线性模型
Linear Model

任课教师：孔雨秋

yqkong@dlut.edu.cn

目 录 (CONTENT)

01

PART ONE

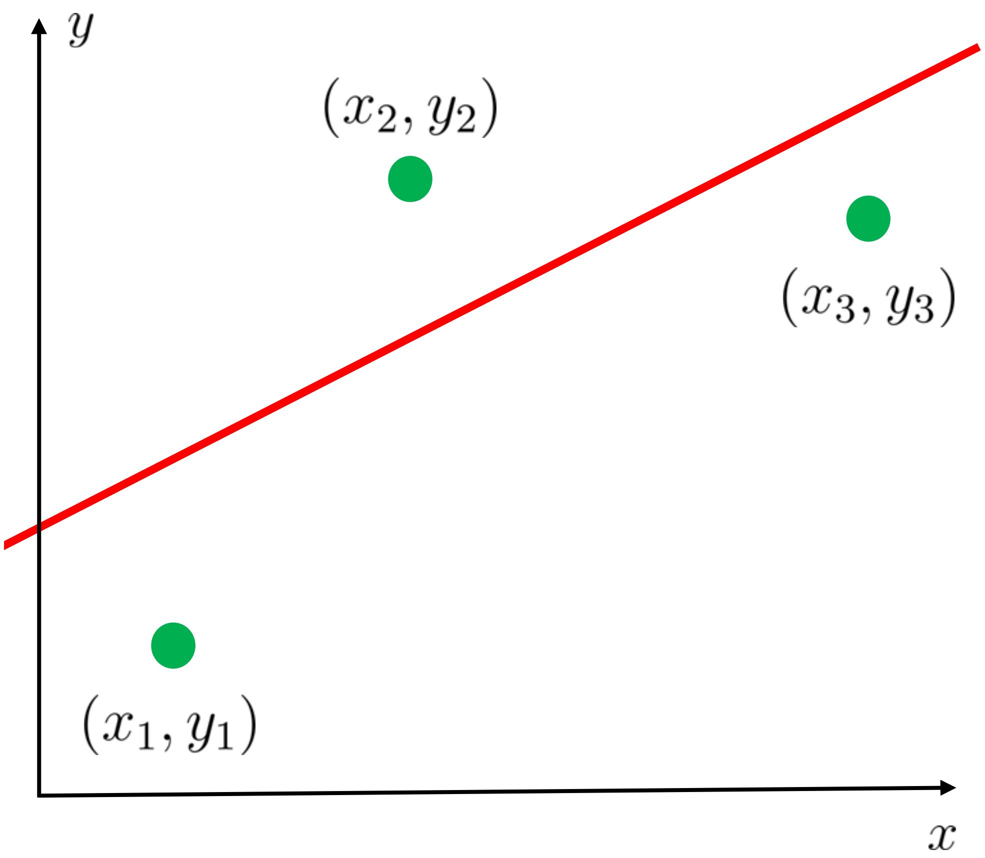
线性回归

Linear Regression

线性回归

■ 一元线性回归

数据 $D = \{(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_N, y_N)\}$



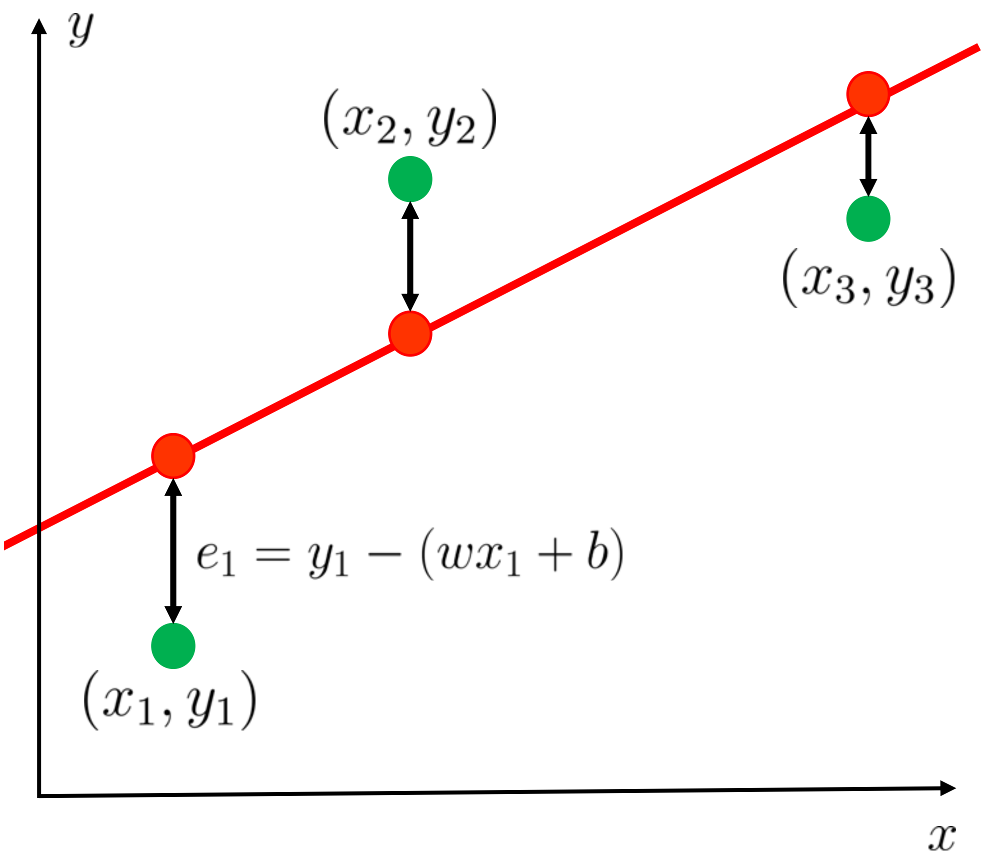
□ 一元线性函数

$$f(x) = wx + b$$

线性回归

■ 一元线性回归

数据 $D = \{(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_N, y_N)\}$



□ 一元线性函数

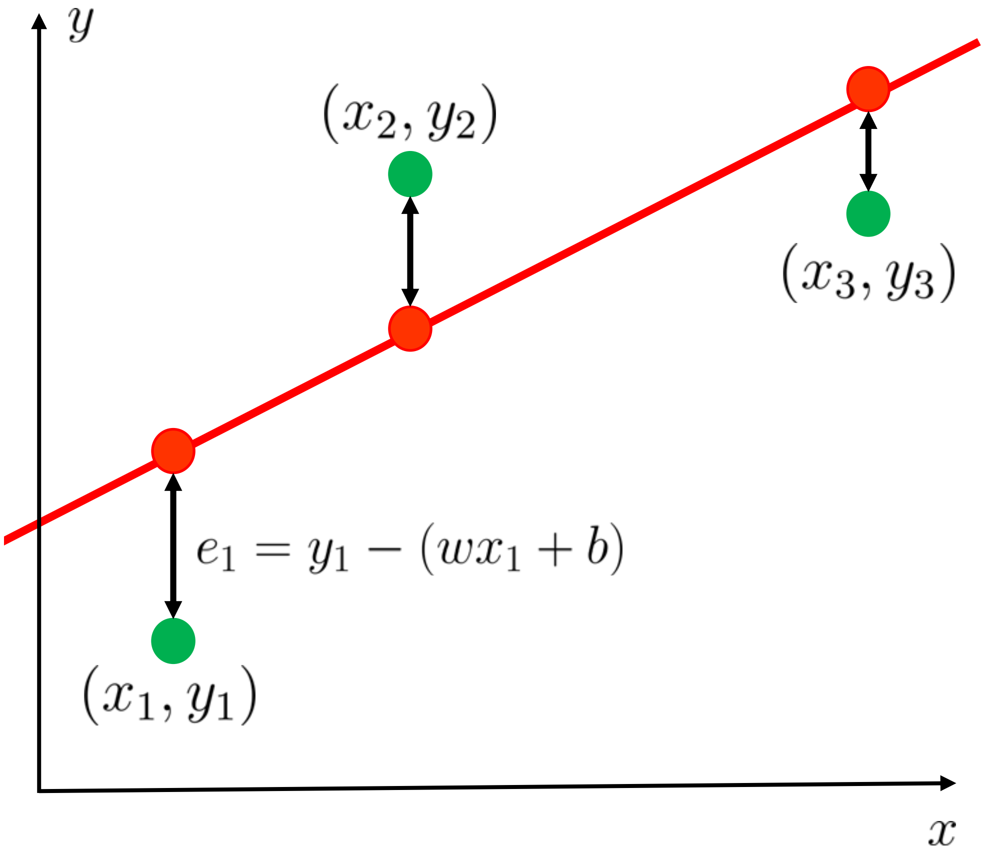
$$f(x) = wx + b$$

□ 均方和误差

$$\sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N} (y_n - wx_n - b)^2$$

■ 一元线性回归

数据 $D = \{(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_N, y_N)\}$



□ 一元线性函数

$$f(x) = wx + b$$

□ 均方和误差

$$\sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N} (y_n - wx_n - b)^2$$

□ 最小二乘法求解

$$(w^*, b^*) = \arg\min_{w,b} \sum_{i=1}^{N} (y_i - wx_i - b)^2$$

线性回归

一元线性回归—最小二乘法

- 一元线性回归函数

$$f(x) = wx + b \Rightarrow wx_i + b \approx y_i$$

- 最小二乘法

$$\left(w^*, b^*\right) = \arg\min_{w,b} \sum_{i=1}^{N} \left(y_i - wx_i - b\right)^2$$

- 最小化均方误差

$$E(w,b) = \sum_{i=1}^{N} \left(y_i - wx_i - b\right)^2$$

线性回归

■ 一元线性回归—最小二乘法

- 最小化均方误差

$$E(w,b)=\sum_{i=1}^{N}\left(y_i-wx_i-b\right)^2$$

- 分别对w和b求导，并令导数等于0

$$\frac{\partial E(w,b)}{\partial w}=2\left(w\sum_{i=1}^{N}x_i^2-\sum_{i=1}^{N}(y_i-b)x_i\right)=0$$

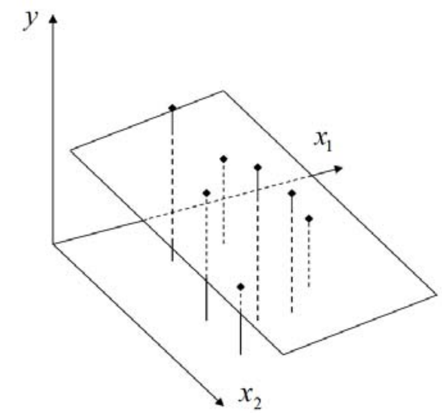
$$\frac{\partial E(w,b)}{\partial b}=2\left(mb-\sum_{i=1}^{m}(y_i-wx_i)\right)=0$$

➡ 有闭形式解 w^*,b^*

线性回归

■ 多元线性回归

$$
\begin{aligned}
y_1 &\approx x_{11}w_1 + x_{12}w_2 + \cdots + x_{1D}w_D + b \\
y_2 &\approx x_{21}w_1 + x_{22}w_2 + \cdots + x_{2D}w_D + b \\
&\cdots \\
y_N &\approx x_{N1}w_1 + x_{N2}w_2 + \cdots + x_{ND}w_D + b
\end{aligned}
$$



$$
\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} & 1 \\ x_{21} & x_{22} & \cdots & x_{2D} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} & 1 \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \\ b \end{bmatrix}}_{\mathbf{w}}
$$

$$\boxed{\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2}$$

多元最小二乘法

线性回归

■ 多元线性回归—最小二乘法

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

□ 误差

$$E(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$

□ 最小二乘法求解

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} = 0$$

□ 闭形式解

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

前提是 $\mathbf{X}^T\mathbf{X}$ 可逆

线性回归

■ 多元线性回归—最小二乘法

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

□ 误差

$$E(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$

□ 最小二乘法求解

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} = 0$$

□ 闭形式解

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \Longrightarrow \mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

正则化、岭回归

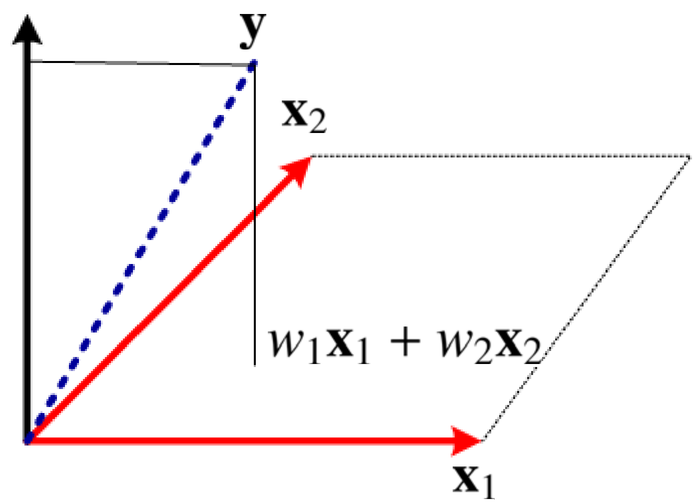
线性回归

■ 最小二乘的几何解释—正交投影

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} & 1 \\ x_{21} & x_{22} & \cdots & x_{2D} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \\ b \end{bmatrix}$$

$$\mathbf{y} = \underline{w_1\mathbf{x}_1 + w_2\mathbf{x}_2} + \cdots + w_D\mathbf{x}_D$$

Y向数据X所张成的子空间的正交投影



■ 最小二乘的概率解释—高斯噪声假设

$$\mathbf{e}=\mathbf{y}-\mathbf{X}\mathbf{w} \quad \Rightarrow \quad \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}=\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}-\begin{bmatrix} x_{1,1}w_1+x_{1,2}w_2+\ldots+x_{1,D}w_D \\ \vdots \\ x_{N,1}w_1+x_{N,2}w_2+\ldots+x_{N,D}w_D \end{bmatrix}$$

- □ 零均值等方差高斯噪声假设

$$e_n \sim \mathcal{N}\left(0,\sigma^2\right) \Rightarrow p\left(e_n\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{e_n^2}{2\sigma^2}\right)$$

- □ 最大似然估计

$$p\left(e_1,e_2,\ldots,e_N\right)=\prod_{n=1}^{N} p\left(e_n\right)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\sum_{n=1}^{N} e_n^2 \Big/ 2\sigma^2\right)$$

- □ 概率解释—假设误差服从零均值等方差的高斯噪声假设

$$\max_{\mathbf{w}} p\left(e_1,e_2,\ldots,e_N;\mathbf{w}\right) \Leftrightarrow \min_{\mathbf{w}} \sum_{n=1}^{N} e_n^2 \Leftrightarrow \min_{\mathbf{w}} \left\|\mathbf{y}-\mathbf{X}\mathbf{w}\right\|_2^2$$

线性回归

- 对数线性回归

$$y \approx \exp(\mathbf{w}^T \mathbf{x} + b)$$

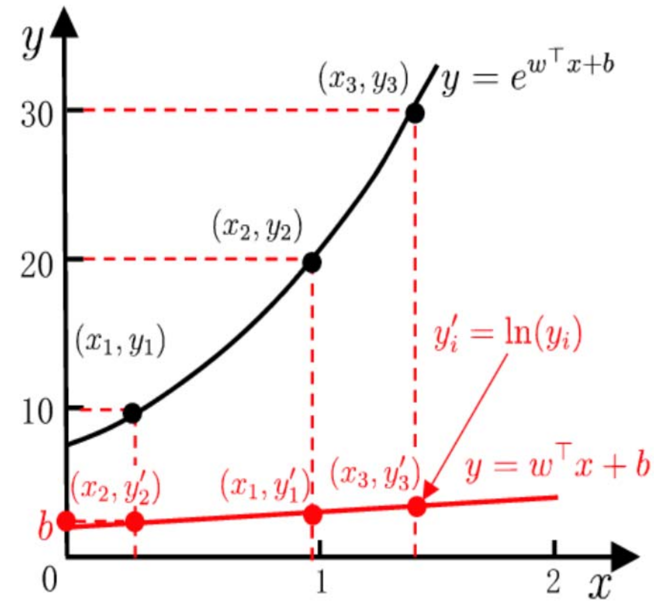
$$\Updownarrow$$

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

$$\Updownarrow$$

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

形式上仍是线性回归，实质上已经是求取输入空间到输出空间的非线性函数映射



- 广义线性模型

$$y = g^{-1}(\mathbf{w}^T \mathbf{x}) \Leftrightarrow y' = g(y) = \mathbf{w}^T \mathbf{x}$$

单调可微函数

02

PART TWO

逻辑斯蒂回归

Logistic Regression

逻辑斯蒂回归

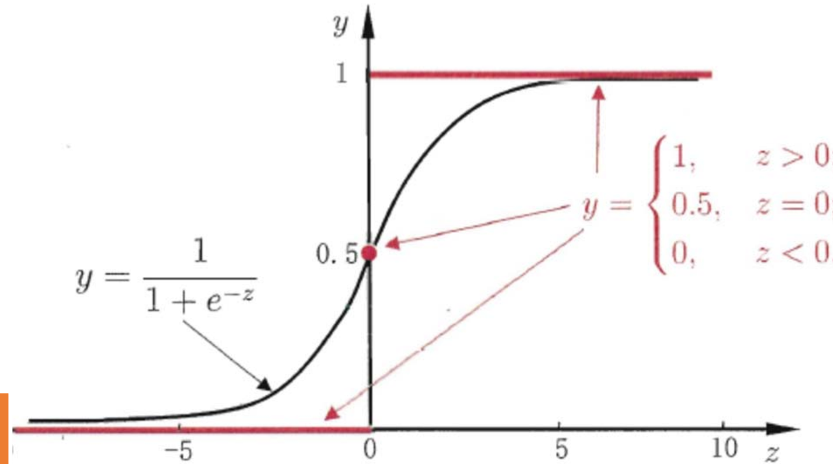
- 能否用线性回归函数做分类?

$$z = \mathbf{w}^T \mathbf{x}$$

□ 分类函数：单位越界函数

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

不连续



□ 分类函数：对数几率函数(logistic function)

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

单调可微
任意阶可导

逻辑斯蒂回归

■ 逻辑斯蒂回归--广义线性回归特例

$$z = \mathbf{w}^T \mathbf{x}$$

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

$$\boxed{\ln \frac{y}{1-y} = \mathbf{w}^T \mathbf{x}} \quad \left(\frac{y}{1-y} = \frac{\frac{1}{1+e^{-z}}}{1-\frac{1}{1+e^{-z}}} = \frac{\frac{1}{1+e^{-z}}}{\frac{e^{-z}}{1+e^{-z}}} = e^{z}\right)$$

$$\ln \frac{y}{1-y} = \ln \frac{p(y=1 \mid \mathbf{x})}{p(y=0 \mid \mathbf{x})} = \mathbf{w}^T \mathbf{x}$$

逻辑斯蒂回归

大连理工大学 Dalian University of Technology

$$\ln \frac{y}{1-y} = \ln \frac{p(y=1\,|\,\mathbf{x})}{p(y=0\,|\,\mathbf{x})} = \mathbf{w}^T\mathbf{x}$$

- **概率解释**
  - 分类概率 $y = \dfrac{1}{1+\exp\left(-\mathbf{w}^T\mathbf{x}\right)}$

$$p_1\left(\mathbf{x};\mathbf{w}\right) = p\left(y=1\middle|\mathbf{x};\mathbf{w}\right) = \frac{\exp\left(\mathbf{w}^T\mathbf{x}\right)}{1+\exp\left(\mathbf{w}^T\mathbf{x}\right)}$$ 属于正样本的概率

$$p_0\left(\mathbf{x};\mathbf{w}\right) = p\left(y=0\middle|\mathbf{x};\mathbf{w}\right) = \frac{1}{1+\exp\left(\mathbf{w}^T\mathbf{x}\right)}$$ 属于负样本的概率

  - 贝叶斯决策理论

$$\ln \frac{p\left(y=1\middle|\mathbf{x};\mathbf{w}\right)}{p\left(y=0\middle|\mathbf{x};\mathbf{w}\right)} = \mathbf{w}^T\mathbf{x} \quad \Rightarrow \begin{cases} \mathbf{w}^T\mathbf{x} > 0 & \rightarrow \quad y=1 \\ \mathbf{w}^T\mathbf{x} < 0 & \rightarrow \quad y=0 \end{cases}$$

逻辑斯蒂回归

■ 极大似然推导

□ 伯努利分布

y_n	1	0
p_n	$p^1(\mathbf{x}_n;\mathbf{w})$	$p^0(\mathbf{x}_n;\mathbf{w})$

$$p^1(\mathbf{x}_n;\mathbf{w}) + p^0(\mathbf{x}_n;\mathbf{w}) = 1$$

□ 最大似然估计

$$p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N;\mathbf{w}) = \prod_{n=1}^{N}\left[p^1(\mathbf{x}_n;\mathbf{w})\right]^{y_n}\left[p^0(\mathbf{x}_n;\mathbf{w})\right]^{1-y_n}$$

$$\max_{\mathbf{w}} p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N;\mathbf{w}) \Leftrightarrow \min_{\mathbf{w}}\left[-\ln p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N;\mathbf{w})\right]$$

逻辑斯蒂回归

■ 极大似然推导

□ 目标函数

$$\begin{aligned} L(\mathbf{w}) &= -\ln p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N; \mathbf{w}) \\ &= -\sum_{n=1}^{N}\left[y_n \ln p^1(\mathbf{x}_n; \mathbf{w}) + (1 - y_n) \ln p^0(\mathbf{x}_n; \mathbf{w}) \right] \\ &= \sum_{n=1}^{N}\left[-y_n \mathbf{w}^T \mathbf{x}_n + \ln\left(1 + \exp\left(\mathbf{w}^T \mathbf{x}_n\right)\right) \right] \end{aligned}$$

极大似然推导

优化问题

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} L(\mathbf{w})$$

$$L(\mathbf{w}) = \sum_{n=1}^{N}\left[-y_n\mathbf{w}^T\mathbf{x}_n + \log\left(1+\exp\left(\mathbf{w}^T\mathbf{x}_n\right)\right)\right]$$

迭代求解

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}$$

梯度下降法
Gradient Descent Method

$$\mathbf{w} \leftarrow \mathbf{w} - \left(\frac{\partial^2 L(\mathbf{w})}{\partial\mathbf{w}\partial\mathbf{w}^T}\right)^{-1}\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}$$

牛顿法
Newton Method

03 PART THREE

正则化
Regularization

正则化

- 一般的目标函数包含两部分：
  - 数据项 (Data Term)：回归/分类的目标，如误差尽可能小或分类尽可能准确等。
  - 正则化项 (Regularization Term)：对参数空间的限制/对解额外属性的追求（比如稀疏）。

$$\underbrace{O(\mathbf{x}) = D(\mathbf{x})}_{\text{数据项}} + \lambda \underbrace{R(\mathbf{x})}_{\text{正则化项}}$$

$$D(\mathbf{x}) \geq 0, R(\mathbf{x}) \geq 0$$

正则化

- 岭回归(Ridge Regression)

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$$

$$Data\ Term: \frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2$$

$$Regularization\ Term: \|\mathbf{x}\|_2^2$$

$$trade-off\ : \lambda$$

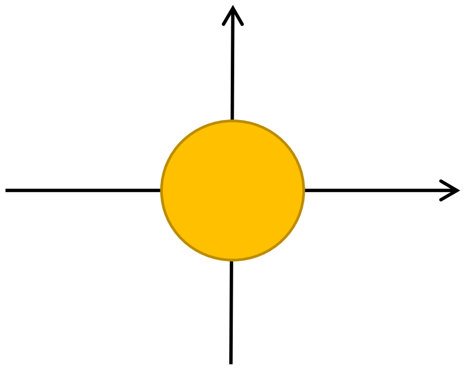
- 正则化是对解空 间的一种限制

$$\min_{\mathbf{x}} D(\mathbf{x}) + \lambda R(\mathbf{x})$$

$$\Leftrightarrow$$

$$\min_{\mathbf{x}} D(\mathbf{x})$$

$$s.t.\ R(\mathbf{x}) \le M$$

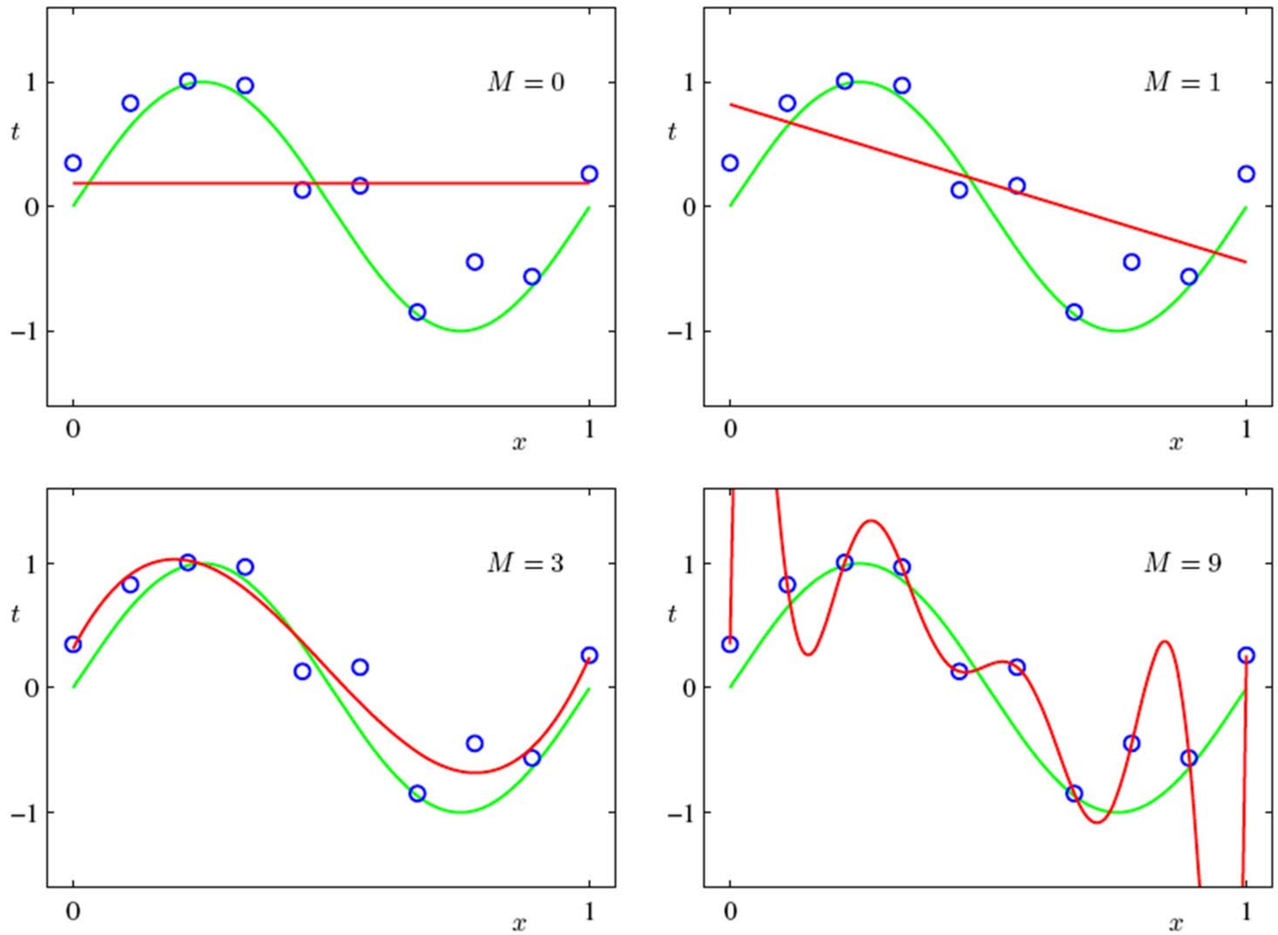


正则化

- 最小二乘模型 $\min_x \frac{1}{2}\|\mathbf{y}-\mathbf{Xw}\|_2^2$

$$y = w_0 + w_1 x + \ldots + w_M x^M$$

M：超参数



正则化

大连理工大学 Dalian University Of Technology

■ 最小二乘模型 $\min_x \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2$

$$y = w_0 + w_1 x + \ldots + w_M x^M$$

M：超参数

Table 1.1 Table of the coefficients $\mathbf{w}^\star$ for polynomials of various order. Observe how the typical magnitude of the coefficients increases dramatically as the order of the polynomial increases.

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

正则化

■ 最小二乘模型 $\min_x \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2$

$$y = w_0 + w_1 x + \ldots + w_M x^M$$

M：超参数

❑ 增加样本数量，防止过拟合

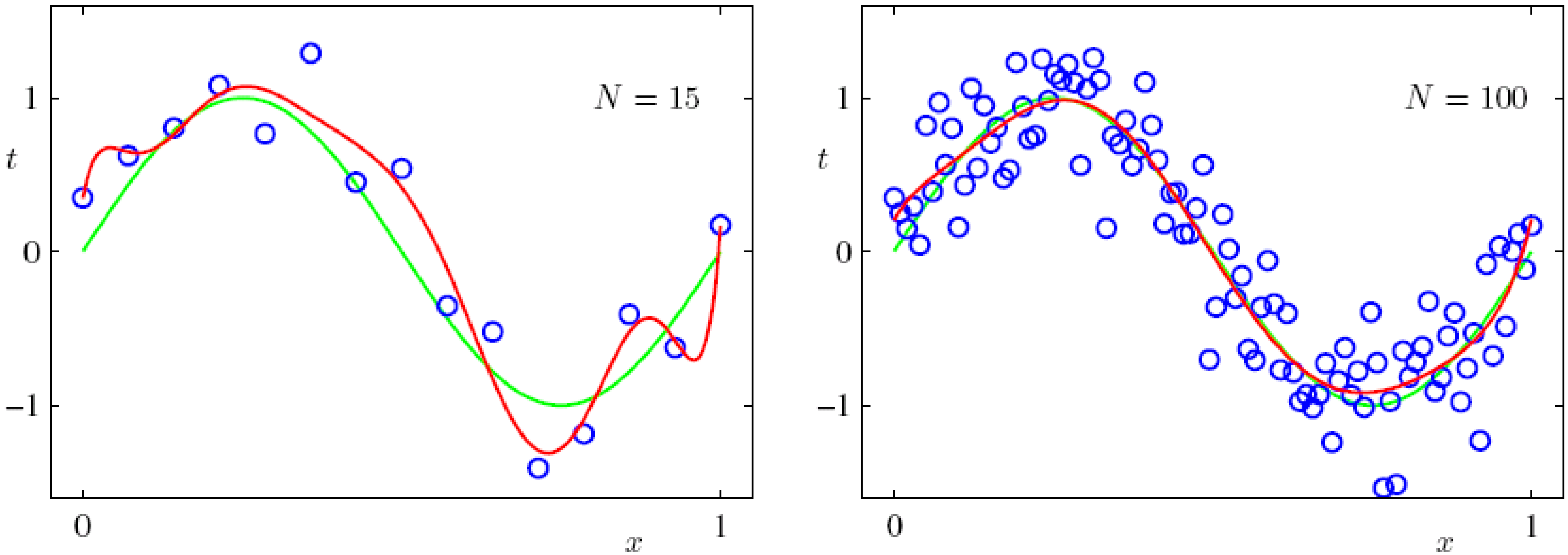


Figure 1.6 Plots of the solutions obtained by minimizing the sum-of-squares error function using the $M = 9$ polynomial for $N = 15$ data points (left plot) and $N = 100$ data points (right plot). We see that increasing the size of the data set reduces the over-fitting problem.

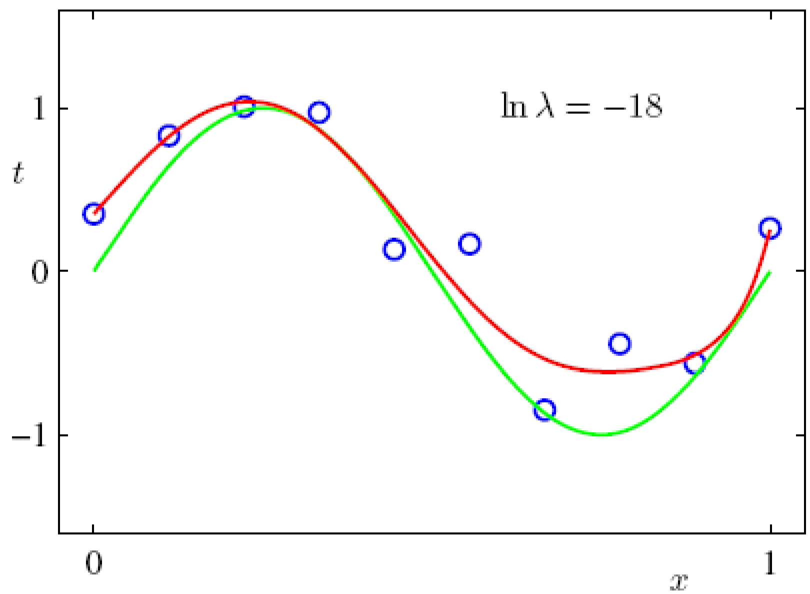
正则化

- 岭回归 $\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + \frac{1}{2}\lambda\|\mathbf{x}\|_2^2$

$$y = w_0 + w_1 x + \ldots + w_M x^M$$

M：超参数

- 正则化，控制模型复杂度



	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Figure 1.7 Plots of $M = 9$ polynomials fitted to the da
function (1.4) for two values of the regularization param
case of no regularizer, i.e., $\lambda = 0$, corresponding to $\ln\lambda =$

正则化

■ 正则化项可以看作参数的先验 (从贝叶斯估计角度)

观测数据：$\mathbf{X}$

模型参数：$\mathbf{w}$

$$p(\mathbf{w}\,|\,\mathbf{X}) = \frac{p(\mathbf{w},\mathbf{X})}{p(\mathbf{X})} = \frac{p(\mathbf{X}\,|\,\mathbf{w})\,p(\mathbf{w})}{p(\mathbf{X})} \propto p(\mathbf{X}\,|\,\mathbf{w})\,p(\mathbf{w})$$

❑ 最大后验概率(Maximum a Posteriori, MAP)

$$\mathbf{w}^{*} = \arg\max_{\mathbf{w}} p(\mathbf{w}\,|\,\mathbf{X}) = \arg\max_{\mathbf{w}} p(\mathbf{X}\,|\,\mathbf{w})\,\boxed{p(\mathbf{w})}$$

❑ 最大似然估计(Maximum Likelihood Estimation, MLE)

$$\mathbf{w}^{*} = \arg\max_{\mathbf{w}} p(\mathbf{X}\,|\,\mathbf{w})$$

参数的先验概率
假设为常数，服从均匀分布

正则化

■ 最大似然估计(MLE) v.s. 最小二乘法

$$\mathbf{y} = \mathbf{X}\mathbf{w}$$

误差/残差 : $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{w}$

假设 $\mathbf{e} = [e_1, e_2, \ldots, e_N]$服从 $i.i.d$ 正态分布$N(0, \sigma^2)$

$\mathbf{w} = [w_1, w_2, \ldots, w_d]$服从均匀分布先验.

$$p(\mathbf{w} \mid \mathbf{X}) \propto p(\mathbf{X} \mid \mathbf{w}) \cancel{p(\mathbf{w})}$$

$$= p(\mathbf{e}) = \prod_{n=1}^{N} p(e_n)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{e_n^2}{2\sigma^2}\right]$$

$$= \left\{\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{\sigma^2}\right)\right\} \exp\left(-\frac{1}{2}\|\mathbf{e}\|_2^2\right)$$

正则化

- 最大后验(MAP) v.s. 岭回归

$\mathbf{y}=\mathbf{X}\mathbf{w}$ 误差/残差 : $\mathbf{e}=\mathbf{y}-\mathbf{X}\mathbf{w}$

假设 $\mathbf{e}=\left[e_1,e_2,\ldots,e_N\right]$服从 $i.i.d$ 高斯正态分布 $N\left(0,\sigma^2\right)$

假设$\mathbf{w}=\left[w_1,w_2,\ldots,w_D\right]$也服从 $i.i.d$ 高斯正态分布 $N\left(0,\sigma_1^2\right)$

$$p(\mathbf{w}\mid\mathbf{X})\propto p(\mathbf{X}\mid\mathbf{w})\,p(\mathbf{w})=p(\mathbf{e})\,p(\mathbf{w})=\left[\prod_{n=1}^{N}p(e_n)\right]\left[\prod_{d=1}^{D}p(w_d)\right]$$

$$=\left\{\prod_{n=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{e_n^2}{2\sigma^2}\right]\right\}\left\{\prod_{d=1}^{D}\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{w_d^2}{2\sigma_1^2}\right]\right\}$$

$$\propto\left\{\prod_{n=1}^{N}\exp\left[-\frac{e_n^2}{2\sigma^2}\right]\right\}\left\{\prod_{d=1}^{D}\exp\left[-\frac{w_d^2}{2\sigma_1^2}\right]\right\}$$

$$\boxed{\max_{\mathbf{x}}p(\mathbf{w}\mid\mathbf{X})\Leftrightarrow\min_{\mathbf{x}}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\frac{\lambda}{2}\|\mathbf{w}\|_2^2}$$

$$=\exp\left(-\frac{\|\mathbf{e}\|_2^2}{2\sigma^2}\right)\exp\left(-\frac{\|\mathbf{w}\|_2^2}{2\sigma_1^2}\right)=\exp\left(\frac{1}{\sigma^2}\right)\exp\left[-\frac{\|\mathbf{e}\|_2^2}{2}-\frac{\sigma^2}{2\sigma_1^2}\|\mathbf{w}\|_2^2\right]$$

$$\propto\exp\left[-\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2-\frac{\lambda}{2}\|\mathbf{w}\|_2^2\right]\qquad\left(ps:\lambda=\frac{\sigma^2}{\sigma_1^2}\right)$$

正则化

■ **最小二乘法(OLS) v.s. 岭回归(RR)**

1.目标函数(*Objective Function*):

$$OLS: \min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{Xw}\|_2^2$$

$$RR: \quad \min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{Xw}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

2.解(*Solution*):

$$OLS: \mathbf{x} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

$$RR: \quad \mathbf{x} = \left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

3.概率(*Probability*):

$OLS: MLE$

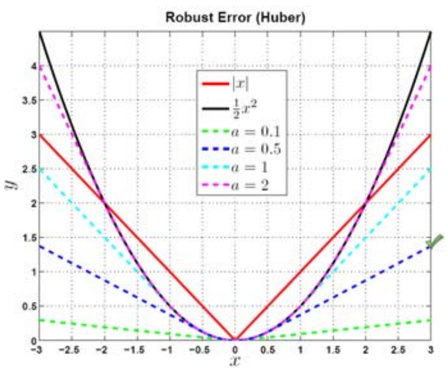
$RR: \quad MAP$

- 吴恩达《机器学习》-正则化 https://www.bilibili.com/video/av55276229
- L1&L2正则化详解 https://www.bilibili.com/video/av77106463?from=search&seid=4369320229005019988
- 什么是L1 L2正则化? https://www.bilibili.com/video/av16009446?from=search&seid=4369320229005019988

正则化

$$\mathbf{e}=\mathbf{y}-\mathbf{X}\mathbf{w} \quad \Rightarrow \quad \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} x_{11}w_1 + x_{12}w_2 + \ldots + x_{1D}w_D \\ \vdots \\ x_{N1}w_1 + x_{N2}w_2 + \ldots + x_{ND}w_D \end{bmatrix}$$

$$\rho\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)+\lambda R\left(\mathbf{w}\right)$$



$$\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\mathbf{w}\right\|_2^2$$

$$\left\|\mathbf{y}-\mathbf{X}\mathbf{w}\right\|_1$$

Huber

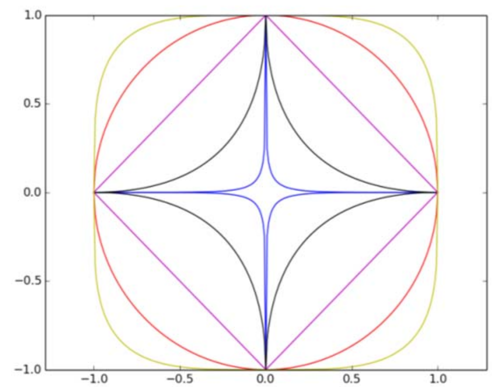
$\vdots$

$$\frac{1}{2}\left\|\mathbf{w}\right\|_2^2$$

$$\left\|\mathbf{w}\right\|_1$$

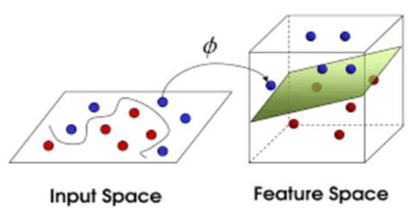
$$\left\|\mathbf{w}\right\|_p := \left(\sum_{i=1}^{n}\left|x_i\right|^p\right)^{1/p}$$

$\vdots$



$$\mathbf{y}\rightarrow\phi\left(\mathbf{y}\right)$$

$$\mathbf{X}\rightarrow\phi\left(\mathbf{X}\right)$$



04

PART FOUR

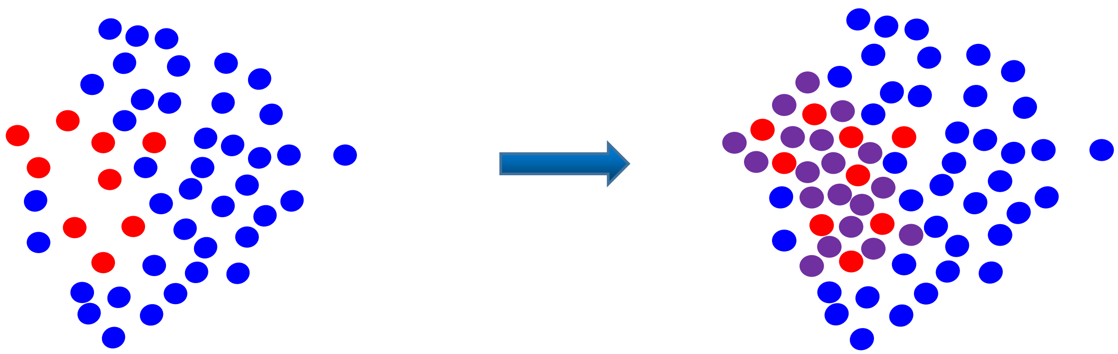
类别不平衡问题

Class-Imbalance

类别不平衡问题

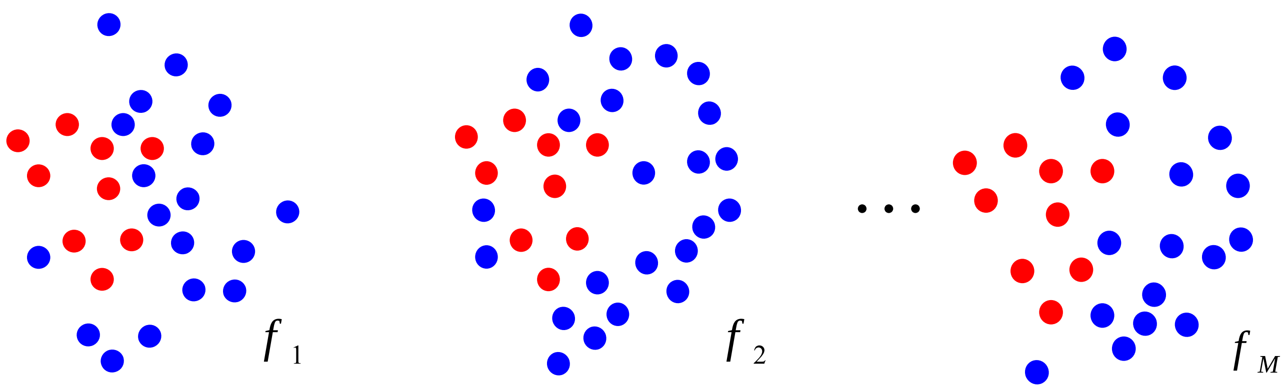
■ 类别不平衡

❑ **过采样(over-sampling)**



- ✓ 样本复制
- ✓ 样本插值
- ✓ 样本生成 **(GAN)**

❑ **降采样(under-sampling)**



- ● 集成学习

$$f = \frac{1}{M}\sum_{m=1}^{M} f_M$$

- ✓ EasyEnsemble[2]
- ✓ BalanceCascade[2]

类别不平衡问题

大连理工大学 Dalian University of Technology

加权损失函数

- 一般损失函数

$$L(\mathbf{X};\mathbf{w})=\frac{1}{N}\sum_{n=1}^{N}L(\mathbf{x}_n,y_n;\mathbf{w})=\sum_{n=1}^{N}\frac{1}{N}L(\mathbf{x}_n,y_n;\mathbf{w})$$

$$=\sum_{n=1}^{N^+}\frac{1}{N}L(\mathbf{x}_n,y_n;\mathbf{w})+\sum_{n=1}^{N^-}\frac{1}{N}L(\mathbf{x}_n,y_n;\mathbf{w})$$

- 加权损失函数

$$L(\mathbf{X};\mathbf{w})=\sum_{n=1}^{N}\rho_n L(\mathbf{x}_n,y_n;\mathbf{w})$$

$$=\sum_{n=1}^{N^+}\rho_n^+ L(\mathbf{x}_n,y_n;\mathbf{w})+\sum_{n=1}^{N^-}\rho_n^- L(\mathbf{x}_n,y_n;\mathbf{w})$$

$$\rho^+=\frac{N^-}{N},\rho^-=\frac{N^+}{N}\quad\Rightarrow\quad\frac{\rho^+}{\rho^-}=\frac{N^-}{N^+}$$

类别不平衡问题

大连理工大学 Dalian University of Technology

■ Focal Loss

❑ 交叉熵损失

$$\mathrm{CE}(p_t) = -\log(p_t) \qquad p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise} \end{cases}$$

$$L = \frac{1}{N}\sum_{n=1}^{N} CE(p_n^t) = -\frac{1}{N}\sum_{n=1}^{N} y_n \log(p_n) + (1-y_n)\log(1-p_n)$$

❑ Focal Loss[1] ⇐ 解决目标检测问题中训练集正负样本极度不平衡情况

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t)$$

✓ 自适应样本加权
✓ 容易分类样本权重降低
✓ 促进分类器学习更关注困难样本