特征选择

特征选择主要从两个方面入手

- **特征是否发散**：特征发散说明特征的方差大，能够根据取值的差异化度量目标信息.
- **特征与目标相关性**：优先选取与目标高度相关性的.
- 对于特征选择，有时候我们需要考虑分类变量和连续变量的不同.

过滤法

先选择特征，后训练模型（通常会指定需要特征的数目K）

需要较大的工程量

方差过滤

- 建议作为数值特征的筛选方法

计算各个特征的方差，然后根据阈值，选择方差大于阈值的特征

自变量方差越大，反映的信息越大

卡方检验

- 建议作为分类问题的分类变量的筛选方法

适用于：
自变量：分类变量
因变量：分类变量

经典的卡方检验是检验定性自变量对定性因变量的相关性。假设自变量有N种取值，因变量有M种取值，考虑自变量等于i且因变量等于j的样本频数的观察值与期望的差距，构建统计量

$$\chi^2 = \sum \frac{(A-E)^2}{E}$$

补充：卡方检验常用来
a：两分类变量是否两两独立；
b：检验某连续变量会否满足某指定分布；
c：检验某分类变量各类出现概率是否等于指定概率；

互信息法

- 建议作为分类问题的分类变量的筛选方法

经典的互信息也是评价定性自变量对定性因变量的相关性的，为了处理定量数据，最大信息系数法被提出，互信息计算公式如下

互信息越大，两者关系越强

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

包装法

递归特征消除

用一个基模型来进行多轮训练，每轮训练后，消除若干权值系数的特征，再基于新的特征集进行下一轮训练

嵌入法

特点：特征选取和模型的训练同时完成

基于L1范数：使用带惩罚项的基模型，除了筛选出特征外，同时也进行了降维

（1）避免过拟合的方法（减少特征、正则化）；
（2）L1范式正则化（LASSO） vs L2范式正则化（岭回归）；
（3）L1正则化更易获得稀疏解（稀疏解意味着参数更少，特征更少，模型更为简单）；