特征的分类

相关特征：模型训练需要的特征

无关特征：模型训练不必要的特征

冗余特征：在其他特征已知的情况下，该特征变得无用（某些特征的组合）

特征选择：给出一系列特征X={X1,X2...Xd},和目标变量Y，去最小化集合S，其实现了Y的最大分类性能（对于给定的分类器和分类性能度量集）

特征选择的方法

Wrappers methods包装法

原则：对所有特征进行枚举子集，寻求性能最好的最小子集

优点：与模型相关，通常可以获得较好的性能

缺点：计算开销大

枚举是2^n，不切合实际，一般使用贪婪搜索

贪婪搜索：

前向搜索算法：一次添加一个特征，直到无法实现进一步的改进

后向搜索算法：从包含全部特征开始，一次删除一个特征，直到无法实现进一步的改进

两者的复杂度都为O(n^2)

Filters methods过滤法

原则：快速计算统计量J(Xf)代替模型评估

与模型无关，与数据的分布有关

1. Score each feature $\mathrm{X_f}$ individually based on the f-th column of the data matrix and label vector Y.
   For each feature $\mathrm{X_f}$
   Compute $J(\mathrm{X_f})$
   End
2. Rank features according to $J(\mathrm{X_f})$.
3. Choose the top k features with the highest scores.

例子：互信息，卡方检验，Person相关系数

优点：复杂度小

缺点：与模型无关，效果较差

Embedded method嵌入法

原则：将选择特征与学习过程同步

增加正则化项

$$E_D(w) + \lambda E_W(w)$$

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} (\boldsymbol{w}^T \boldsymbol{x_i} - y_i)^2 + \lambda \|\boldsymbol{w}\|_1$$

欠拟合与过拟合

概化：在非训练集上模型表现良好性能的能力

经常通过测量模型在测试集上的表现来评估概化误差

模型的性能好坏取决于：

使训练误差减小

时训练集误差与测试集误差的差距减小

欠拟合：模型在训练集上不能得到足够小的误差

过拟合：训练集误差和测试机误差太大

根据模型参数的个数可以避免过拟合和欠拟合

参数越多，在训练集上的拟合效果越好，但可能出现过拟合现象

参数过少，则可能在训练集上不能拟合训练数据

一种控制机器学习算法容量的方式：使用假设空间

Linear regression	$y = b + wx$
Introduce x^2 (quadratic model)	$y = b + w_1 x + w_2 x^2$
Continue to add more powers of x	$y = b + \sum_{i=1}^{9} w_i x^i$

当机器学习算法在以下方面的能力合适时，它们通常会表现得最好：

需要执行的任务的真正复杂性

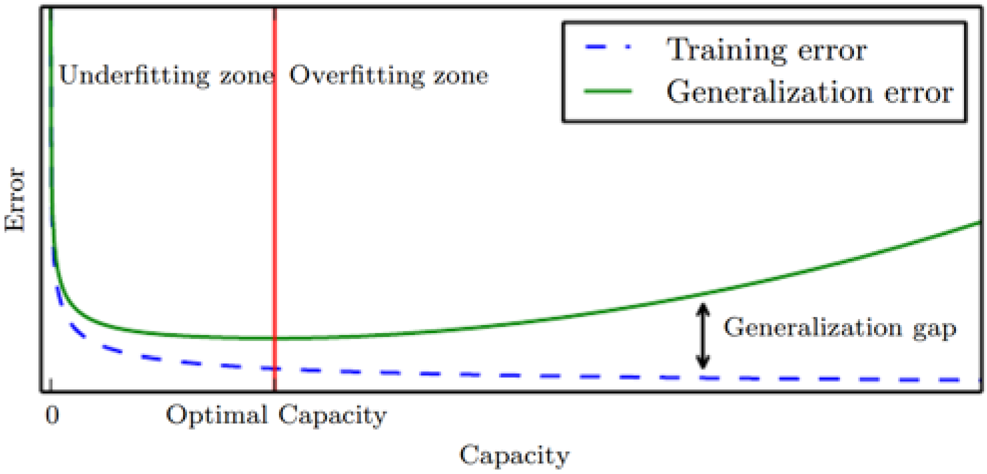
提供的训练数据量

模型参数数量的选择，依据问题本身的复杂程度和数据量大小

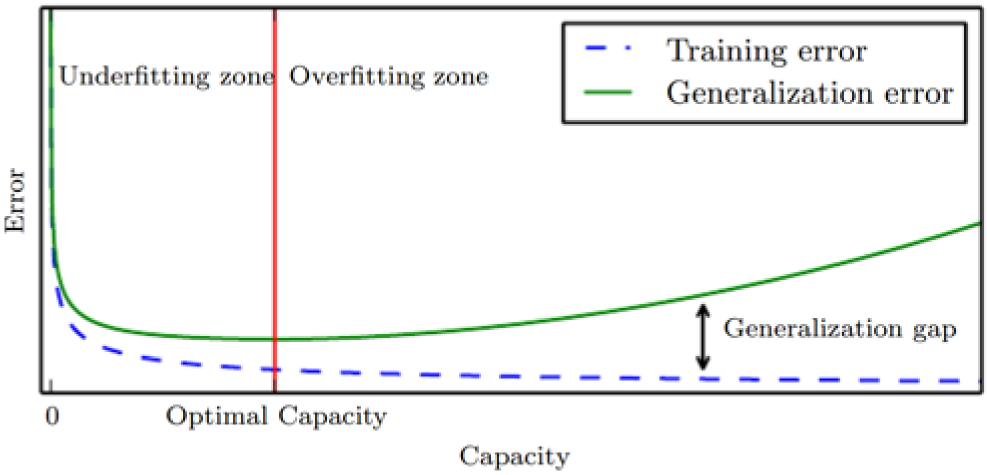
具有相似泛化能力的概化误差，优先选择模型简单的，对于复杂的模型，由于数据错误而意外拟合的可能性更大。

训练误差随着模型容量的增加而减小，直到它逐渐接近可能的最小误差值。

概化误差呈U形曲线，是模型容量的函数。



增加训练集的数量，可以减少过拟合，越大的数据集，能够承受拟合数据的模型就越复杂



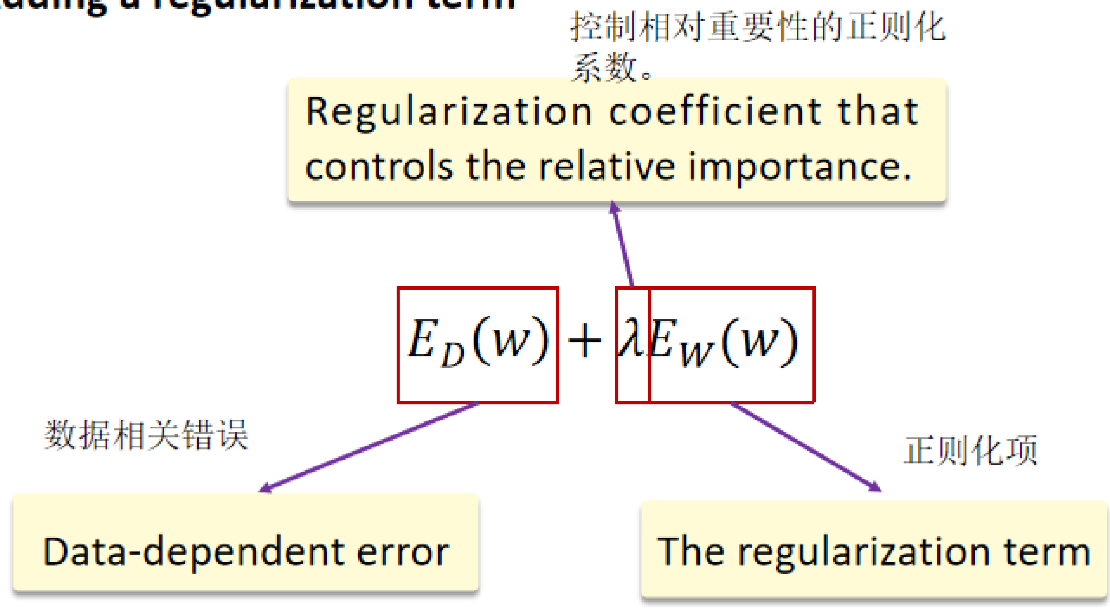
避免过拟合采取的措施是为了减少概化误差而不是训练误差

增加正则化项避免过拟合

原理：在损失函数上加上某些规则（限制），缩小解空间，从而减少求出过拟合解的可能性

通过正则化，增加的正则化项可以让损失函数在迭代过程中尽量使得w模最小，所以使得w中有的项为0，也就是相当于删去了一些特征，避免了过多参数导致的过拟合现象，简化了模型

Adding a regularization term



$$E_D(\boldsymbol{w}) = \sum_{i=1}^{m} \left(\boldsymbol{w}^T \boldsymbol{x}_i - y_i\right)^2$$

L2范数

权重衰减：归权重值向0衰减

参数收缩：参数向0收缩

优点：保留w的二次函数，所以它的精确极小值可以在闭合形式中找到。

One simple form of regularizer (L2 norm)

$$E_W(w) = \frac{1}{2} w^T w$$

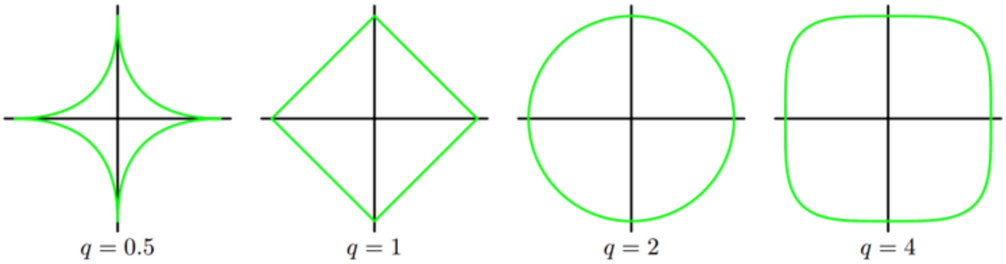
$$E_T(w) = \sum_{i=1}^{m} (\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} w^T w \quad \textbf{Ridge regression}$$

广泛的正则化

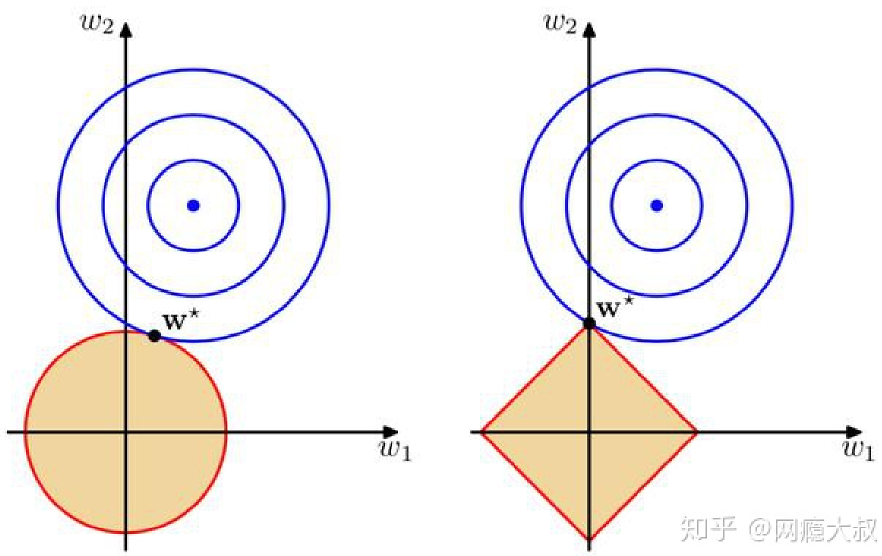
More General Regularizer

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

$\boldsymbol{w} = [w_1, w_2]$



Contours of the regularization term for various values of the *q*.



目标函数最小化的几何展示

可以看到，L1正则化的最优参数值 w^* 恰好是 $w_1 = 0$ 的时候，意味着我们剔除了模型中一个特征（系数为0等价于剔除该特征），**从而达到了降低模型复杂度的目的。**在这个意义上L1正则化效果要优于L2正则化，**但L1存在拐点不是处处可微，从而L2正则化有更好的求解特性。**

q=1对应于lasso(最小绝对收缩与选择算子)

如果**λ足够大，则一些系数wj被驱动为零，从而产生稀疏模型**

$$\boxed{\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2} + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

$E_D(w)$

Note that minimizing the above function is equivalent to minimizing the unregularized sum-of-squares error $E_D(w)$ subject to the constraint注意，最小化上述函数等价于最小化受约束的非正则化平方和误差ED（w）

$$\sum_{j=1}^{M}|w_j|^q \leqslant \eta$$

for an appropriate value of the parameter η, where the two approaches can be related using Lagrange multipliers.

对于参数η的适当值，其中两种方法可以使用拉格朗日乘子相关联

L1范数

L1 Regularization

- L1 regularization→ sparse solution.
- It can be considered analogous to performing embedded feature selection, where the trained model implicitly performs feature selection.它可以被认为类似于执行嵌入特征选择，其中训练的模型隐含地执行特征选择
- Specifically, the entries of the weight vector w_i's which are non-zero (or practically outside a low threshold $|w_i| > \epsilon$, where $\epsilon > 0$) represent features that are important for the classification task. w>0，则代表在模型中起到重要作用

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} (\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 + \lambda \|\boldsymbol{w}\|_1$$

Optimization for L1 norm

- **ISTA (Iterative Shrinkage-Thresholding Algorithms) 迭代收缩阈值算法**
- **Fast ISTA (Fast Iterative Shrinkage-Thresholding Algorithms)**
  - **Amir Beck, Marc Teboulle: A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. SIAM J. Imaging Sciences 2(1): 183-202 (2009)**

The objective function of ISTA has the form of

$$arg\ \min F(\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{X\alpha} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 = f(\boldsymbol{\alpha}) + g(\boldsymbol{\alpha})$$