# 模型评估与选择

## 经验误差与过拟合

**错误率**：m个样本中有a个分类错误，即错误率E=a/m；

**精度**：与错误率相对，即1-错误率；

**误差**：学习器的预测输出和真实值的差异；

**训练误差、经验误差**：在训练集上的误差；

**泛化误差**：在新样本中的误差；

**过拟合**：训练样本自身的特点当作所有潜在样本都会有的一般性质，即训练过于完美，这样会导致泛化能力下降；无法完全避免，只能缓解；

**欠拟合**：与过拟合相对，即训练样本的一般性质尚未学好。

问题：现实任务中，学习算法有很多种，即使一种算法不同的参数也会有不同的模型产生，如何评估模型的好坏？

理想解决：对候选模型的泛化误差进行评估，选择泛化误差最小的。然而我们无法直接获得泛化误差，而训练误差又由于过拟合现象不适合作为标准。

## 评估方法

需要使用**测试集**的测试误差作为泛化误差的近似。需要注意测试集的选择：测试样本不要在训练样本中出现。

**留出法**：直接在数据集中选择一部分作为训练集，剩下的作为测试集。

**交叉验证法**：将数据集划分成k个子集，尽可能的大小相似，数据分布一致；然后选择k-1个数据集作为训练集，剩下一个作为测试集，如此得到k组训练/测试集，最终返回k个结果的均值，又称k折交叉验证。

**留一法**：交叉验证法的特例，即有m个样本，则设k=m。

**自助法**：设m个样本的数据集D，有放回的抽取m次放到数据集D’，这是D’可以作为训练集，而样本中不被抽中的概率为，这些数据作为测试集。由于样本有放回抽样，训练集存在重复数据，引发估计偏差，适用于小数据量。

## 性能度量

**错误率与精度**：1.1有提到过

**查全率(召回率)与查准率(准确率)**：对于二分类问题则有

|  |  |  |
| --- | --- | --- |
| 真实情况 | 预测情况 | |
| 正例 | 反例 |
| 正例 | TP(真正例) | FN(假反例) |
| 反例 | FP(假正例) | TN(真反例) |

查准率(准确率)：

查全率(召回率)：

F1度量：，由调和平均数得出

，由加权平均数得出，其中度量了查准率与查全率的相对重要性，时退化为标准的，时查准率有更大的影响，时查全率有更大的影响。

**PR曲线、ROC曲线、代价曲线**：略

## 比较检验

## 偏差与方差

# 线性模型

## 基本形式

假设特征值，其中表示特征数，表示样本数，目标值。

创建模型：假设参数矩阵，则有，表示预测值。

变换矩阵，，则有。

## 线性回归模型

**普通最小二乘法**：，求最小值得

**问题**：当X不是列满秩(即特征数大于样本数)时，或者某些列之间的线性相关性比较大时， IMG_256 的行列式接近于0，即 IMG_257 接近于奇异，上述问题变为一个不适定问题，此时，计算 IMG_258 时误差会很大，传统的最小二乘法缺乏稳定性与可靠性。

**岭回归(L2正则化)**

，求最小值得。

随着  的增大， 各元素的绝对值均趋于不断变小，它们相对于正确值 IMG_259 的偏差也越来越大。  趋于无穷大时，  趋于0。其中， 随  的改变而变化的轨迹，就称为**岭迹**。实际计算中可选非常多的  值，做出一个岭迹图，看看这个图在取哪个值的时候变稳定了，那就确定  值了。岭回归是对最小二乘回归的一种补充，它损失了无偏性，来换取高的数值稳定性，从而得到较高的计算精度。

的取值可以通过**交叉验证**的方式获取。

**Lasso回归(L1正则化)**

，求最小值得

**评估方法**

RMSE 指的是**均方根误差**：



MSE 指的是**均方误差**：



MAE 指的是评价**绝对误差**：



R Squared也叫**R方**：



如果结果是0，就说明我们的模型跟瞎猜差不多。  
如果结果是1。就说明我们模型无错误。  
如果结果是0-1之间的数，就是我们模型的好坏程度。  
如果结果是负数。说明我们的模型还不如瞎猜。（其实导致这种情况说明我们的数据其实没有啥线性关系）

R-squared（值范围0-1）描述的 输入变量对输出变量的解释程度。在单变量线性回归中R-squared 越大，说明拟合程度越好。

然而只要曾加了更多的变量，无论增加的变量是否和输出变量存在关系，则R-squared 要么保持不变，要么增加。

So, 需要adjusted R-squared ,它会对那些增加的且不会改善模型效果的变量增加一个惩罚向。

结论，**如果单变量线性回归，则使用 R-squared评估，多变量，则使用adjusted R-squared**。

在单变量线性回归中，R-squared和adjusted R-squared是一致的。

另外，如果增加更多无意义的变量，则R-squared 和adjusted R-squared之间的差距会越来越大，Adjusted R-squared会下降。但是如果加入的特征值是显著的，则adjusted R-squared也会上升

# 决策树

# 神经网络

# 支持向量机

# 贝叶斯分类

# 集成学习

# 聚类

# 降维与度量学习

# 特征选择与稀疏学习

# 计算学习理论

# 半监督学习

# 概率图模型

# 规则学习

# 强化学习