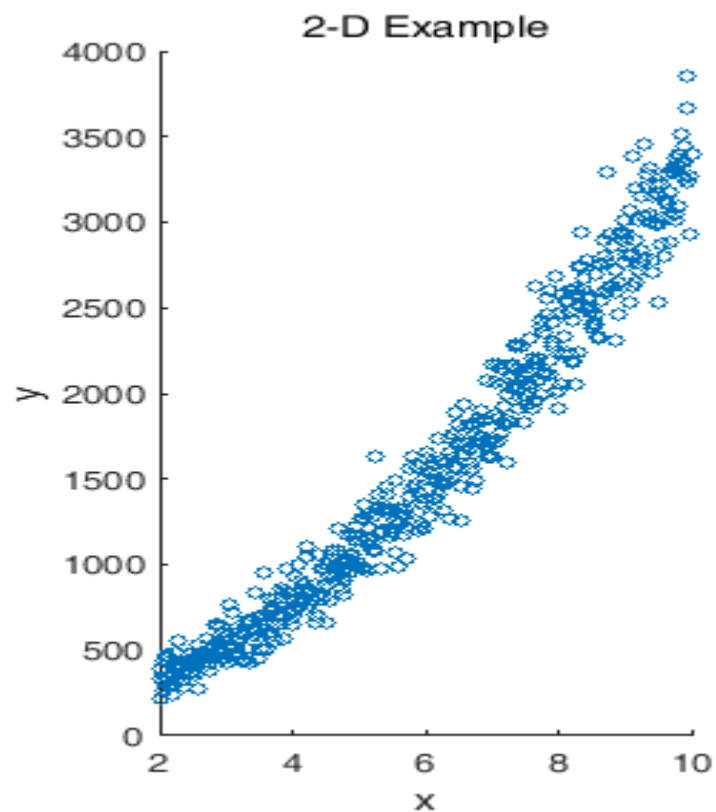
$$\operatorname{argmin}_f \frac{1}{n} \sum_{i=1}^n [y_i - \textcolor{red}{f}(x_i)]^2$$

极大似然法:

$$\operatorname{argmax}_p \ln \prod_{i=1}^n \textcolor{red}{p}(y_i | x_i)$$

当 f 是线性函数， p 满足高斯分布时，两个模型等价。

线性回归模型（延伸）



广义线性回归模型

线性回归模型：

$$y = w^T x + b$$

广义线性回归模型（Generalized linear model）：

$$g(y) = w^T x + b$$

或者

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

其中 g 被称为 **联系函数（link function）**。

回归任务应用专题

回归模型的应用

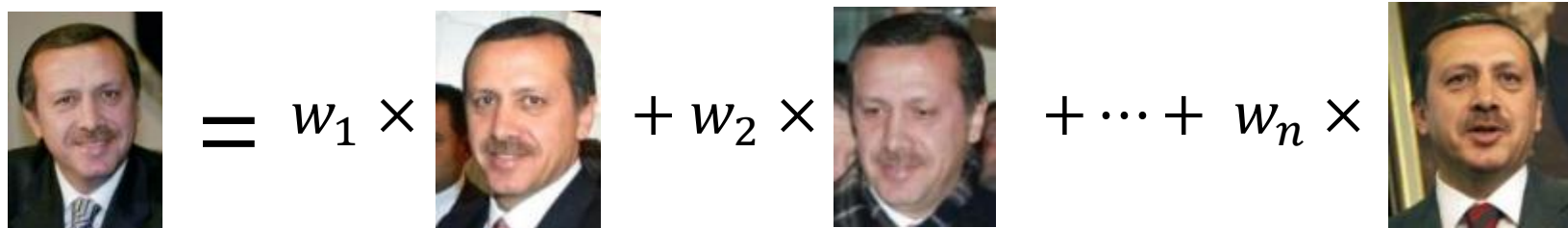
陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

Linear regression for face recognition

主要思想：一个样本可以被其他同类样本线性表示。


$$\text{Target Face} = w_1 \times \text{Face 1} + w_2 \times \text{Face 2} + \dots + w_n \times \text{Face n}$$

Naseem I, Togneri R, Bennamoun M. Linear regression for face recognition[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2010, 32(11):2106-2112.

Linear regression for face recognition

- 根据前面假设，我们可以利用一个线性方程来描绘这种线性关系：
 - $\tilde{x} \approx X_c \omega_c \in \mathbb{R}^{d \times 1}$, (1)
 - $X_c = [x_1, x_2, \dots, x_m] \in \mathbb{R}^{d \times m}$
 - $\omega_c = [\omega_1, \omega_2, \dots, \omega_m] \in \mathbb{R}^{1 \times m}$
- \tilde{x} 是输入测试样本， X_c 是所有 c 类样本组成的样本矩阵， ω_c 是 X_c 关于 \tilde{x} 的线性表示的系数。如果假设成立且输入样本属于 c 类样本，则
 - $c = \underset{i}{\operatorname{argmin}} ||\tilde{x} - X_i \omega_i||_2^2, i \in \{1, 2, \dots, c, \dots, C\}$

Linear regression for face recognition

- 其他用于人脸识别的经典线性回归算法:
 - Sparse Representation (稀疏表示)[1]
 - Collaborative Representation (协同表示)[2]
- 假设调整: 与所有人脸样本线性相关, 但同类样本在线性表示中贡献最大:

$$\tilde{x} \approx X\omega \in \mathcal{R}^{d \times 1}, \quad \tilde{X}_i \subset X, \quad \tilde{\omega}_i \subset \omega$$

$$c = \underset{i}{\operatorname{argmin}} ||\tilde{x} - \tilde{X}_i \tilde{\omega}_i||_2^2, i \in \{1, 2, \dots, c, \dots, C\}$$

[1] Wright J, Yang A Y, Ganesh A, et al. Robust Face Recognition via Sparse Representation[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2008, 31(2):210-227.

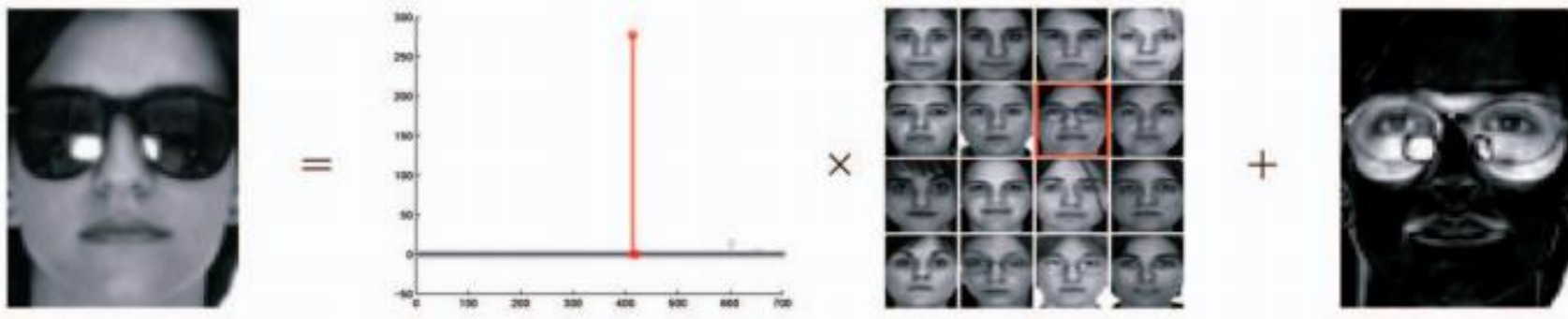
[2] Zhang L, Yang M, Feng X. Sparse representation or collaborative representation: Which helps face recognition?[C] // *IEEE International Conference on Computer Vision (ICCV)*, 2012:471-478.

Sparse Representation

- 思想:

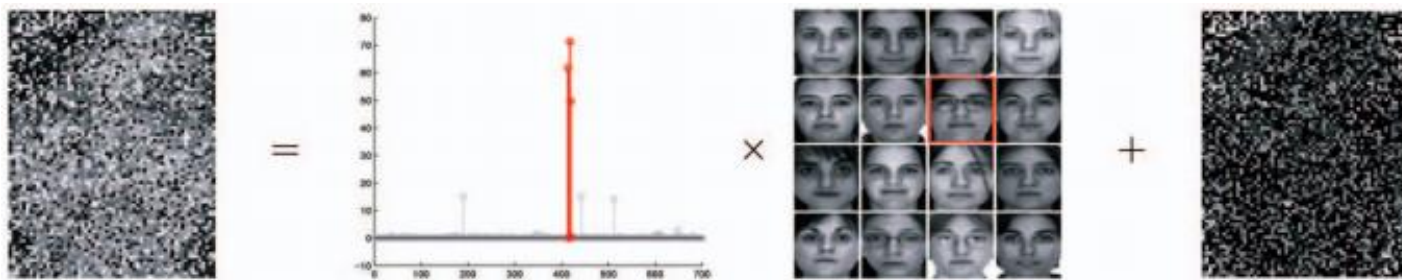
- ① 强制选择极少数训练样本线性表示输入样本。
- ② 被最终选择样本应与输入样本具有很强的关联性。
- ③ 强调线性表示系数 ω 的稀疏性。

$$\min_w ||\tilde{x} - X\omega||_2^2 + \beta ||\omega||_1$$

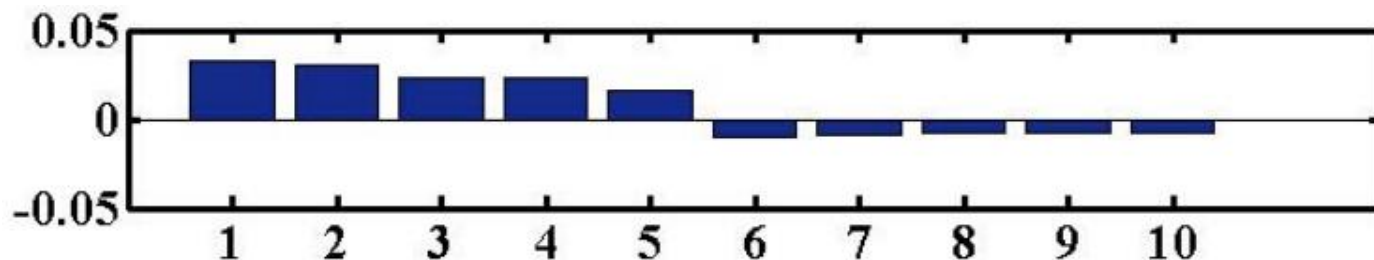


Sparse Representation

- 优点：
 - 对噪声和遮挡等影响因素比较鲁棒。



- 极强的相关样本选择能力。



Collaborative Representation

- 稀疏表示的缺点：
 - 计算代价高
 - 假设过强，表示不够光滑(Representation is not smooth)，忽略其他样本的贡献。

- 解决思路：
 - 把1范数约束项进行松弛：

$$\min_w ||\tilde{x} - X\omega||_2^2 + \beta ||\omega||_2^2$$

- 典型的最小二乘问题，直接求解。
- 协同表示的特性与稀疏表示相似，对噪声与遮挡也具有一定的鲁棒性

Evaluation of LR approaches

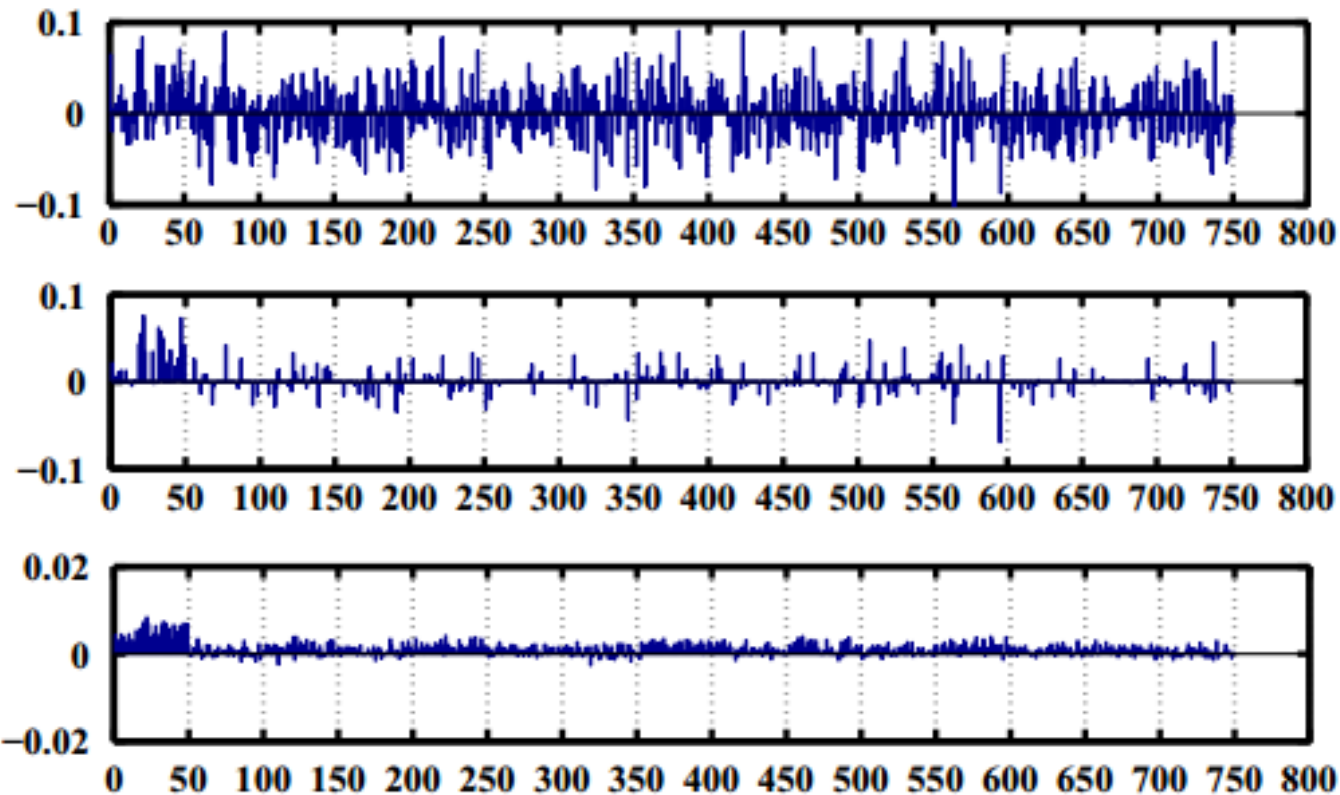
- 性能评估（2折交叉验证）：

Table 1
Classification accuracy comparison on ORL, AR, Scene15 and Caltech256 databases.

Methods	Classification accuracy (Mean \pm STD,%)				
	ORL	AR	Scene15	Caltech256	COIL20
NNC	87.25 \pm 1.06	66.01 \pm 0.08	61.47 \pm 1.41	49.25 \pm 1.23	84.51 \pm 1.47
AHC [24]	88.21 \pm 0.41	69.35 \pm 0.25	74.49 \pm 1.60	56.70 \pm 2.14	89.87 \pm 1.41
RFC [29]	88.53 \pm 2.13	69.73 \pm 0.88	72.27 \pm 2.59	56.32 \pm 3.00	88.11 \pm 2.37
LIBSVM [30]	90.25 \pm 3.18	68.10 \pm 0.67	74.60 \pm 2.17	60.85 \pm 2.47	87.50 \pm 1.18
LRC [21]	88.75 \pm 3.18	68.75 \pm 0.43	60.33 \pm 3.30	43.00 \pm 0.85	88.82 \pm 1.08
SRC [2]	92.00 \pm 3.54	63.87 \pm 0.42	67.20 \pm 1.13	48.05 \pm 0.64	88.19 \pm 0.98
SGC [31]	88.50 \pm 2.83	73.27 \pm 0.42	71.27 \pm 2.17	50.60 \pm 3.45	88.33 \pm 0.59
CRC [12]	92.75 \pm 3.89	68.25 \pm 0.42	67.60 \pm 3.58	50.06 \pm 0.71	88.81 \pm 1.10
CSSRC	94.25 \pm 3.18	77.14 \pm 0.34	74.60 \pm 2.36	61.15 \pm 1.63	89.17 \pm 1.37

Observation from real data

- the regression coefficients of LR, SR, CR(from top to down, the test sample belongs to the category of the first 50 training samples)



线性回归小结 I

- 我们介绍了回归任务的思想以及如何建立模型。

分析变量之间的关系、线性回归模型

- 我们定义了损失函数来评估线性回归模型的好坏。

平方损失函数、极大似然函数

- 我们推导了线性回归模型中两个参数的显式表达。

最小二乘法（向量与矩阵形式）

线性回归小结 II

- 我们**使用了**线性回归算法对两组数据进行分析。

人工例子、奥运会男子100米

- 我们**应用了**线性回归算法进行预测，并检验结果。

2016、2020奥运会男子100米

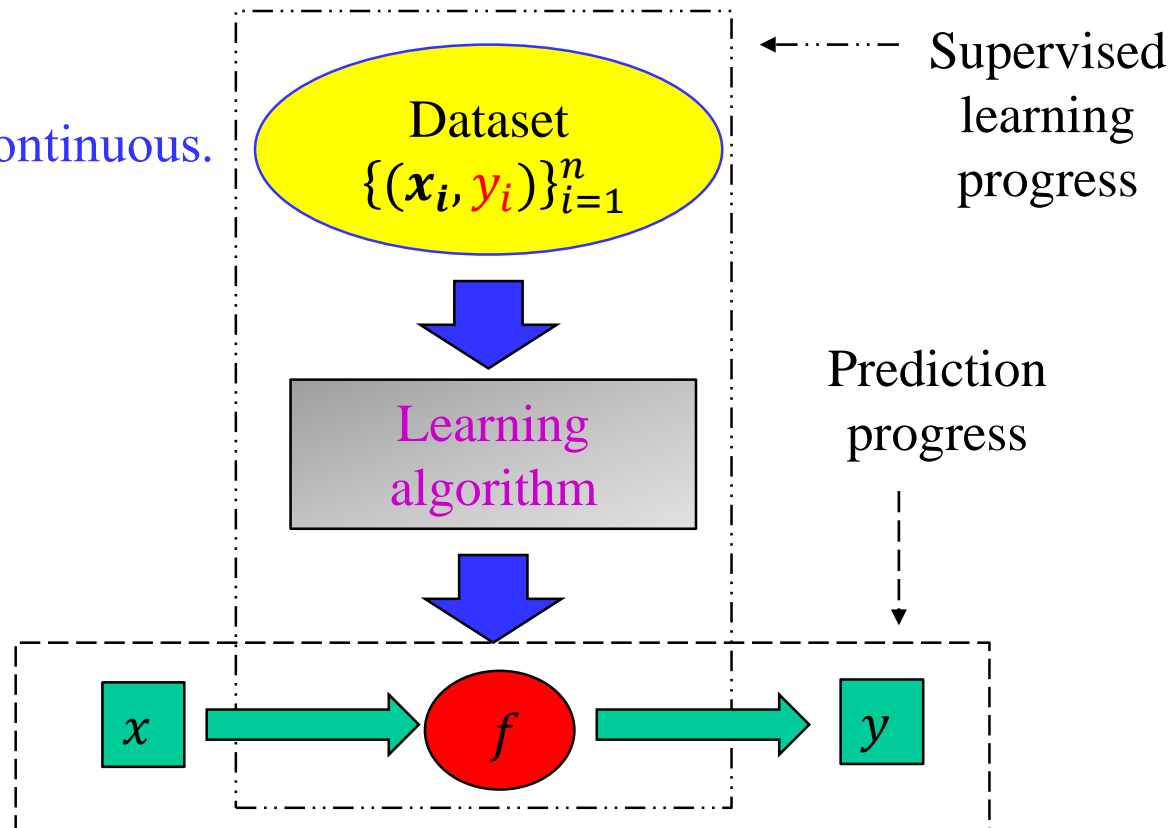
- 我们**描述了**线性回归模型的概率解释。

极大似然法

线性回归小结 III

Learning with a **teacher**

Regression:
Label y_i 's are continuous.



思考题

多元线性回归模型

$$\operatorname{argmin}_{\hat{\mathbf{w}}} \mathcal{L}(\hat{\mathbf{w}}) = \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2$$

该问题的最优解:

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

如果矩阵 $\mathbf{X}^T \mathbf{X}$ 的条件数太大, 会导致求逆不准确, 或者 $\hat{\mathbf{w}}^*$ 存在很多个解, 怎么办?

练习题

对于线性回归问题，给定

$$w_0^* = \left(\frac{1}{n} \sum_i y_i \right) - w_1^* \left(\frac{1}{n} \sum_i x_i \right)$$

$$w_1^* = -\frac{1}{n} \sum_i x_i (w_0^* - y_i) / \frac{1}{n} \sum_i x_i^2$$

试推导：

$$w_1^* = \frac{\sum_i y_i (x_i - \frac{1}{n} \sum_i x_i)}{\sum_i x_i^2 - \frac{1}{n} (\sum_i x_i)^2}$$

练习题

对于一维数据，试证明下列两种解法的等价性。

公式1.

$$w_1^* = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad w_0^* = \bar{y} - w_1^* \bar{x}$$

公式2.

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

练习题

用 Matlab 实现线性回归的算法：

1. 构造人工数据。提示：(x, y) 要呈线性分布。
2. 利用公式1和公式2求出直线方程。
3. 评价两种方法的优劣（运行时间、目标函数等）
4. 画图。（画出原始数据点云、直线）

逻辑回归

陈飞宇

fchen@cqu.edu.cn

办公室：软件学院529

前情提要

线性回归模型：

$$\operatorname{argmin}_{w_0, w_1} \frac{1}{n} \sum_{i=1}^n [y_i - (w_0 + w_1 x_i)]^2$$

建模，计算，应用。

概率解释（极大似然法）：

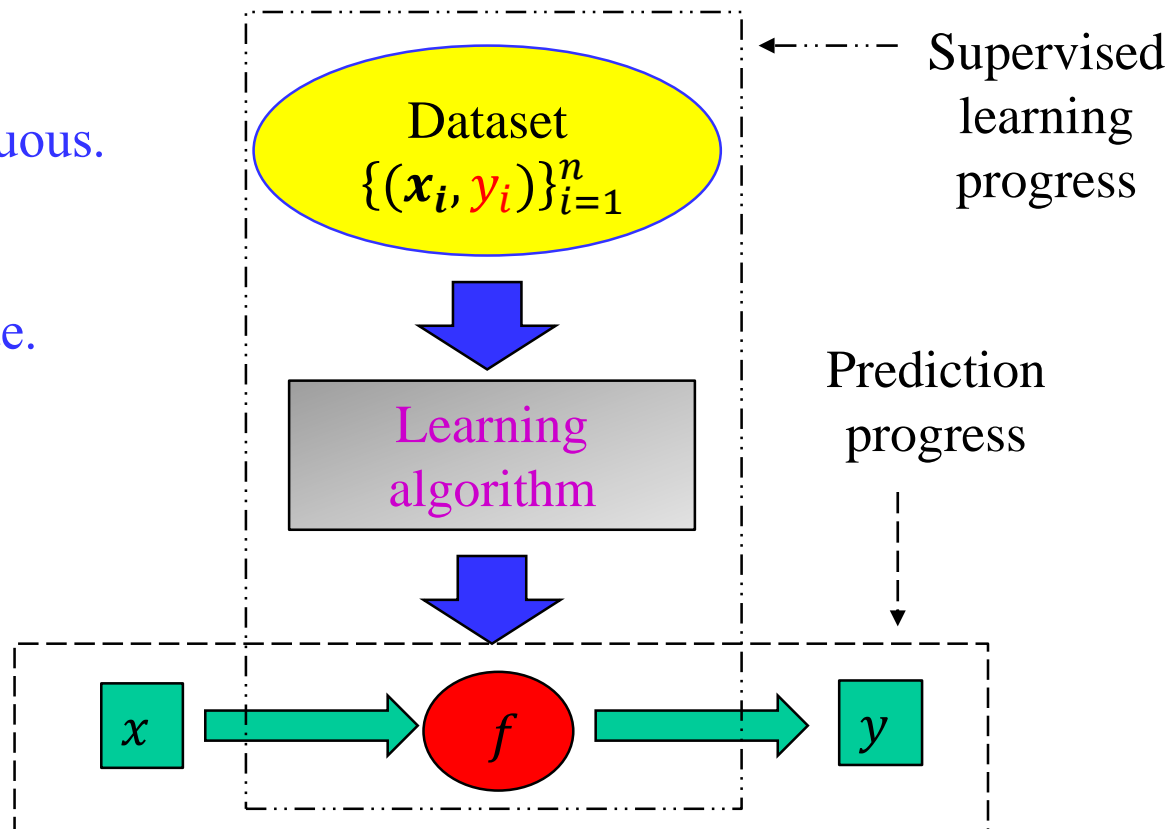
$$\operatorname{argmax}_{\hat{w}} \sum_{i=1}^n \ln p(y_i | \hat{x}_i; \hat{w})$$

监督学习 (Supervised Learning)

Learning with a **teacher**

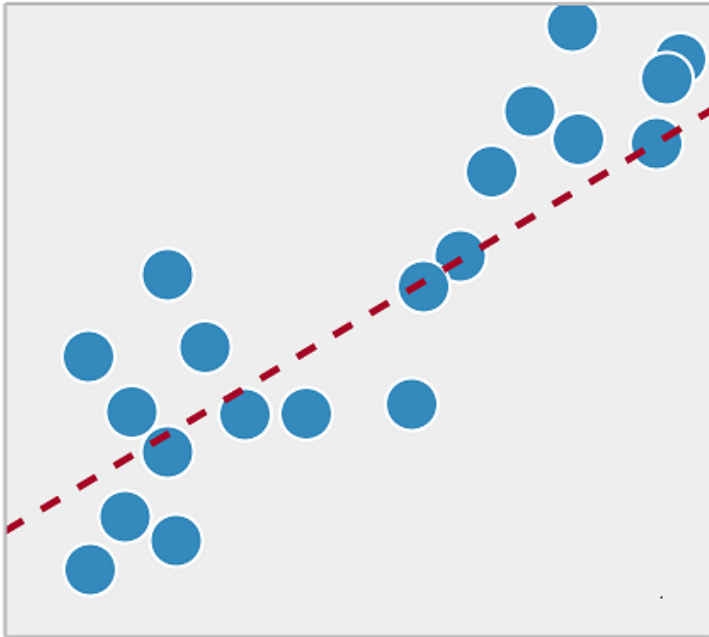
Regression:
 y_i 's are continuous.

Classification:
 y_i 's are discrete.

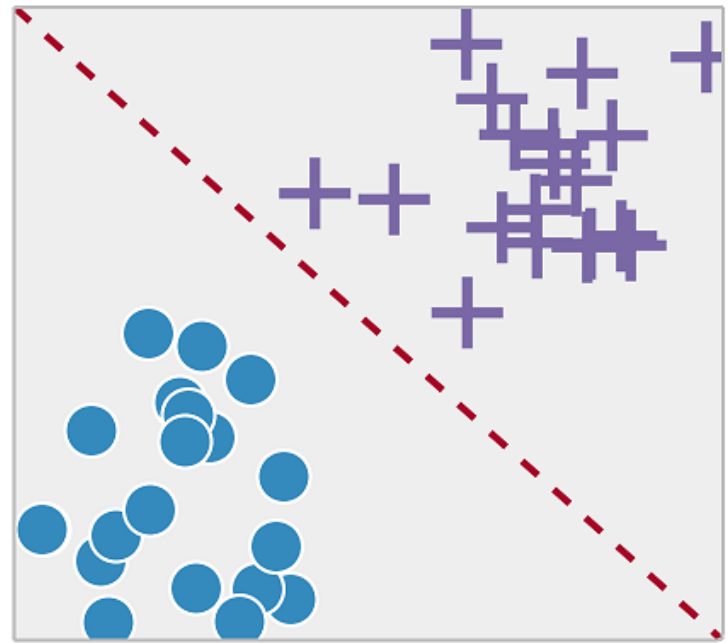


回归与分类

Regression



Classification



我们能否利用线性回归的思想解决分类任务（二分类）？

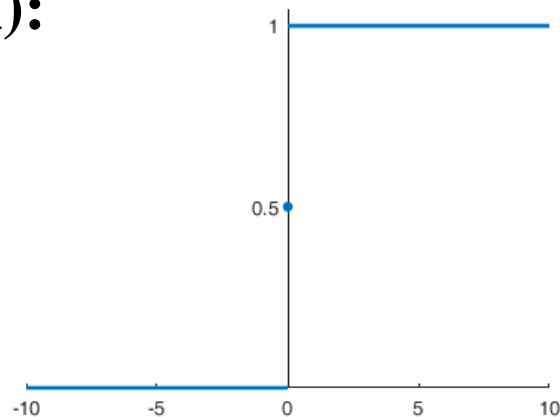
回归与分类

直观上说，可以规定直线上方的点为**正样本(Positive)**，
直线下方的点为**负样本(Negative)**。

本质上说，需要把连续值转化为离散值(例如: $\{0, 1\}$)。

单位阶跃函数 (Unit-step function):

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



对数几率函数（Logistic function）

由于单位阶跃函数不是一个连续函数，我们通常选择一些性质好的函数作为替代函数。

Unit-step function and logistic function

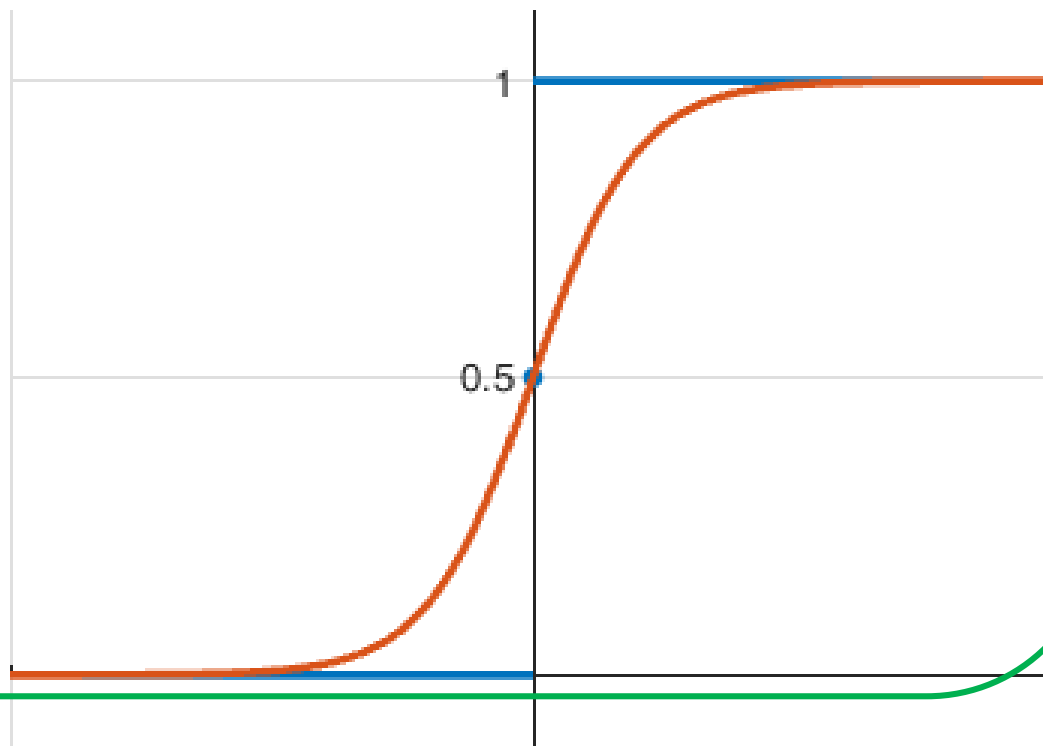
Logistic function:

$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g(+\infty) = 1$$

$$g(0) = 0.5$$

$$g(-\infty) = 0$$



对数几率函数（Logistic function）

对数几率函数 g 为任意阶可导函数，它的一阶导数为：

$$g' = \frac{1}{(1+e^{-z})^2} e^{-z} = \frac{1}{1+e^{-z}} \frac{e^{-z}}{1+e^{-z}} = g(1-g)$$

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

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

$$\ln \frac{y}{1-y} = z$$

逻辑回归（ Logistic regression ）

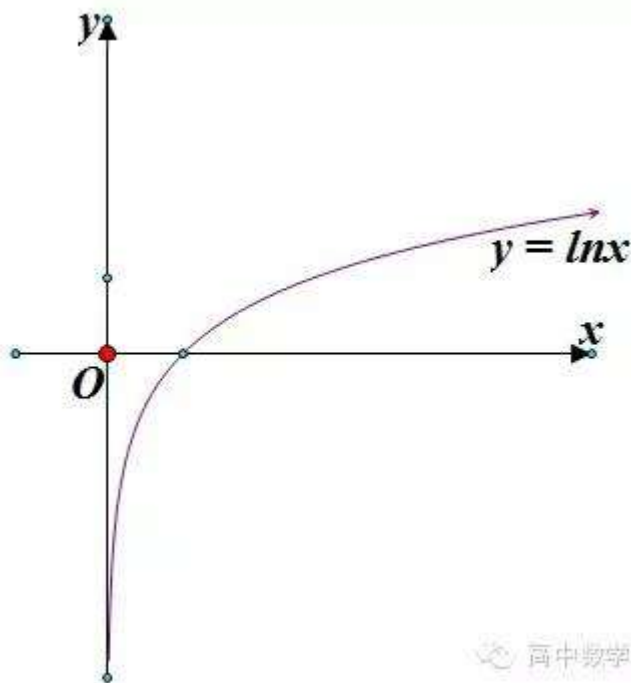
结合线性回归的思想与对数几率函数的特点，我们得到对数回归模型（Logistic regression/ logit regression）：

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

将 y 视为样本 x 作为正例的可能性，则 $1 - y$ 是其为反例的可能性。两者的比值 称为“几率” (Odds)， 反映了 x 作为正例的相对可能性。

逻辑回归（Logistic regression）

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



说明：

如果 x 在直线上方，则 $w^T x + b > 0$,

$\frac{y}{1-y} > 1$ 。意味着 x 作为正例的可能性大于其为反例的可能性。

如果 x 在直线下方， $w^T x + b < 0$, 则

$\frac{y}{1-y} < 1$ 。意味着 x 作为反例的可能性大于其为正例的可能性。

逻辑回归的本质是用线性回归的预测结果去逼近真实标记的对数几率。

逻辑回归 (Logistic regression)

对于二分类问题，输出变量 $y \in \{0, 1\}$ 。我们将 y 写成后验概率估计 $p(y = 1|x)$ ，则 $1 - y$ 可写成后验概率估计 $p(y = 0|x)$ ，我们有

$$p_1(x; \mathbf{w}) = p(y = 1|x) = \frac{e^{\mathbf{w}^T x + b}}{1 + e^{\mathbf{w}^T x + b}}$$

$$p_0(x; \mathbf{w}) = p(y = 0|x) = \frac{1}{1 + e^{\mathbf{w}^T x + b}}$$

任意样本的似然函数可以写成：

$$p(y_i | \hat{x}_i; \hat{\mathbf{w}}) = [p_1(\hat{x}_i; \hat{\mathbf{w}})]^{y_i} [p_0(\hat{x}_i; \hat{\mathbf{w}})]^{(1-y_i)}$$

极大似然法 (Maximum Likelihood)

逻辑回归模型:

$$\operatorname{argmax}_{\hat{\mathbf{w}}} \sum_{i=1}^n \ln p(\mathbf{y}_i | \hat{\mathbf{x}}_i; \hat{\mathbf{w}})$$

其中:

$$\begin{aligned} \ln p(\mathbf{y}_i | \hat{\mathbf{x}}_i; \hat{\mathbf{w}}) &= \ln ([p_1(\hat{\mathbf{x}}_i; \hat{\mathbf{w}})]^{y_i} [p_0(\hat{\mathbf{x}}_i; \hat{\mathbf{w}})]^{(1-y_i)}) \\ &= y_i \ln p_1(\hat{\mathbf{x}}_i; \hat{\mathbf{w}}) + (1 - y_i) \ln p_0(\hat{\mathbf{x}}_i; \hat{\mathbf{w}}) \end{aligned}$$

极大似然法 (Maximum Likelihood)

$$\operatorname{argmax}_{\hat{\mathbf{w}}} \sum_{i=1}^n (\mathbf{y}_i \ln \mathbf{p}(\mathbf{y} = \mathbf{1} | \hat{\mathbf{x}}_i; \hat{\mathbf{w}}) + (1 - \mathbf{y}_i) \ln \mathbf{p}(\mathbf{y} = \mathbf{0} | \hat{\mathbf{x}}_i; \hat{\mathbf{w}}))$$

说明:

当 $\mathbf{y}_i = 1$, 即 \mathbf{x}_i 是正样本时, 我们希望由参数 $\hat{\mathbf{w}}$ 确定的直线方程可以使得 $\mathbf{p}(\mathbf{y} = \mathbf{1} | \hat{\mathbf{x}}_i; \hat{\mathbf{w}})$ 最大化, 此时 $\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i$ 的值越大越好。

当 $\mathbf{y}_i = 0$, 即 \mathbf{x}_i 是负样本时, 我们希望由参数 $\hat{\mathbf{w}}$ 确定的直线方程可以使得 $\mathbf{p}(\mathbf{y} = \mathbf{0} | \hat{\mathbf{x}}_i; \hat{\mathbf{w}})$ 最大化, 此时 $\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i$ 的值越小越好。

极大似然法 (Maximum Likelihood)

分别把 p_i 的表达式代入 $\ln p_i$, 得到:

$$\ln p_1 = \ln \left(\frac{e^{\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i}}{1 + e^{\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i}} \right) = \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i - \ln \left(1 + e^{\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i} \right)$$

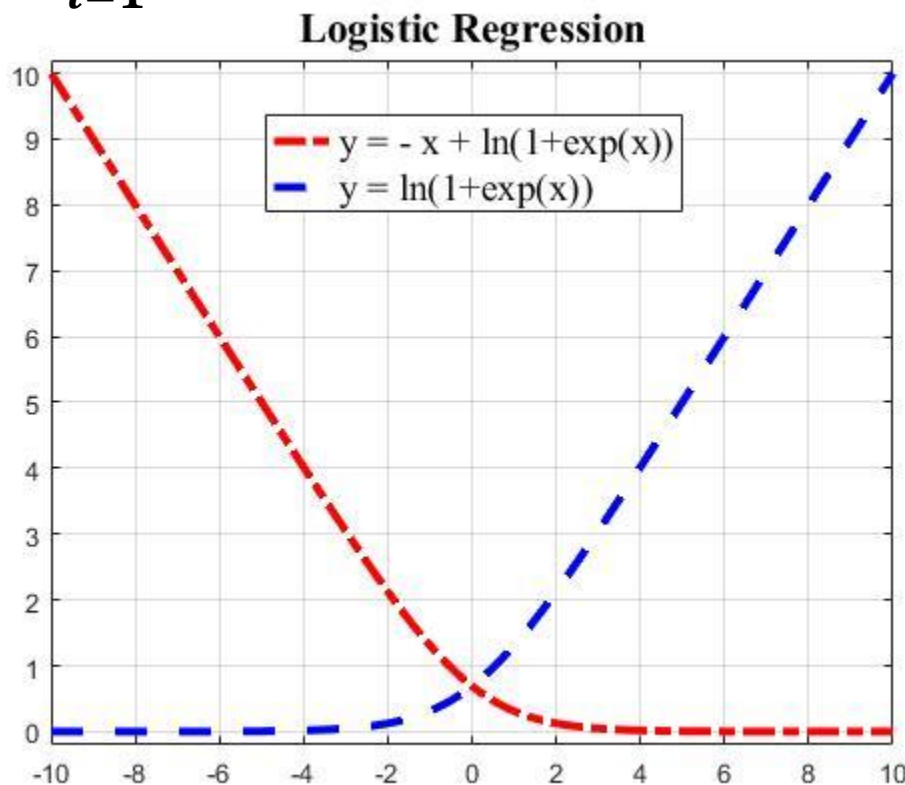
$$\ln p_0 = \ln \left(\frac{1}{1 + e^{\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i}} \right) = - \ln \left(1 + e^{\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i} \right)$$

特别地, 我们可以得到 $\ln p(y_i | \hat{\mathbf{x}}_i; \hat{\mathbf{w}})$ 的表达式

$$\ln p = y_i \ln p_1 + (1 - y_i) \ln p_0 = y_i \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i - \ln \left(1 + e^{\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i} \right)$$

逻辑回归 (Logistic Regression)

$$\operatorname{argmin}_{\hat{\mathbf{w}}} \sum_{i=1}^n [-y_i \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i + \ln(1 + e^{\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i})]$$



逻辑回归 (Logistic Regression)

求解无约束优化问题:

$$\operatorname{argmin}_{\hat{\mathbf{w}}} \sum_{i=1}^n [-\mathbf{y}_i \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i + \ln(1 + e^{\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i})]$$

数值方法 I: 牛顿法 (Newton's method)

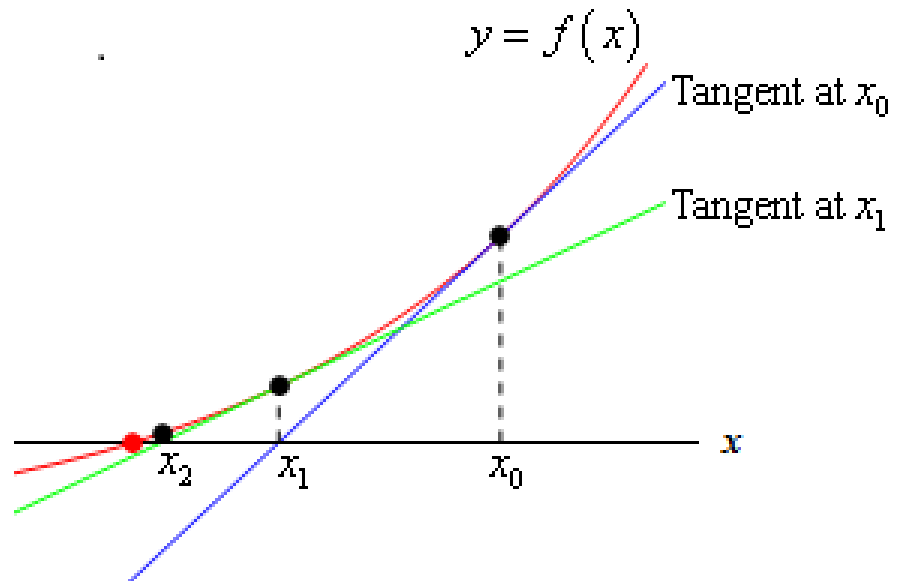
数值方法 II: 梯度下降法 (Gradient Decent Method)

牛顿法 (Newton's Method)

方程求根问题:

$$f(x) = 0$$

$$f'(x_n) = \frac{f(x_n) - f(x_{n+1})}{x_n - x_{n+1}}$$



$$x_{n+1} = x_n - \frac{f(x_n) - f(x_{n+1})}{f'(x_n)} \approx x_n - \frac{f(x_n)}{f'(x_n)}$$

牛顿法 (Newton's Method)

最小化问题:

$$\mathbf{x}^* = \operatorname{argmin}_x \ell(\mathbf{x}) \quad \longleftrightarrow \quad \ell'(\mathbf{x}^*) = \mathbf{0}$$

令 $f(\mathbf{x}) = \ell'(\mathbf{x})$, 则 $f'(\mathbf{x}) = \ell''(\mathbf{x})$

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{f(\mathbf{x}_n)}{f'(\mathbf{x}_n)} = \mathbf{x}_n - \frac{\ell'(\mathbf{x}_n)}{\ell''(\mathbf{x}_n)}$$

梯度下降法

迭代公式:

$$x_{n+1} = x_n - \gamma_n \nabla f(x_n)$$

其中 γ_n 是第n步下降时选取的步长，也称学习率。

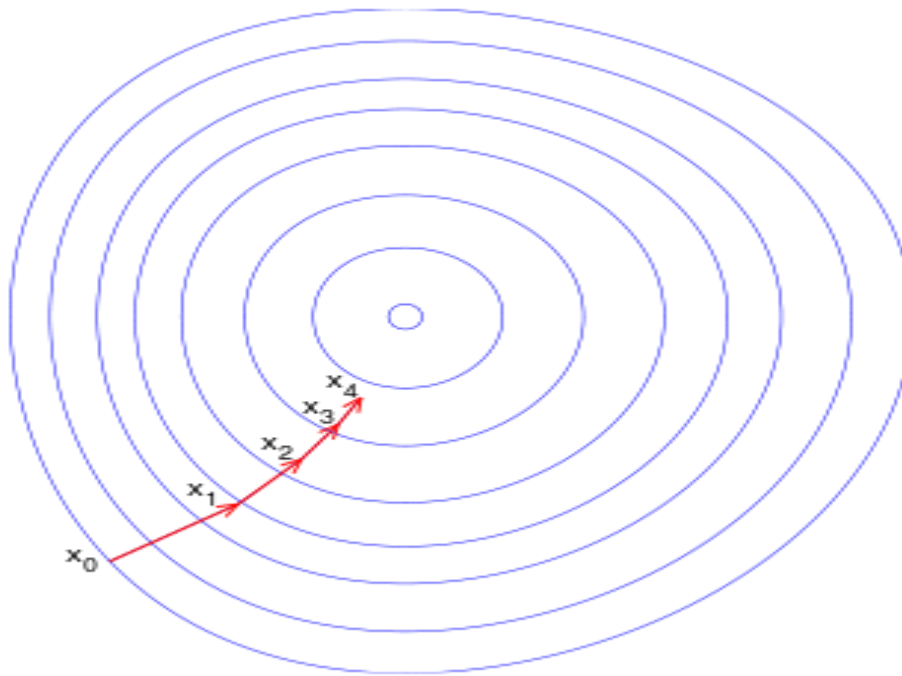
线搜索 (Barzilai-Borwein Step):

$$\gamma_n = \frac{(x_n - x_{n-1})^T (\nabla f(x_n) - \nabla f(x_{n-1}))}{\|\nabla f(x_n) - \nabla f(x_{n-1})\|^2}$$

梯度下降法

基本思想:

$$f(x_{n+1}) - f(x_n) = (x_{n+1} - x_n)^T \nabla f(x_n) = -\gamma_n \nabla^2 f(x_n) < 0$$



算法总结

牛顿法

$$x_{n+1} = x_n - \frac{\ell'(x_n)}{\ell''(x_n)}$$

梯度下降法

$$x_{n+1} = x_n - \gamma_n \ell'(x_n)$$

- 牛顿法和梯度下降法是求解最优化问题的常见的两种算法。
- 前者使用割线逐渐逼近最优解，后者使得目标函数逐渐下降。
- 牛顿法的收敛速度快，但是需要二阶导数信息。
- 梯度下降法计算速度快，但是需要人工确认步长参数。
- BB步实际上借助了二阶导数信息(x_{n-1} 和 x_n 的梯度的差)。

极大似然法 (Maximum Likelihood)

$$\operatorname{argmin}_{\widehat{\mathbf{w}}} - \sum_{i=1}^n (y_i \ln p_1 + (1 - y_i) \ln p_0)$$

利用对数几率函数的性质 $p_1' = p_1(1 - p_1)$ ，可以得到目标函数的导数。

$$\begin{aligned} (\ln p(y_i | \hat{x}_i; \widehat{\mathbf{w}}))' &= y_i \frac{1}{p_1} p_1(1 - p_1) \hat{x}_i + (1 - y_i) \frac{1}{1 - p_1} (-p_1(1 - p_1)) \hat{x}_i \\ &= \hat{x}_i(y_i - p_1) \end{aligned}$$

$$(\ln p(y_i | \hat{x}_i; \widehat{\mathbf{w}}))'' = (\hat{x}_i(y_i - p_1))' = -\hat{x}_i p_1(1 - p_1) \hat{x}_i^T$$

监督学习 (Supervised Learning)

对于给定训练集 $\{(x_i, y_i)\}_{i=1}^n$ ，我们希望找到数据 x_i 与其对应的标签 y_i 之间的函数关系 f ，即

$$y_i \approx f(x_i)$$

从概率角度来说，我们希望找到恰当的分布函数 p ，使得在给定 x_i 时 y_i 发生的概率越大越好，即

$$p(y_i|x_i) \approx 1$$

两种方法

最小二乘法:

$$\operatorname{argmin}_f \frac{1}{n} \sum_i (y_i - f(x_i))^2$$

极大似然法:

$$\operatorname{argmax}_p \ln \prod_{i=1}^m p(y_i | x_i)$$

无论是 f 还是 p 都需要合适的假设。

最小二乘法

最小二乘法:

$$\operatorname{argmin}_f \frac{1}{n} \sum_i (y_i - f(x_i))^2$$

- 线性回归: $f(x_i) = \mathbf{w}^T \mathbf{x}_i + b = \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i$
- 逻辑回归: $f(x_i) = \frac{1}{1+e^{-(\mathbf{w}^T \mathbf{x}_i + b)}} = \frac{1}{1+e^{-\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i}}$

极大似然法

极大似然法:

$$\operatorname{argmax}_p \sum_{i=1}^n \ln p(\mathbf{y}_i | \mathbf{x}_i)$$

- 线性回归: $\ln p(\mathbf{y}_i | \mathbf{x}_i) = \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} (\mathbf{y}_i - \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i)^2$
- 逻辑回归: $\ln p(\mathbf{y}_i | \mathbf{x}_i) = -\mathbf{y}_i \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i + \ln (1 + e^{\hat{\mathbf{w}}^T \hat{\mathbf{x}}_i})$