LiVE Group
视觉智能与学习研究中心

机器学习（第5讲）

主讲：张 磊

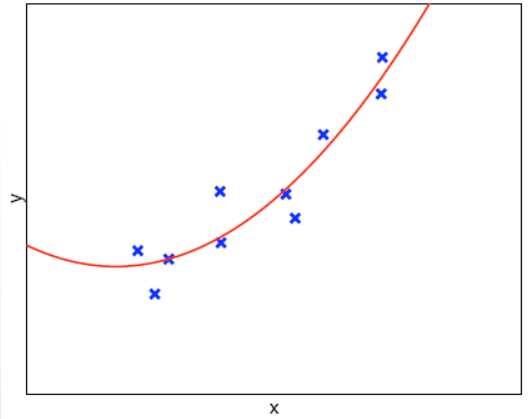
E-mail: leizhang@cqu.edu.cn
Lab Website: http://www.leizhang.tk

第五章： 正则化

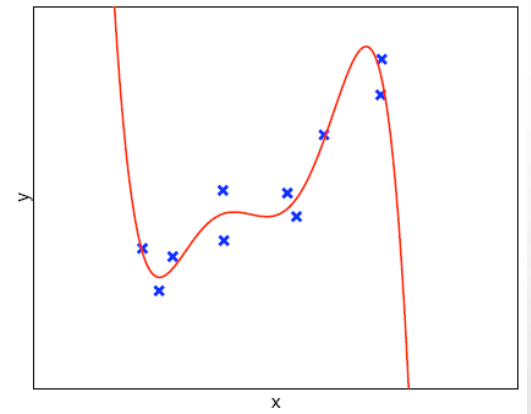
正则化是机器学习中的重要一环，其目的是为了降低模型复杂度，增强模型的抗噪声能力，防止模型对训练数据集的过拟合（过学习）。换句话说，正则化是用于对模型参数的一种约束。

举例：一个2阶和5阶的拟合函数

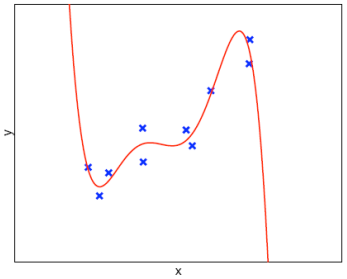
奥卡姆剃刀定律（Occam‘s Razor, Ockham’s Razor）：大道至简



$$h_\theta(x)=\theta_0+\theta_1 x+\theta_2 x^2$$



$$h_\theta(x)=\theta_0+\theta_1 x+\theta_2 x^2+\theta_3 x^3+\theta_4 x^4+\theta_5 x^5$$



原线性模型的最小化表达式为

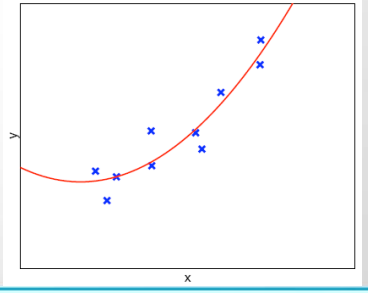
$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x_i)-t_i\right)^2 \Rightarrow h_\theta(x)=\theta_0+\theta_1 x+\theta_2 x^2+\theta_3 x^3+\theta_4 x^4+\theta_5 x^5$$

为了获得一个较好的模型，我们希望模型复杂度越小越好（阶数越小），也就是希望得到 $\theta_3,\theta_4,\theta_5 \approx 0$ (参数变少，得到简单的模型假设)

那么模型应当改进如下

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x_i)-t_i\right)^2 +10000\cdot\theta_3^2+10000\cdot\theta_4^2+10000\cdot\theta_5^2$$

$$h_\theta(x)=\theta_0+\theta_1 x+\theta_2 x^2$$



一个好的机器学习模型，既要反映数据中的蕴含信息，又要反映先验信息。

将上述模型写成以下形式：

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left(h_{\theta}(x_i)-t_i\right)^2+\lambda\cdot\sum_{j=1}^{m}\theta_j^2$$

正则化项

正则化参数 λ 是一个可调的参数（误差项和正则项之间的平衡系数）,如果该值非常大,则获得的参数就会非常小! 容易产生"欠拟合"，因此在算法编程中，可调试该参数。

1）正则化实际上是在经验误差函数上加约束，这种约束的解释就是先验知识，对参数进行正则化等价于对参数引入某种先验分布。

2）正则化解决了逆问题的不适定性，产生的解依赖于数据，而不易受噪声影响。

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left(h_{\theta}(x_i)-t_i\right)^2+\lambda\cdot\sum_{j=1}^{m}\theta_j^2$$

正则化项

实际上,正则化是对模型参数的约束,带有约束的模型可以进一步改写

正则化是机器学习也是机器学习中的重要一环，其目的是为了降低模型复杂度，增强模型的抗噪声能力，防止模型对训练数据集的过拟合（过学习）。换句话说，正则化是用于对模型参数的一种约束。

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x_i)-t_i\right)^2 + \lambda\cdot\sum_{j=1}^{m}\theta_j^2$$

正则化项

等价于

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x_i)-t_i\right)^2$$

$$s.t.\sum_{j=1}^{m}\theta_j^2 \le \xi$$

相当于在原模型基础上，加了一个约束项

等价性证明

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x_i)-t_i\right)^2+\lambda\cdot\sum_{j=1}^{m}\theta_j^2$$

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x_i)-t_i\right)^2$$

$$s.t.\sum_{j=1}^{m}\theta_j^2\le\xi$$

对下面的模型利用拉格朗日乘子法构造目标函数

$$\min_{\theta} J(\theta)=\frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x_i)-t_i\right)^2+\lambda\cdot\left(\sum_{j=1}^{m}\theta_j^2-\xi\right)$$

□多维下的线性回归模型表达

定义一组观测数据集（$\mathbf{X},\mathbf{Y}$）,其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$, $\mathbf{X}$为属性，$\mathbf{Y}$为响应标签。$\mathbf{x}_i$ (i=1,...,N) 是d-维的向量。

$$\min_{\boldsymbol{\theta}} \frac{1}{2N}\sum_{i=1}^{N}\left\|h_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right)-\mathbf{y}_i\right\|_2^2$$

其中，$\boldsymbol{\theta}$ 是模型参数（向量或矩阵）。

□多维下的正则化线性回归模型表达

定义一组观测数据集（$\mathbf{X},\mathbf{Y}$）,其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$, $\mathbf{X}$为属性，$\mathbf{Y}$为响应标签。$\mathbf{x}_i$ (i=1,...,N) 是d-维的向量。

$$\min_{\boldsymbol{\theta}} \frac{1}{2N}\sum_{i=1}^{N}\left\|h_{\boldsymbol{\theta}}(\mathbf{x}_i)-\mathbf{y}_i\right\|_2^2+\lambda\cdot\|\boldsymbol{\theta}\|_2^2$$

其中，$\boldsymbol{\theta}$ 是模型参数（向量或矩阵）。

与非正则化相比，目标函数中多了一项，即在最小化误差的同时，还要保持参数值比较小。

◘ 正则化最小二乘模型（岭回归）

定义一组观测数据集（$\mathbf{X},\mathbf{Y}$）,其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$, $\mathbf{X}$为属性，$\mathbf{Y}$为响应标签。$\mathbf{x}_i$ (i=1,...,N) 是d-维的向量。

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left\| \boldsymbol{\theta}^{\mathrm{T}} \mathbf{x}_i - \mathbf{y}_i \right\|_2^2 + \lambda \cdot \|\boldsymbol{\theta}\|_2^2$$

等价于

$$\min_{\boldsymbol{\theta}} \left\| \mathbf{X}^{\mathrm{T}} \boldsymbol{\theta} - \mathbf{Y} \right\|_{\mathrm{F}}^2 + \lambda \cdot \|\boldsymbol{\theta}\|_{\mathrm{F}}^2$$

其中，$\boldsymbol{\theta}$ 是模型参数（向量或矩阵）。

◘ 正则化最小二乘模型（岭回归）

定义一组观测数据集（$\mathbf{X},\mathbf{Y}$）,其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$, $\mathbf{X}$为属性，$\mathbf{Y}$为响应标签。$\mathbf{x}_i$ (i=1,...,N) 是d-维的向量。

$$\min_{\boldsymbol{\theta}} \left\| \mathbf{X}^{\mathrm{T}}\boldsymbol{\theta} - \mathbf{Y} \right\|_{\mathrm{F}}^2 + \lambda \cdot \left\| \boldsymbol{\theta} \right\|_{\mathrm{F}}^2$$

模型求解：

令 $J(\boldsymbol{\theta}) = \left\| \mathbf{X}^{\mathrm{T}}\boldsymbol{\theta} - \mathbf{Y} \right\|_{\mathrm{F}}^2 + \lambda \cdot \left\| \boldsymbol{\theta} \right\|_{\mathrm{F}}^2$ 为目标函数。

求导：

$$\frac{dJ(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta} - \mathbf{Y}\right) + \lambda \cdot \boldsymbol{\theta} = 0 \quad \Rightarrow \quad \boldsymbol{\theta} = \left(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda \cdot \mathbf{I}\right)^{-1}\mathbf{X}\mathbf{Y}$$

直接写出解析解

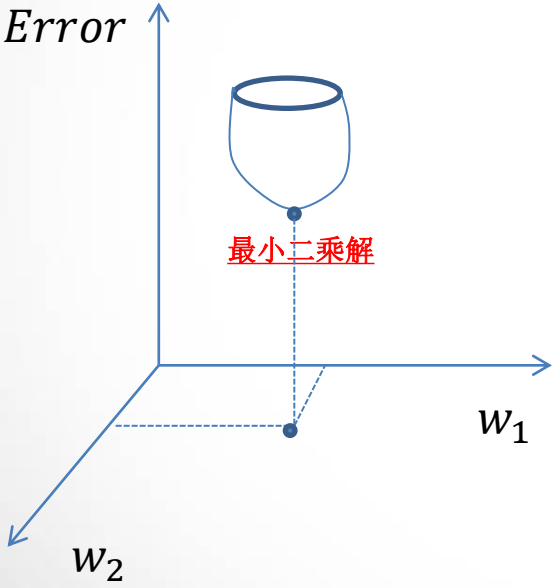
□多维下的正则化线性回归模型表达

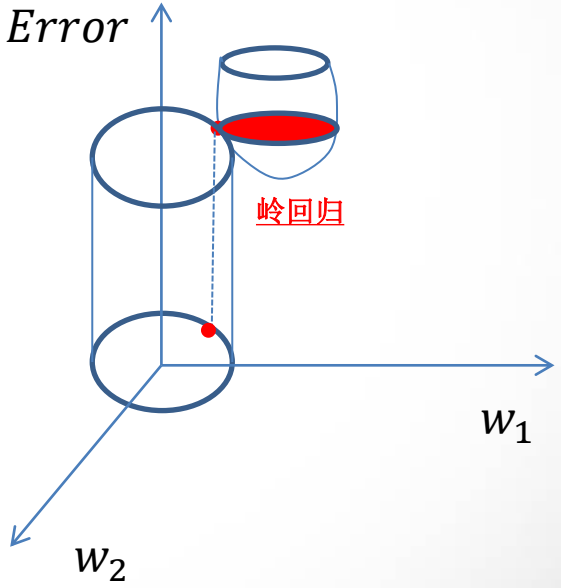
定义一组观测数据集（$\mathbf{X},\mathbf{Y}$），其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$，$\mathbf{X}$为属性，$\mathbf{Y}$为响应标签。$\mathbf{x}_i$ (i=1,...,N) 是d-维的向量。

$$\min_{\boldsymbol{\theta}} \frac{1}{2N}\sum_{i=1}^{N}\left\|h_{\boldsymbol{\theta}}(\mathbf{x}_i)-\mathbf{y}_i\right\|_2^2+\lambda\cdot\left\|\boldsymbol{\theta}\right\|_2^2$$

其中，$\boldsymbol{\theta}$ 是模型参数（向量或矩阵）。

注：这里正则化项的设计采用的是L2-norm(Tikhonov regularization**)，是一种很平滑的约束，但并不稀疏。**

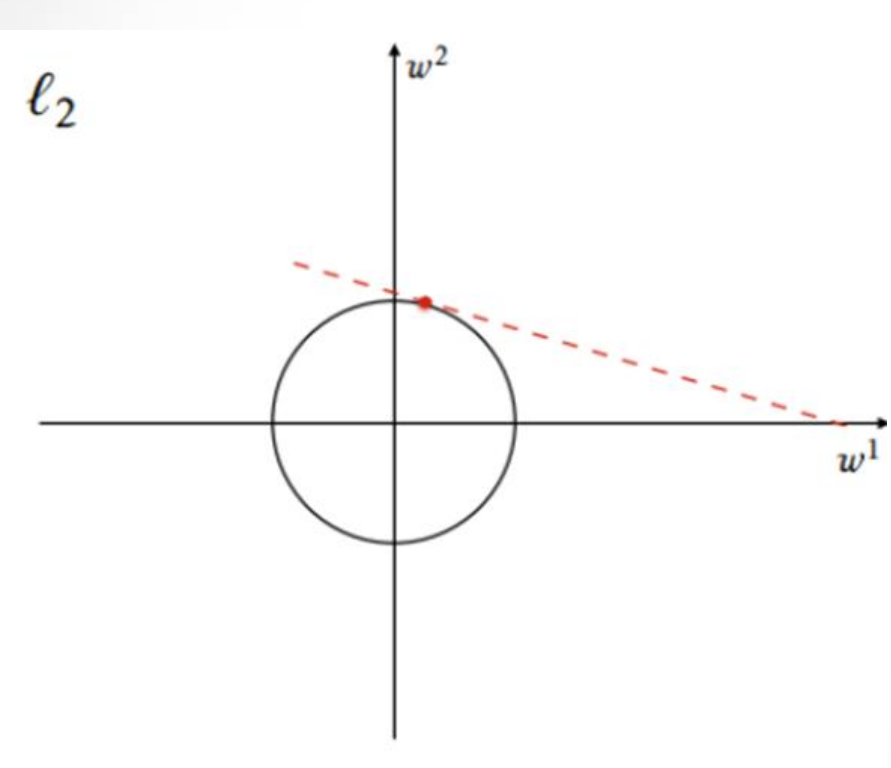


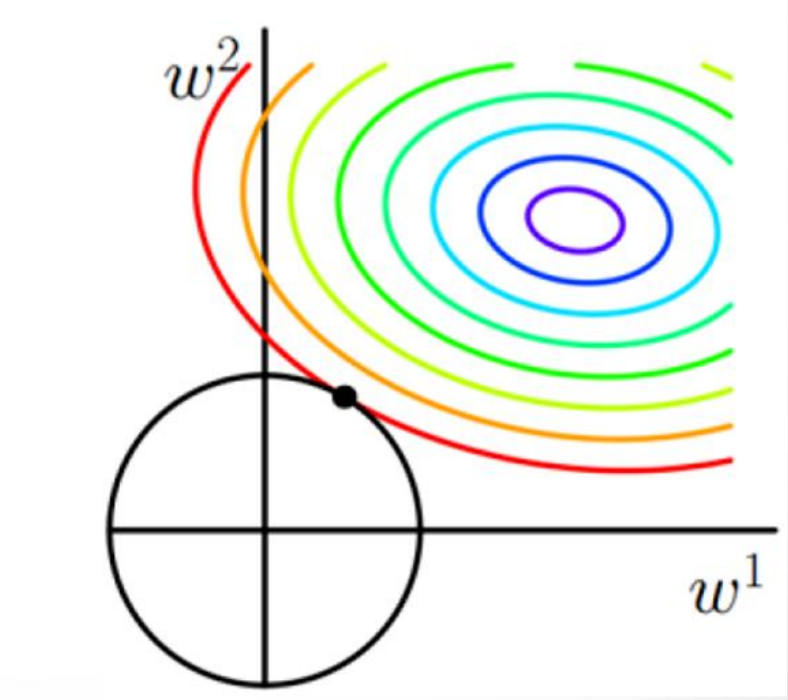


$$w_1^2 + w_2^2 = r^2$$

解的几何分析

第五章：正则化regularization





□吉洪诺夫正则化(Tikhonov Regularization, 1943, 1963)

定义一组观测数据集（$\mathbf{X},\mathbf{Y}$）,其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$, $\mathbf{X}$为属性，$\mathbf{Y}$为响应标签。$\mathbf{x}_i$ (i=1,...,N) 是d-维的向量。

$$\min_{\boldsymbol{\theta}} \frac{1}{2N}\sum_{i=1}^{N}\left\|h_{\boldsymbol{\theta}}(\mathbf{x}_i)-\mathbf{y}_i\right\|_2^2+\lambda\cdot\left\|\Gamma\boldsymbol{\theta}\right\|_2^2$$

其中，Γ是**吉洪诺夫**矩阵,在许多情况下$\Gamma=\alpha\cdot\mathbf{I}$。

对于L2-Norm,解的稀疏性不好，（如何能保证求解的稀疏性？）

□多重共线性问题

多重共线性是指模型中解释变量之间存在精确相关关系或高度相关关系而使模型失真或难以估计。

有效数据是获得精确模型的前提条件。在最小二乘问题中，当多重共线性问题出现时，$|X'X|$接近于0，导致奇异发生，从而其矩阵的逆不存在。

在正则化最小二乘问题中，通过引入λI可较大程度上降低奇异矩阵的发生概率。

$$若X=\begin{bmatrix}1 & 3 & 4\\ 2 & 6 & 8\end{bmatrix}, 则\ det\left(X^TX\right)=0$$

数据矩阵X非列满秩，如何理解“秩”的意义？

□多重共线性问题

特征的选择，即强解释性变量的选择。

采用稀疏正则化，对于删除模型中无效变量（导致共线性）或无意义变量具有较好的效果。

在L2范式下的约束中，使得模型的解收缩到一个较小的值，但收缩到0的概率较低。对于无效变量的删除作用不强。

□多维下的稀疏正则化线性回归模型表达

定义一组观测数据集（$\mathbf{X},\mathbf{Y}$）,其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$, $\mathbf{X}$为属性，$\mathbf{Y}$为响应标签。$\mathbf{x}_i$ $(i=1,\ldots,N)$ 是d-维的向量。

$$\min_{\boldsymbol{\theta}} \frac{1}{2N}\sum_{i=1}^{N}\left\|h_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right)-\mathbf{y}_i\right\|_2^2+\lambda\cdot\|\boldsymbol{\theta}\|_0$$

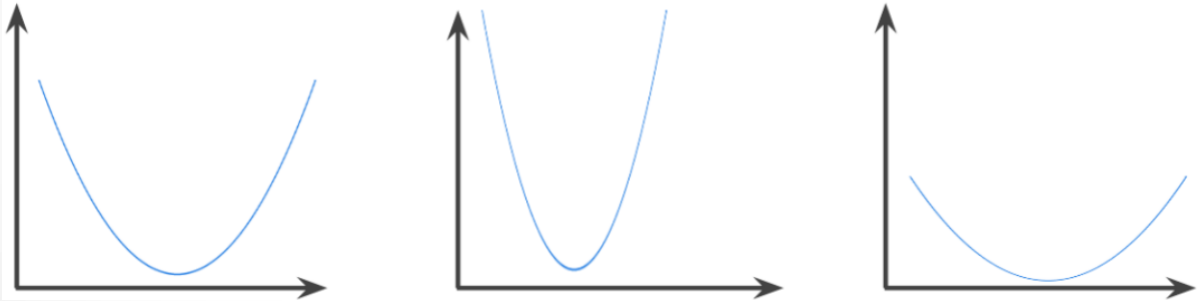
L0-范数表示向量 θ 中非零元素的个数。

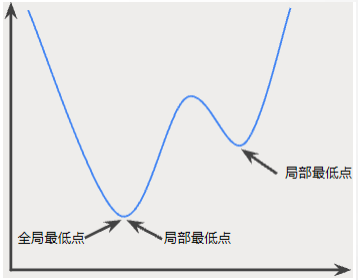
注：这里正则化项的设计采用的是L0-norm,是一种强稀疏约束。

但求解时，由于L0的非凸性，难以求解NP-hard。通常采用松弛化。

□基本概念

凸函数、凸优化、凸集





$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

Jensen不等式（凸函数的充要条件）

若f(x)是[a,b]的凸函数，对于区间内的任意x1,x2,…,xn，有

$$f\left(\sum_{i=1}^{n} t_i x_i\right) \leq \sum_{i=1}^{n} t_i f(x_i), where \sum_{i=1}^{n} t_i = 1$$

Least absolute shrinkage and selection operator, Lasso

□多维下的稀疏正则化线性回归模型表达 (Lasso Regression)

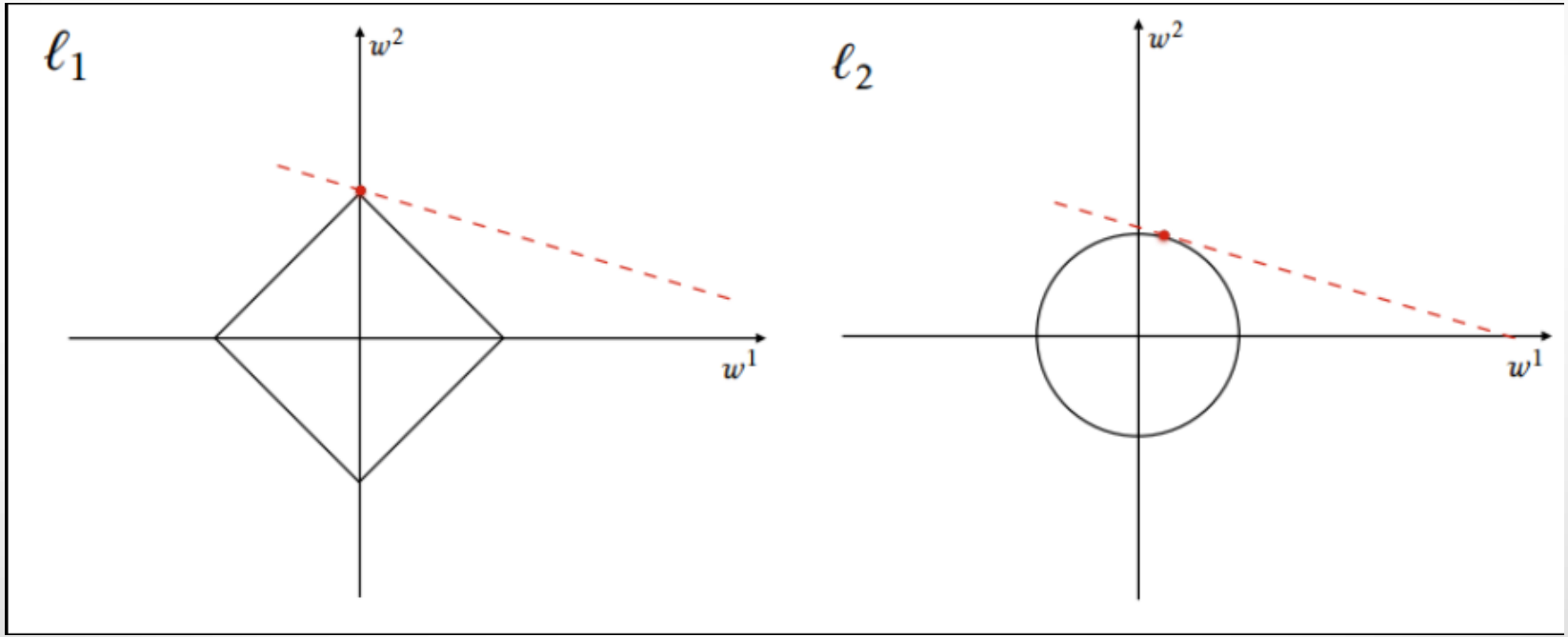
定义一组观测数据集（$\mathbf{X},\mathbf{Y}$）,其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$, $\mathbf{X}$为属性，$\mathbf{Y}$为响应标签。$\mathbf{x}_i$ $(i=1,\ldots,N)$ 是d-维的向量。

$$\min_{\boldsymbol{\theta}} \frac{1}{2N}\sum_{i=1}^{N}\left\|h_{\boldsymbol{\theta}}(\mathbf{x}_i)-\mathbf{y}_i\right\|_2^2+\lambda\cdot\|\boldsymbol{\theta}\|_1$$

L1-范数的表达式。 $\|\boldsymbol{\theta}\|_1=\sum_{d=1}^{D}|\theta_d|$

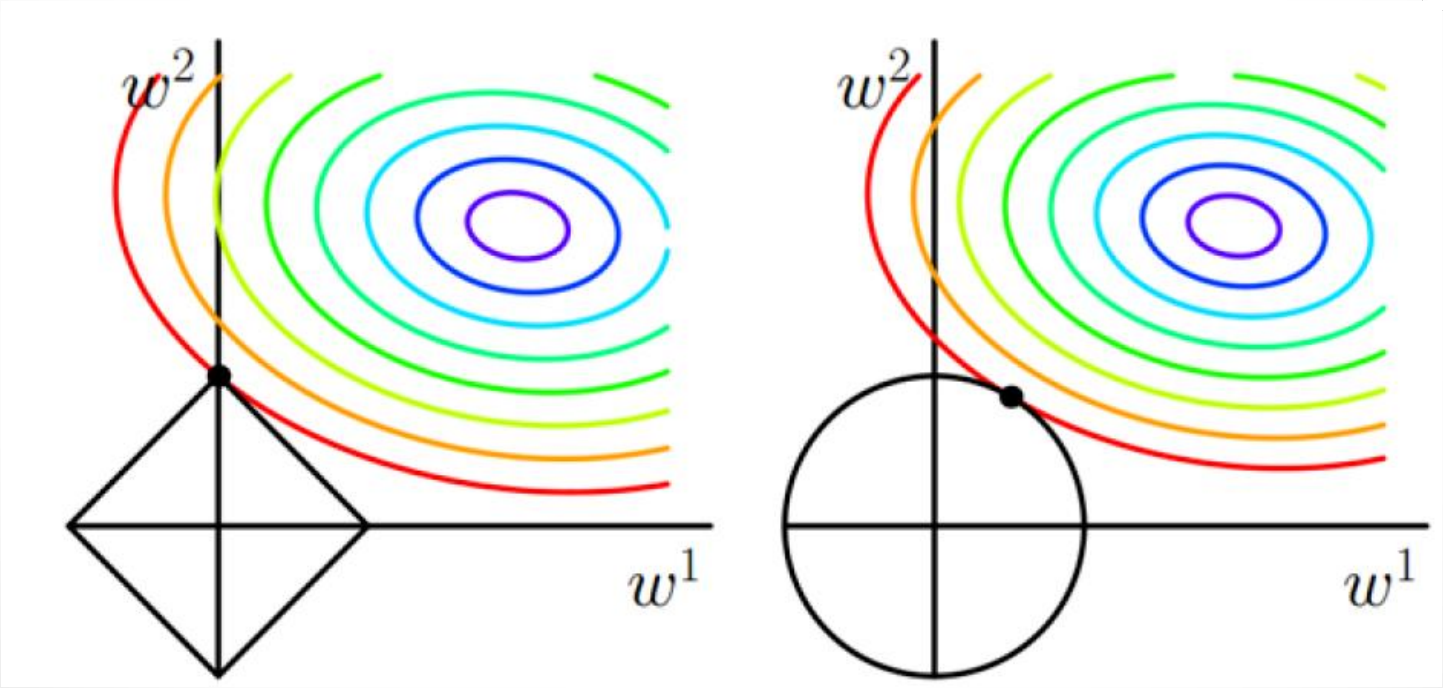
注：这里正则化项的设计采用的是L1-norm,是L0的凸松弛，近似的稀疏解，可以通过Lasso算法迭代求解(虽然凸，但在x=0处不严格可导/可微，或者说在x=0处不存在导数)。

第五章：正则化regularization

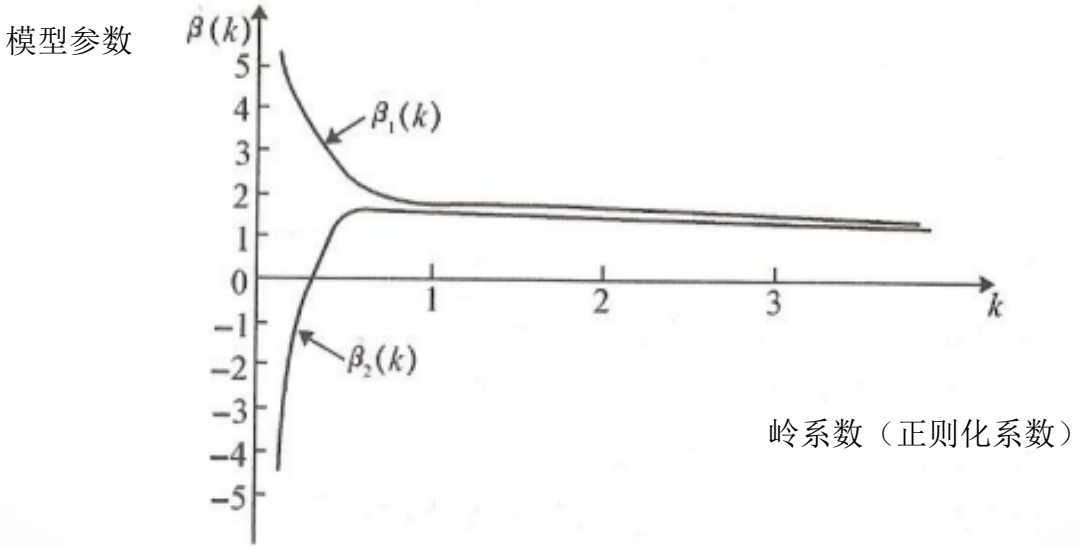


以两变量w1,w2为例。Ridge和Lasso都是有偏估计，最小二乘是无偏估计

第五章：正则化regularization



L1有角，解落在角点的可能性更大，而L2没有角，因此不可能落在坐标系上。等高线抛物面在水平面上的投影。



岭回归名字的由来：模型的解与正则化参数的关系图类似一个山岭。

ElasticNet模型(Lasso Regression+Ridge Regression)

定义一组观测数据集（$\mathbf{X},\mathbf{Y}$）,其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$, $\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]$, $\mathbf{X}$为属性，$\mathbf{Y}$为响应标签。$\mathbf{x}_i$ (i=1,...,N) 是d-维的向量。

$$\min_{\theta} \frac{1}{2N} \sum_{i=1}^{N} \left\| h_\theta(\mathbf{x}_i) - \mathbf{y}_i \right\|_2^2 + \lambda \cdot \left\| \theta \right\|_1 + u \cdot \left\| \theta \right\|_2$$

注：这里正则化项的设计采用的是L1-norm和L2-norm的混合体。

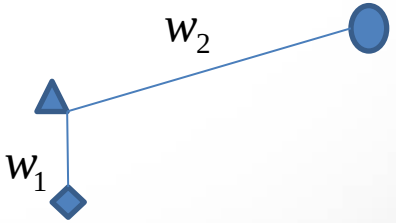
◘半监督正则化（流形正则）

定义一组新的无标签的观测数据集$\mathbf{X}$,其中$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]$，$\mathbf{X}$为属性，无标签。$\mathbf{x}_i$ $(i=1,\ldots,N)$ 是d-维的向量。

因为样本无标签，则不能再用损失函数对函数 f 进行定义。

不妨做出以下假设：在欧式空间内，距离较近的两个样本$\mathbf{x}_1,\mathbf{x}_2$，其具有相同响应即$f(\mathbf{x}_1)= f(\mathbf{x}_2)$的概率更大。

设两点之间的连线，w存在一个权重即重要性系数。



◘半监督正则化（流形正则）

局部线性嵌入（local linear embedding, Science 2000）。

一种非线性降维模型，使得流形学习(manifold learning)得到迅速发展。

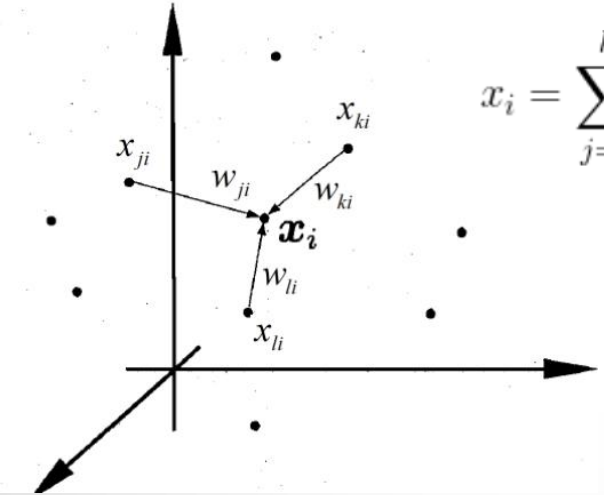
假设数据是均匀采样于一个高维欧氏空间中的低维流形，流形学习就是从高维采样数据中恢复低维流形结构，即找到高维空间中的低维流形，并求出相应的嵌入映射，以实现维数约简或者数据可视化。

所谓局部线性是指数据的某个局部是线性的，但整体是非线性的。

流形学习即学习从高维到低维的嵌入映射，使得数据的局部拓扑结构不变。

◘半监督正则化（流形正则）

局部线性嵌入（local linear embedding, Science 2000）。



$$x_i = \sum_{j=1}^{k} w_{ji} x_{ji} \longrightarrow \arg\min_{w} \sum_{i=1}^{N} (\|x_i - \sum_{j=1}^{k} w_{ji} x_{ji}\|_2^2)$$

$$\arg\min_{Y} \sum_{i=1}^{N} (\|y_i - \sum_{j=1}^{k} w_{ji} y_{ji}\|_2^2)$$

□半监督正则化（流形正则）

按照上面的分析，半监督正则项的表达式可以写成

$$R(f)=\sum_{i,j} w_{i,j}\left(f(x_i)-f(x_j)\right)^2$$

其中，$w_{i,j}$ 是连接样本x_i和x_j之间的权重（重要性系数），属于给定的值，其定义方法可以根据欧式距离来表达。比如，

$$w_{i,j}=\exp\left(-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2\right)$$

显然，权重矩阵W是一个对称阵。

□基于半监督正则化（流形正则）的线性建模

按照上面的分析，半监督正则项的表达式可以写成

$$\min_f \sum_k \left(f(x_k)-y_k\right)^2+\alpha\cdot\|f\|_2+\beta\cdot R(f)$$

$$\min_f \sum_k \left(f(x_k)-y_k\right)^2+\alpha\cdot\|f\|_2+\beta\cdot\sum_{i,j} w_{i,j}\left(f(x_i)-f(x_j)\right)^2$$

□Dropout

Dropout是深度神经网络（深度学习）框架中的基本操作(Hinton, 2012)，是一种典型的深度学习模型正则化技术，防止overfitting。

核心思想：通过多模型/网络组合的形式改善算法鲁棒性。

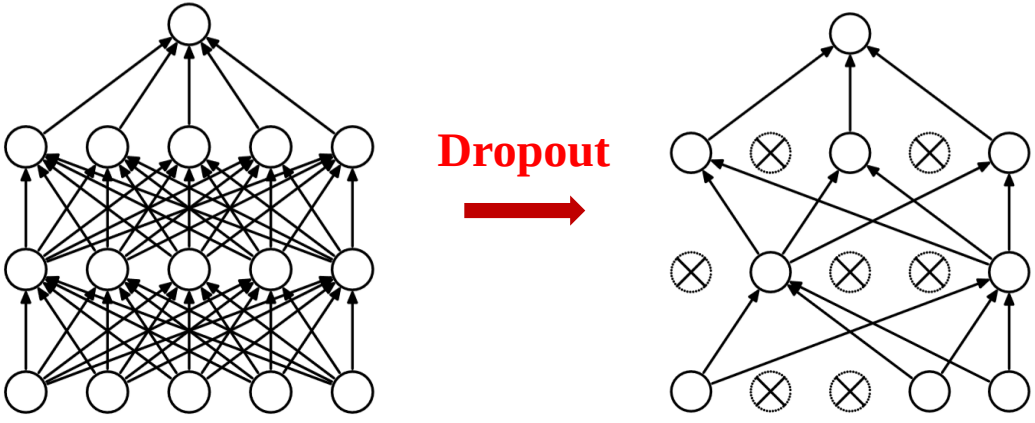
$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M}\hat{Y}_m$$

其中，$\hat{Y}_m$表示每个模型的预测结果。

存在问题：如果要训练M个不同结构的模型/网络，成本高、效率低。

□Dropout

Dropout就是在优化过程中，以一定的概率，随机“删除”网络模型中的一些特征节点（隐层神经元）。

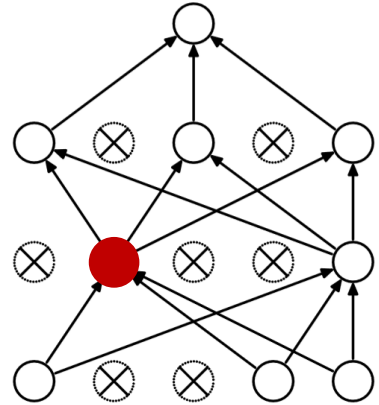


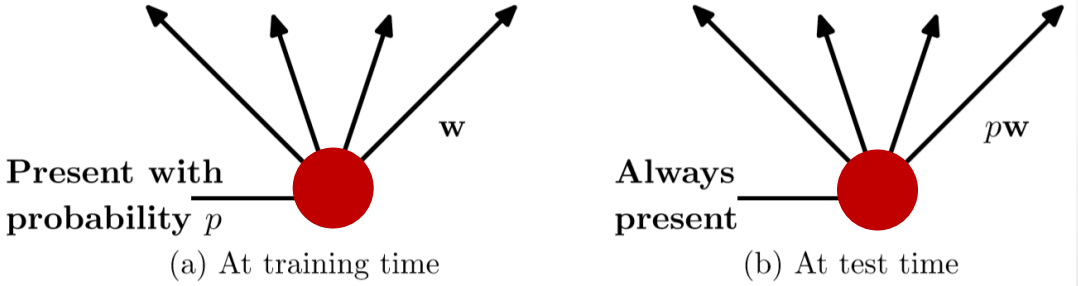
(a) Standard Neural Net　　(b) After applying dropout.

Hinton, et al. “Improving neural networks by preventing co-adaptation of feature detectors,” arXiv, 2012.
Hinton, et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research, 2014.

□Dropout

Dropout就是在优化过程中，以一定的概率，随机“删除”网络模型中的一些特征节点（隐层神经元）。





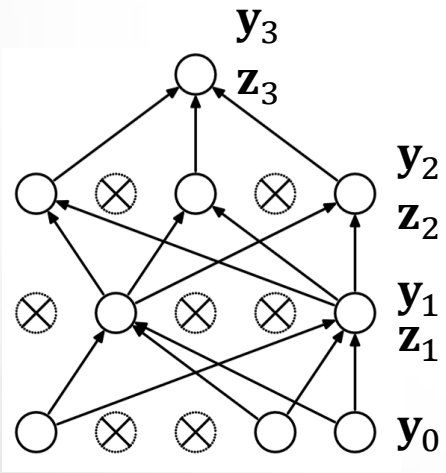
注1：Dropout只出现在训练/学习阶段

注2：测试阶段，采用pw，保持期望不变px+(1-p)0=px

从模型参数控制/约简角度，Dropout与L1, L2正则相似。

□Dropout

Dropout就是在优化过程中，以一定的概率，随机“删除”网络模型中的一些特征节点（隐层神经元）。



z表示每层输入, y表示每层输出

$$
\begin{aligned}
z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)}\mathbf{y}^l + b_i^{(l+1)}, \\
y_i^{(l+1)} &= f(z_i^{(l+1)}),
\end{aligned}
$$

$$
\begin{aligned}
r_j^{(l)} &\sim \text{Bernoulli}(p), \\
\widetilde{\mathbf{y}}^{(l)} &= \mathbf{r}^{(l)} * \mathbf{y}^{(l)},
\end{aligned}
$$

$$
\begin{aligned}
z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)}\widetilde{\mathbf{y}}^l + b_i^{(l+1)}, \\
y_i^{(l+1)} &= f(z_i^{(l+1)}).
\end{aligned}
$$

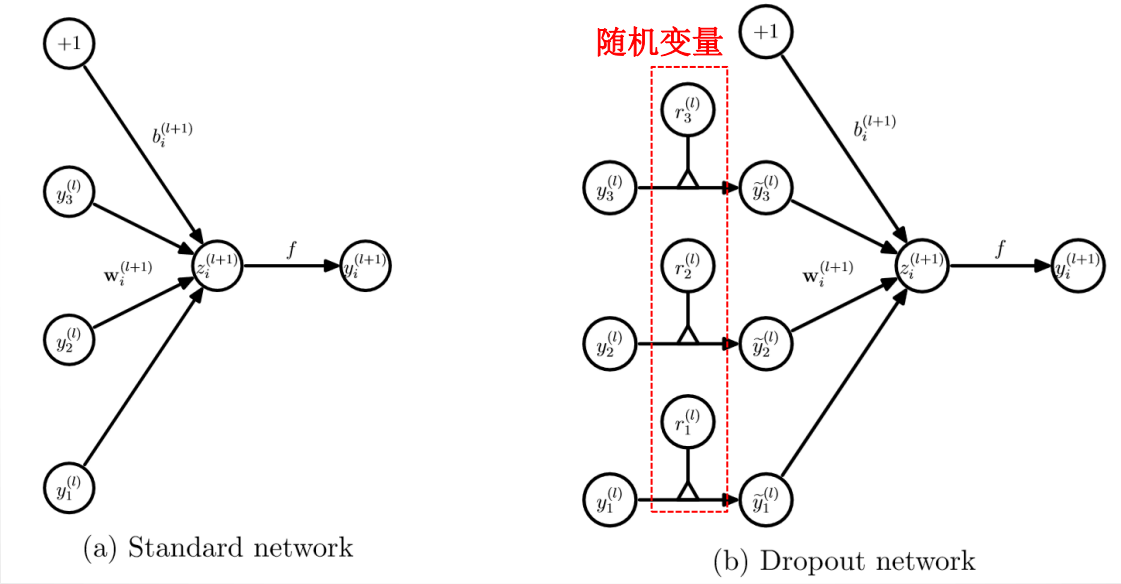
标准前向计算单元

Dropout单元

带有Dropout的基本前向计算单元

Dropout

Dropout就是在优化过程中，以一定的概率，随机“删除”网络模型中的一些特征节点（隐层神经元）。

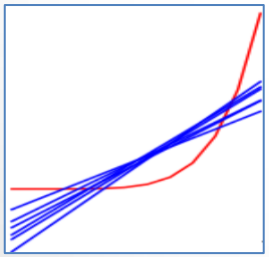
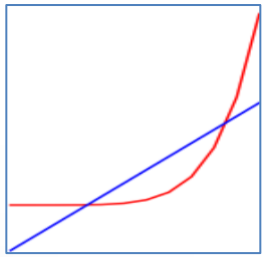


(a) Standard network　　(b) Dropout network

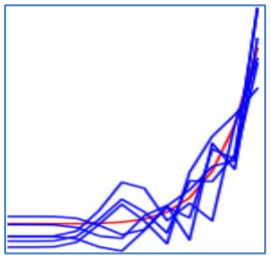
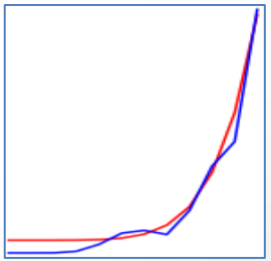
正则化、泛化能力、过拟合三者关系

正则化的提出是用于提升模型的泛化能力，避免过拟合问题。

1. 泛化能力是指，训练模型不仅要保证训练集的误差最小化，同时还要保证测试性能，通常称为“偏差-方差”平衡问题（或者折中问题，Bias-Variance trade-off）。偏差是针对训练集，方差针对验证集。
2. Total error=bias^2+variance+noise

偏差大
方差小
欠拟合

偏差小
方差大
过拟合

总体误差= $\mathrm{E}\left[(y-\hat{f})^2\right]$ 偏差= $(f-\mathrm{E}[\hat{f}])^2$ 方差= $\mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right]$ 噪声=E[(f-y)^2] (人工标注误差)

$$Err(x)=\left(E[\hat{f}(x)]-f(x)\right)^2+E\left[\hat{f}(x)-E[\hat{f}(x)]\right]^2+\sigma_e^2$$

正则化、泛化能力、过拟合三者关系

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\hat{f}(x) - E[\hat{f}(x)]\right]^2 + \sigma_e^2$$

推导过程

$$Err(x) = E\left[(\hat{f} - y)^2\right] = E\left[(\hat{f} - \bar{f} + \bar{f} - y)^2\right]$$

$$= E\left[(\hat{f} - \bar{f})^2\right] + E\left[(\bar{f} - y)^2\right] + 2E\left[(\hat{f} - \bar{f}) \cdot (\bar{f} - y)\right] = E\left[(\hat{f} - \bar{f})^2\right] + E\left[(\bar{f} - y)^2\right]$$

$$= E\left[(\hat{f} - \bar{f})^2\right] + E\left[(\bar{f} - f + f - y)^2\right]$$

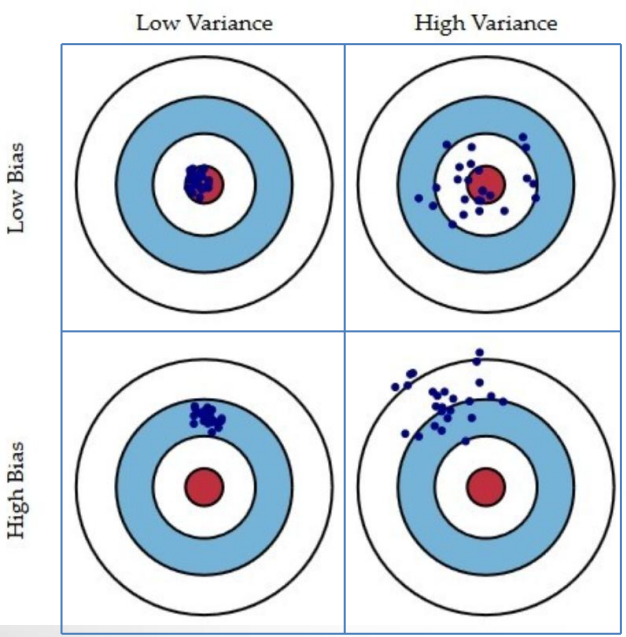
$$= E\left[(\hat{f} - \bar{f})^2\right] + E\left[(\bar{f} - f)^2\right] + E[(f - y)^2] + 2E\left[(\bar{f} - f) \cdot (f - y)\right]$$

$$= E\left[(\hat{f} - \bar{f})^2\right] + (\bar{f} - f)^2 + E[(f - y)^2]$$

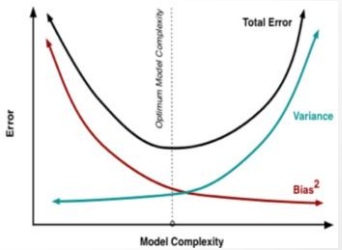
$$= Variance + Bias^2 + \sigma_e^2$$

其中，y为真实值，f为人工标定值，$\hat{f}$为预测值，$\bar{f}$为多次模型预测的均值

正则化、泛化能力、过拟合三者关系



机器学习是通过有限的训练数据，利用数学模型获得对无限数据的近似。



根本原因：

1. 模型的不完美
2. 训练数据有限

解决方法：k-fold cross-validation、大数据集、深度学习

寻找泛化、欠拟合、过拟合之间的平衡问题，就是“偏差-方差”折中问题。