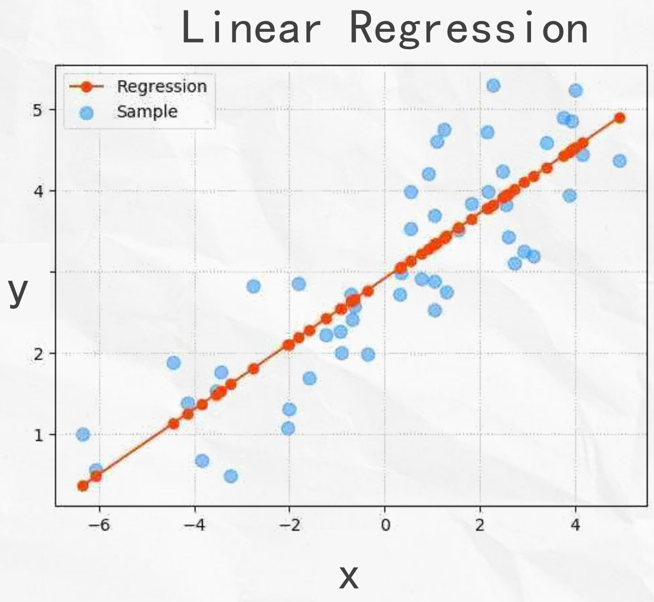
# 机器学习

## 线性回归

涂文婷
tu.wenting@mail.shufe.edu.cn

# 线性回归

- 定义




$$f_{\text{瓜甜}}(\boldsymbol{x}) = 0.2 \cdot x_{\text{色泽}} + 0.5 \cdot x_{\text{根蒂}} + 0.3 \cdot x_{\text{敲声}} + 1$$

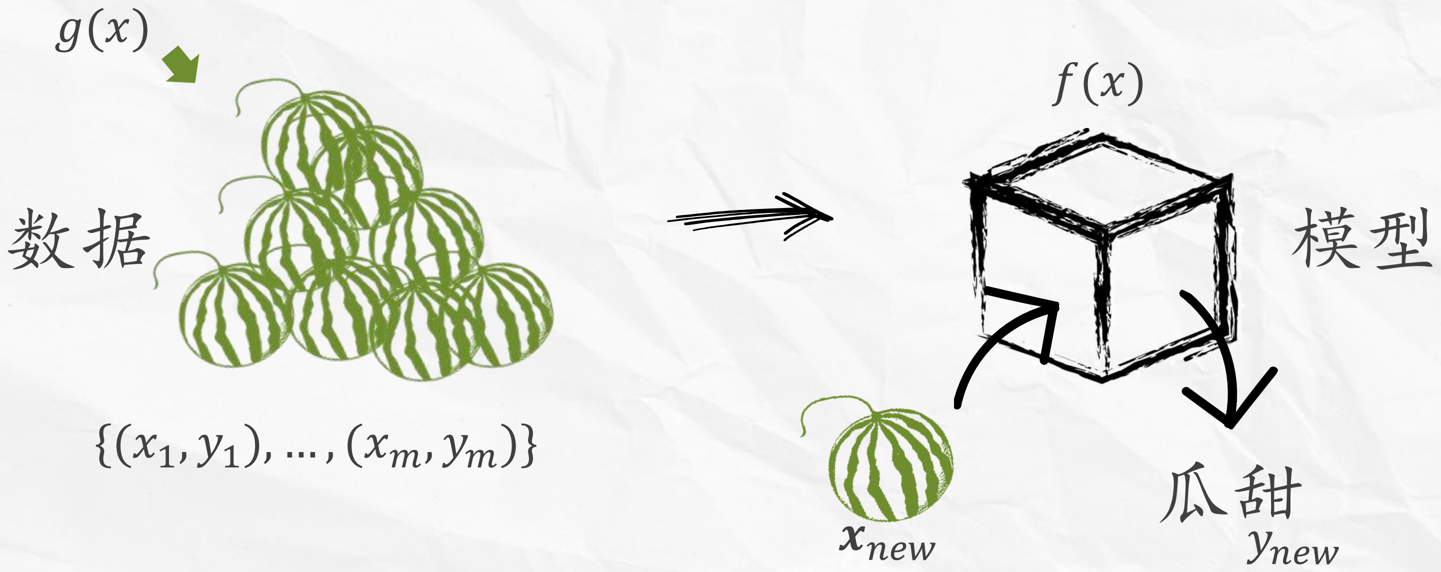
# 线性回归

- 最小二乘法

设定模型的形式：$f(\boldsymbol{x}) = w_1x_1 + w_2x_2 + \cdots + w_dx_d + b$

设定误差的形式：$\ell(f(\boldsymbol{x}_i), y_i) = (f(\boldsymbol{x}_i) - y_i)^2$

利用最小化训练误差求解模型参数：$\arg\min\limits_{(w,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2 = \arg\min\limits_{(w,b)} \sum_{i=1}^{m} (y_i - wx_i - b)^2$



# 线性回归

- 最小二乘法

设定模型的形式：$f(\boldsymbol{x}) = w_1x_1 + w_2x_2 + \cdots + w_dx_d + b$

设定误差的形式：$\ell(f(\boldsymbol{x}_i), y_i) = (f(x_i) - y_i)^2$

利用最小化训练误差求解模型参数：

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

解析解：

$$w = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2}, \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$b = \frac{1}{m}\sum_{i=1}^{m} (y_i - wx_i)$$

# 线性回归

- 最小二乘法

$$\mathbf{X}=\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} & 1 \\ x_{21} & x_{22} & \dots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{md} & 1 \end{pmatrix}=\begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} & 1 \\ \boldsymbol{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} & 1 \end{pmatrix}$$

$$\boldsymbol{y}=(y_1;y_2;\dots;y_m)$$

$$\hat{\boldsymbol{w}}^*=\arg\min_{\hat{\boldsymbol{w}}}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{w}})$$

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

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}}=2\mathbf{X}^{\mathrm{T}}(\mathbf{X}\hat{\boldsymbol{w}}-\boldsymbol{y})$$

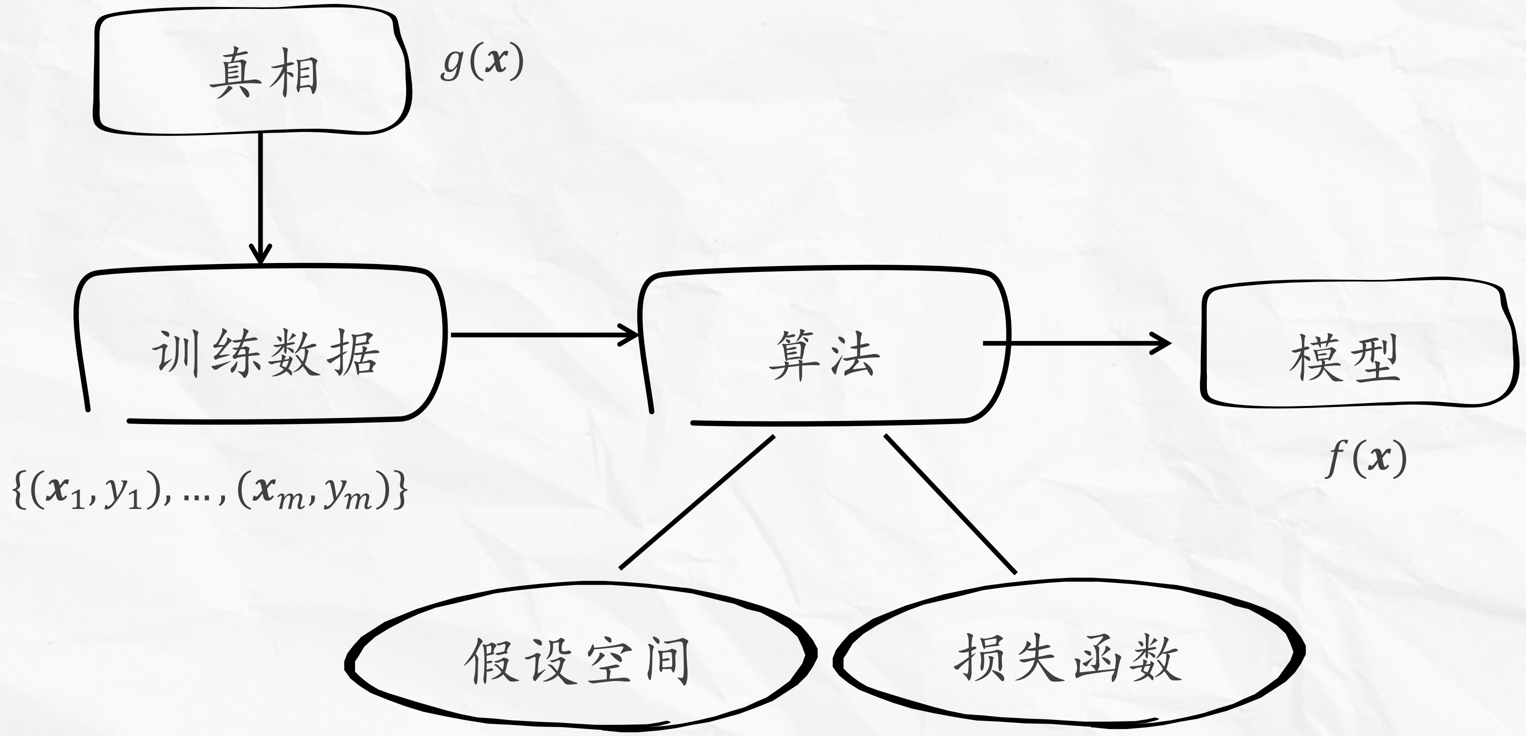
$$\hat{\boldsymbol{w}}^*=\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

$$f\left(\hat{\boldsymbol{x}}_i\right)=\hat{\boldsymbol{x}}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y},\ \hat{\boldsymbol{x}}_i=(\boldsymbol{x}_i,1)$$

现实任务中 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 往往不是满秩矩阵，此时此时可解出多个$\hat{\boldsymbol{w}}$，都能使均方误差最小化. 选择哪一个解作为输出将由学习算法的归纳偏好决定，常见的做法是引入正则化（regularization）项.

# 线性回归

- 机器学习框架




e.g. $f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$

$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b$

e.g. $\ell(f(\boldsymbol{x}_i), y_i) = (f(\boldsymbol{x}_i) - y_i)^2$

# 线性回归

- 正则化技术
  - 一般形式

$$\min_f \Omega(f) + \sum_{i=1}^{m} \ell(f(\boldsymbol{x}_i), y_i)$$

结构风险用于描述模型的某些性质

经验风险用于描述模型与训练数据的契合程度

# 线性回归

- 正则化技术

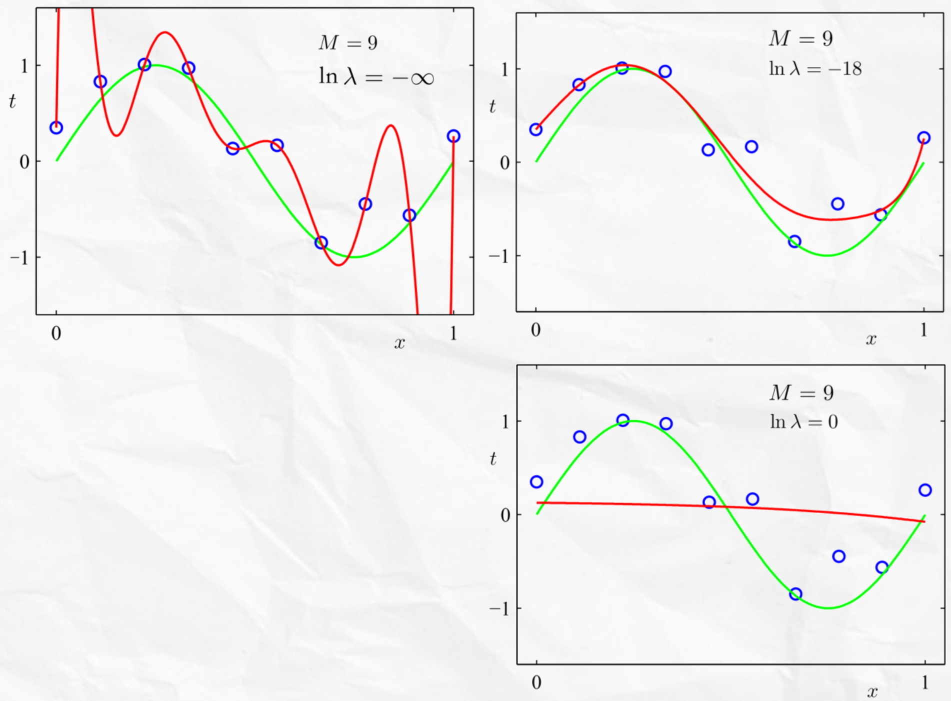
> Ridge 岭回归

$$\min_{\boldsymbol{w},b} \| \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}} \|_2^2 + \lambda \| \boldsymbol{w} \|_2^2 \quad , \quad \| \boldsymbol{w} \|_2^2 = \sum_{j=1}^{d} w_j^{\ 2}$$

$$\min_{\boldsymbol{w},b} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} \left( \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}} \right) + \lambda \| \boldsymbol{w} \|_2^2$$

$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_M x^M$$

$$\min_{\boldsymbol{w},b} \| \boldsymbol{y} - \mathbf{X}\boldsymbol{w} \|_2^2 + \lambda \| \boldsymbol{w} \|_2^2$$



# 线性回归

- 正则化技术

> LASSO

$$\min_{\boldsymbol{w},b} \parallel \boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{w}} \parallel_2^2 + \lambda \parallel \boldsymbol{w} \parallel_1 \quad , \parallel \boldsymbol{w} \parallel_1 = \sum_{j=1}^{d} |w_j|$$

$$\min_{\boldsymbol{w},b} (\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}}\left(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{w}}\right) + \lambda \parallel \boldsymbol{w} \parallel_1$$

> ElasticNet

$$\min_{\boldsymbol{w},b} \parallel \boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{w}} \parallel_2^2 + \lambda_1 \parallel \boldsymbol{w} \parallel_1 + \lambda_2 \parallel \boldsymbol{w} \parallel_2^2$$

# 线性回归

- Lasso in Sklearn

  https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

- ElasticNet in Sklearn

  https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

# 扩展：梯度下降法求解的线性回归

- 梯度下降法

考虑无约束优化问题$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$
若能构造一个序列$\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots$
满足$\mathcal{L}\left(\boldsymbol{\theta}^{t+1}\right) < \mathcal{L}(\boldsymbol{\theta}^t)$, $t$=0,1,2 , …
则不断执行该过程即可收剑到局部极小点
根据泰勒展式有$\mathcal{L}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) \simeq \mathcal{L}(\boldsymbol{\theta}) + \Delta\boldsymbol{\theta}^{\mathrm{T}} \nabla \mathcal{L}(\boldsymbol{\theta})$
于是, 欲满足$\mathcal{L}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) < \mathcal{L}(\boldsymbol{\theta})$
可选择$\Delta\boldsymbol{\theta} = -\eta \nabla \mathcal{L}(\boldsymbol{\theta})$
其中步长$\eta$是一个小常数. 这就是梯度下降法

# 扩展：梯度下降法求解的线性回归

- 梯度下降法求解最小二乘法

> 批量梯度下降

$$w^{(t+1)} = w^{(t)} - \eta \nabla \mathcal{L}$$
$$w^{(t+1)} = w^{(t)} + \eta \sum_{i=1}^{m} \left(y_i - w^{(t)T} x_i\right) x_i$$

> 随机梯度下降

$$w^{(t+1)} = w^{(t)} - \eta \nabla \mathcal{L}_i$$
$$w^{(t+1)} = w^{(t)} + \eta \left(y_i - w^{(t)T} x_i\right) x_i$$

# 扩展：利用线性回归实现多项式回归

- 多项式回归

原始特征集

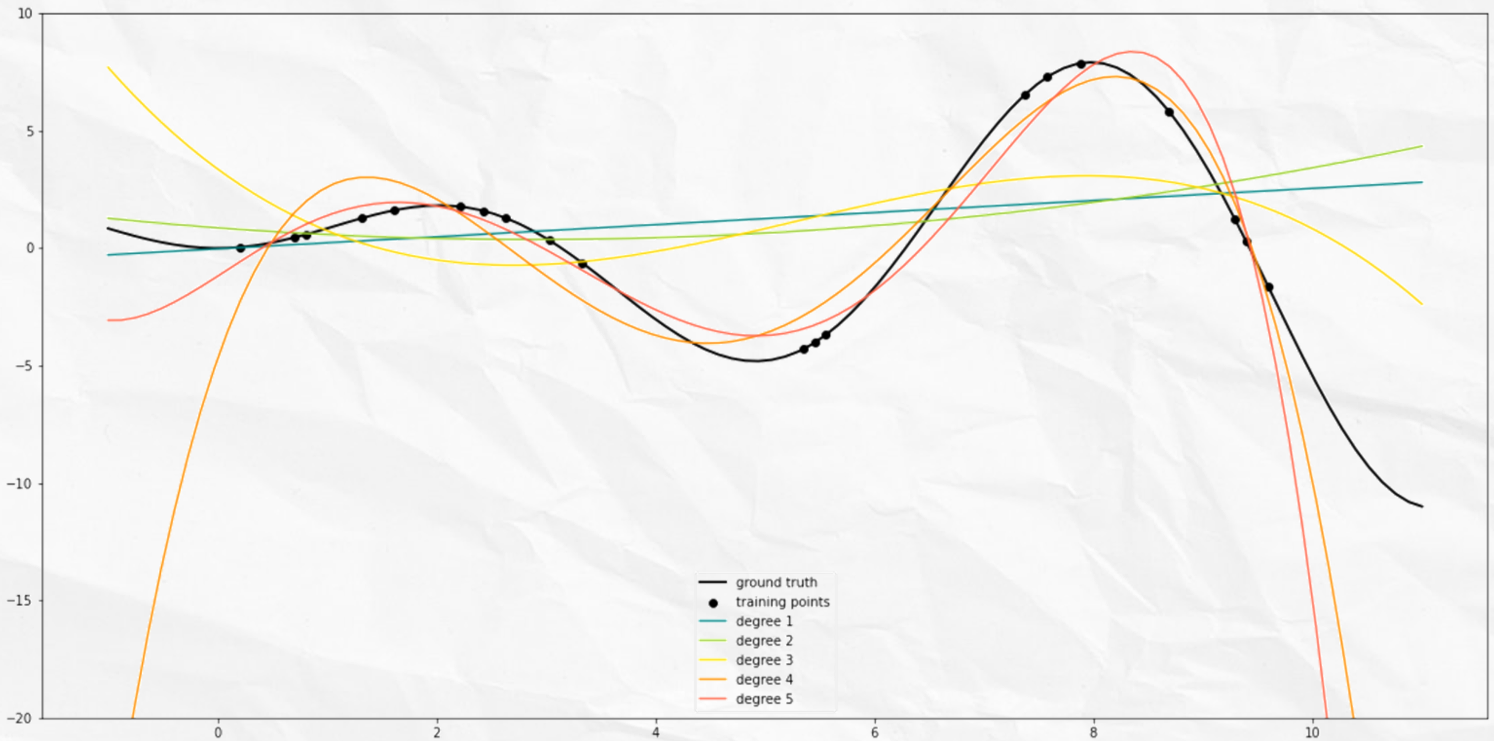
$(x_1, x_2)$

$(x_1, x_2, x_3)$

变换后特征集

$(1, x_1, x_2, x_1^2, x_1x_2, x_2^2)$

$(1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)$



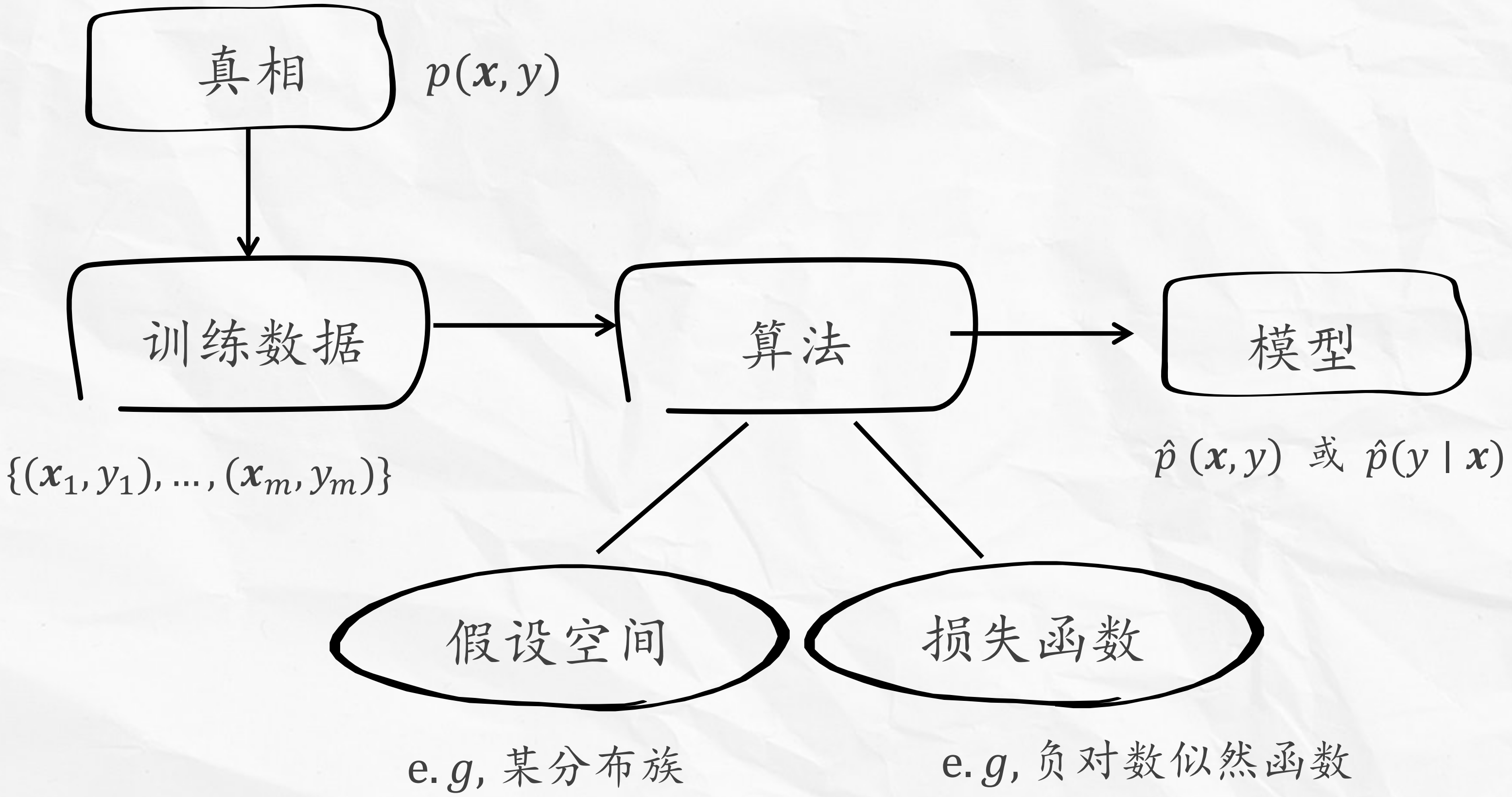
# 扩展：利用线性回归实现多项式回归

- PolynomialFeatures in Sklearn

  https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

# 扩展：概率论角度下的线性回归

- 机器学习框架 (概率论角度)



# 扩展：概率论角度下的线性回归

## ○ 概率论角度的最小二乘法

· 假设$p(y \mid \boldsymbol{x})$服从高斯分布，高斯分布的均值参数由线性函数$f(\boldsymbol{x}, \boldsymbol{w})$给出，其中$\boldsymbol{w}$为模型参数，方差记为$\beta^{-1}$

$$p(y \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}\left(y \mid f(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}\right),\ y = f(\boldsymbol{x}, \boldsymbol{w}) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \beta^{-1})$$

# 扩展：概率论角度下的线性回归

## 概率论角度的最小二乘法

· 给定训练样本集：$\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$，$\mathbf{y} = \{y_1, \cdots, y_N\}$
· $\boldsymbol{w}$的(条件)似然函数：$p(\mathbf{y} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}\left(y_i \mid f(\boldsymbol{x}_i, \boldsymbol{w}), \beta^{-1}\right)$
· 对数似然函数：$\ln p(\mathbf{y} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \sum_{i=1}^{m} \ln \mathcal{N}(y_i \mid f(\boldsymbol{x}_i, \boldsymbol{w}), \beta^{-1})$
· 改写对数似然函数：

$$\frac{m}{2}\ln\beta - \frac{m}{2}\ln(2\pi) - \beta E_D(\boldsymbol{w}),\ \ E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{m} \{y_i - f(\boldsymbol{x}_i, \boldsymbol{w})\}^2$$

· $\boldsymbol{w}$ 的最大似然估计(MLE)：

$$\boldsymbol{w}_{MLE}^{\star} = \arg\max_{w} \ln p(\mathbf{y} \mid \boldsymbol{X}, \boldsymbol{w}, \beta)$$

· 有结论：

$$\boldsymbol{w}_{MLE}^{*} = \boldsymbol{w}_{LS}^{*}$$

# 扩展：概率论角度下的线性回归

- 最大似然估计与最大后验估计

> 最大似然估计

假设 $p(\boldsymbol{z} \mid \boldsymbol{\theta})$ 的概率密度函数形式（ $\boldsymbol{\theta}$ 的似然分布）： $q(\boldsymbol{z}; \boldsymbol{\theta})$

$\mathcal{D} = \{\boldsymbol{z}\}_{i=1}^{n}$
Likelihood $p(\mathcal{D} \mid \boldsymbol{\theta})$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D} \mid \boldsymbol{\theta}) = \arg\max \prod_{i=1}^{m} q(\boldsymbol{z}; \boldsymbol{\theta})$$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \log L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} [\sum_{i=1}^{m} \log q(\boldsymbol{z}_i; \boldsymbol{\theta})]$$

$$p(\boldsymbol{z}) = q(\boldsymbol{z}; \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}})$$

# 扩展：概率论角度下的线性回归

## • 最大似然估计与最大后验估计

> 最大后验估计

假设 $p(\boldsymbol{z} \mid \boldsymbol{\theta})$的概率密度函数形式（ $\boldsymbol{\theta}$ 的似然分布）： $q(\boldsymbol{z}; \boldsymbol{\theta})$
假设$\boldsymbol{\theta}$的先验概率分布：$p(\boldsymbol{\theta})$
得到$\boldsymbol{\theta}$的先验概率分布：$p(\boldsymbol{\theta} \mid \mathcal{D})$

$\mathcal{D} = \{\boldsymbol{z}\}_{i=1}^{n}$
Posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathcal{D})$$

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} (\sum_{i=1}^{m} \log q(\boldsymbol{z}_i \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}))$$

$$p(\boldsymbol{z}) = q(\boldsymbol{z}; \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})$$

# 扩展：概率论角度下的线性回归

## 概率论角度的岭回归

· 给定训练样本集：$\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$，$\mathbf{y} = \{y_1, \cdots, y_N\}$

· $\boldsymbol{w}$的(条件)似然函数：$p(\mathbf{y} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}\left(y_i \mid f(\boldsymbol{x}_i, \boldsymbol{w}), \beta^{-1}\right)$

· $\boldsymbol{w}$的先验分布：$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}\left(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\boldsymbol{I}\right)$

· 推出$\boldsymbol{w}$的后验分布：

$$p(\boldsymbol{w} \mid \mathbf{y}) \propto \prod_{i=1}^{m} \mathcal{N}\left(y_i \mid f(\boldsymbol{x}_i, \boldsymbol{w}), \beta^{-1}\right) \mathcal{N}\left(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\boldsymbol{I}\right)$$

· $\boldsymbol{w}$ 的后验似然估计(MAP)：

$$\boldsymbol{w}_{MAP}^{*} = \arg\max_{w} \ln p(\boldsymbol{w} \mid \mathbf{y})$$

$$= \arg\max_{w}\left(-\frac{\beta}{2}\sum_{i=1}^{m} \{y_i - f(\boldsymbol{x}_i, \boldsymbol{w})\}^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{const}\right)$$

· 有结论：

$$\boldsymbol{w}_{MAP}^{*} = \boldsymbol{w}_{RR}^{*}$$