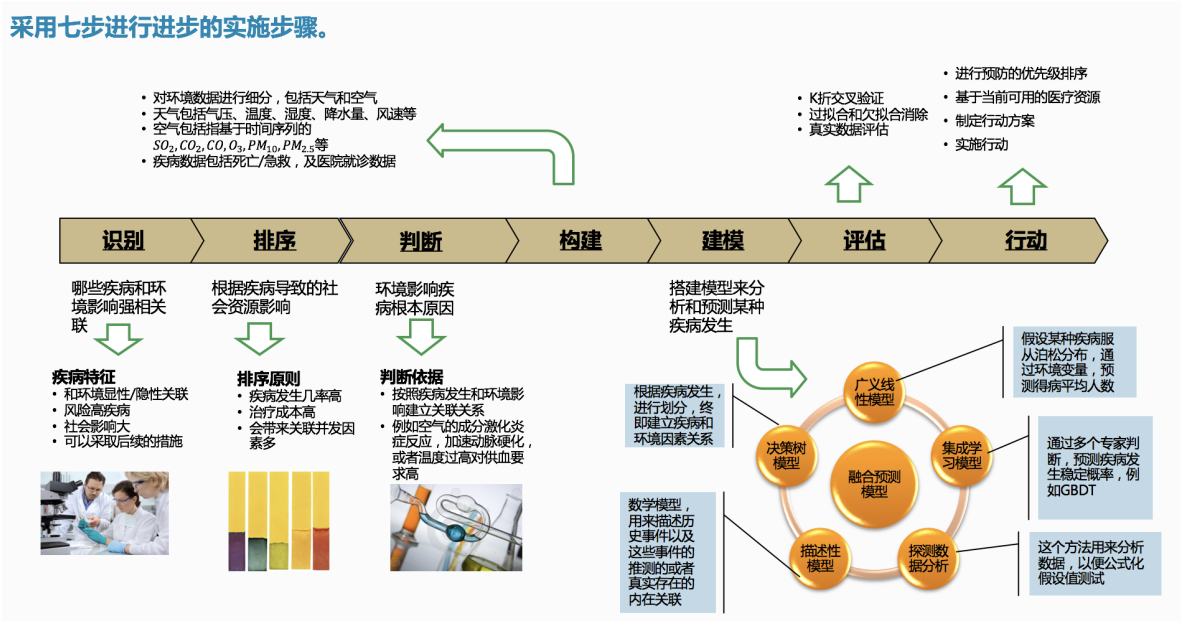
目标2：构建基于环境污染相关的呼吸系统或心血管系统疾病风险预测模型



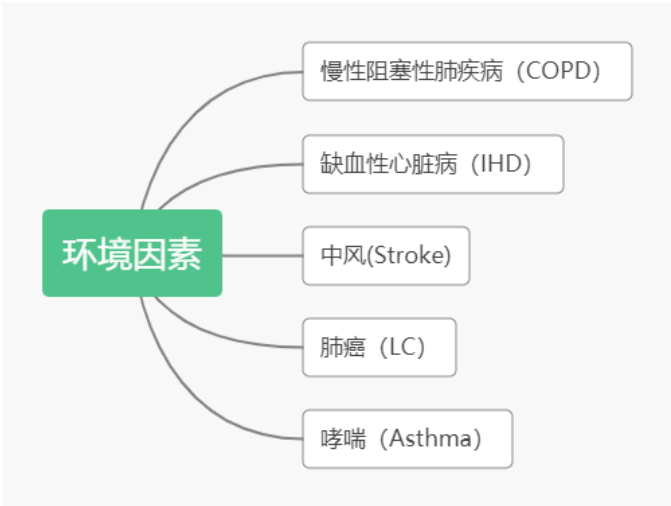
1. 目的

（1）通过环境等因素预测某类人群患呼吸系统或心血管系统疾病的风险。

（2）通过建模找到增大某类人群患呼吸系统或心血管系统疾病风险的空气污染物及气象因素。

2. 背景介绍

基于近五年相关文献的研究发现以下五种疾病和环境因素相关度比较高。其中COPD和IHD在多篇论文中作为研究对象。



以COPD为例，大量研究表明，环境细颗粒物$PM_{2.5}$(直径<2.5um)与慢性阻塞性肺疾病（COPD）的发病率和死亡率增加相关。但是，对于PM 2.5如何诱导并进一步加重COPD发育和进展的这些作用的潜在机制知之甚少[2]。其中一种解决方案是从数据中提取特征并且预测结果，便于为大家提供更好的预防措施。

3. 数据

- 所用数据
  - 患呼吸道疾病或者心血管疾病的病人资料（如性别，年龄，就诊医院，常住地址，其他基础疾病等等）
  - 急救数据：ICD编码 ，呼救地址，初步诊断（疾病类型如创伤类），出生日期，年龄，接诊日期
  - 死因数据：出生日期，生前常住地址类型
  - 空气质量数据
  - pm2.5成分数据
  - 气象数据
- 数据处理
  - 清洗所有的数据表
  - 连接不同的数据表
- 特征选取
  - 处理相关性高的变量
  - 选择与所研究目标相关度高的变量
- 样本复杂度估计：

  本次研究目标为全体广州市民以及所有气象数据M, 目前我们使用的训练数据为m（有关病人和气象的各项参数）。事实上，抽样的数据永远无法与真实数据100%一致。

  那么，m到底在多大程度上能够准确代表M？为了得到答案，需要对训练数据m进行可行性估计。泛化性能通常由两个参数定义：ε误差容限，即允许的泛化误差；δ置信度参数，确定错误的频率。

  假设模型为N(x)，可知$\mu = N(M)$为全体广州市民真实的疾病风险指数（无法直接得到），而$v = N(m)$为使用现有抽样数据得到的疾病风险指数。根据霍夫丁不等式, ∀ε > 0:

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 m}。$$

可以观察到，随着m的增长，|ν−μ|>ε变得越来越不可能 。为了以1 −δ的概率获得最多ε的泛化误差，我 们必须让使用m个训练示例的可信度高于1 −δ：

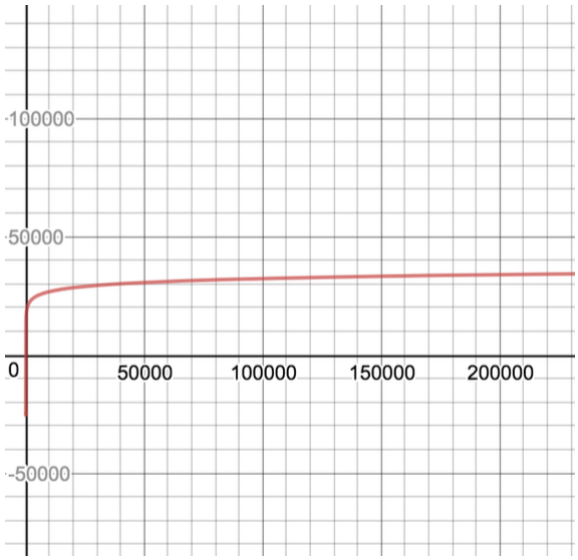
$$P[|\mu \leq v + \epsilon] \geq 1 - \delta$$

利用Vapnik-Chervonenkis维度对霍夫丁公式的进一步推导与简化，得到：

$$m \geq \frac{8}{\epsilon^2} ln \frac{4((2m)^{vc} + 1)}{\delta},$$

其中，vc是Vapnik-Chervonenkis维度。需要使用迭代方法（即从0开始增加m直到m收敛）获得m的数值。

例如在VC等于3的模型评估情况下，我们希望泛化误差最大为0.1，置信度为90%：



上图横坐标为迭代次数x，纵坐标为m，可以在上例中观察样本数m在3万附近趋近稳定，即至少需要3万的样本数据。

4. 研究方法

简要描述

由于医院患者总体比较稳定，可以将其作为整体研究对象，通过建立疾病与环境的关系来预测患某种疾病的人数及百分比。

其他可以参考的指标如下。

指标（1）：

根据某种疾病死亡率推导出某种特定人群早死所致的寿命损失年数(Years of Life Lost，**YLL**)并和全国平均、同期平均比较，作为某种疾病的总体风险指数：

$$YLL = \sum_{i=1}^{N} a_i \times d_i$$

其中，a_i为当死亡发生在年龄组i和i+1之间时，存活到该区域预期寿命的剩余寿命年数；d_i为代表在i和i+1岁之间调查人口中观察到的死亡人数。即每10万人中，实际死亡年数与低死亡人群中该年龄的预期寿命之差。

指标（2）：

某种疾病死亡比例与YLL（如肺癌与全国平均的比值），作为该疾病相对风险指数：

$$RR = n\frac{\Delta R}{\bar{R}},$$

其中，n是修正参数。右边比值大于1说明该疾病高于全国平均。

指标（3）：

该疾病在某群体中的死亡/确诊率和全市人群/全国的平均值比较并取对数，作为该疾病针对某一群体的风险系数，将该系数作为特定群体的风险指数：

$$R_i = e^{\frac{a_1 D + a_2 E}{a_1 \bar{D} + a_2 \bar{E}}},$$

其中，D为死亡率，E为确诊率。

4.1 模型

变量

- X_i：对某类人群患呼吸道疾病或者心血管类疾病产生影响的变量i的值，例如某污染物的浓度值
- Y：患某类疾病的人数

在本目标中，模型输出的结果将是某类人生某种病或由某种疾病致死的人数，据此可计算出其占全市人口的比例。

模型一：广义线性模型(generalized linear model, GLM)

根据数据的特性使用广义线性模型中的一种特殊模型 - 对数线性模型（log linear model）。假设某类人得某种疾病的数量服从泊松分布，则模型可表示为

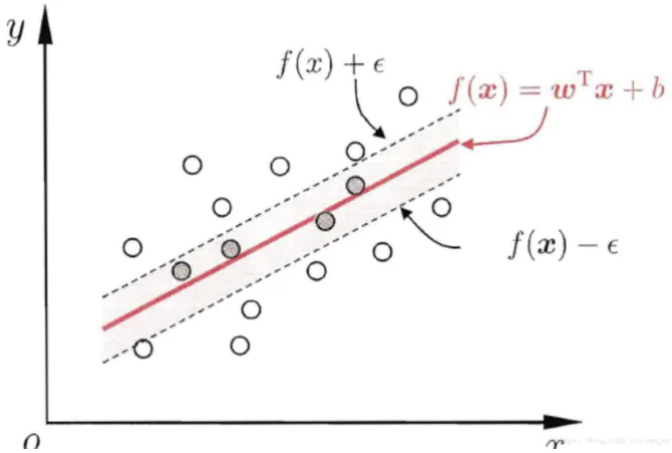
$$Y|X;\mu \sim Poisson(\mu)$$
$$log(\mu) = \sum_{i=1}^{n} \beta_i x_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

通过环境变量等，此模型可以预测出某类人得某种疾病的平均人数。

模型二：支持向量机-回归(SVR)

支持向量机也可以用作回归方法，支持向量机(SVM)本身是针对二分类问题提出的，而支持向量回归（SVR）是支持向量机（SVM）中的一个重要的应用分支。支持向量回归所寻求的最优超平面是使所有的样本点离超平面的“距离”最小。[3] 只要$f(x)$与y偏离程度在阈值ε的范围内，即可认为预测正确。如下图中，红色线为预测值，小圆圈为观测值，所有在虚线之间的都被认为是预测正确了。$f(x)$的优化可通过计算虚线外的损失$|f(x)-y|>\varepsilon$得到。

例子：假设x为$PM_{2.5}$的平均浓度值，某类人患某种疾病的预测人数$f(x)$即为下图橙色直线。



模型三：决策树（Decision Tree）

决策树就是一种基本的分类与回归方法，若它的内部结点特征的取值为“是”和“否”， 为二叉树结构。决策树的关键在于两个：划分点和输出均值，即如何确定每一个节点的位置以及划分后各区域的输出（以及最终划分的区域输出）。经典的一维回归决策树采用均方误差（MSE）和均值分别作为划分依据和输出值，即找到两边MSE之和最小的点作为每一次划分节点的标准：

$$min[\sum_{R_1(j,s)}(y_i - c_1)^2 + \sum_{R_2(j,s)}(y_i - c_2)^2],$$

其中，

$$c_1 = \frac{1}{N1}\sum_{R_1(j,s)} y_i, c_2 = \frac{1}{N2}\sum_{R_2(j,s)} y_i$$

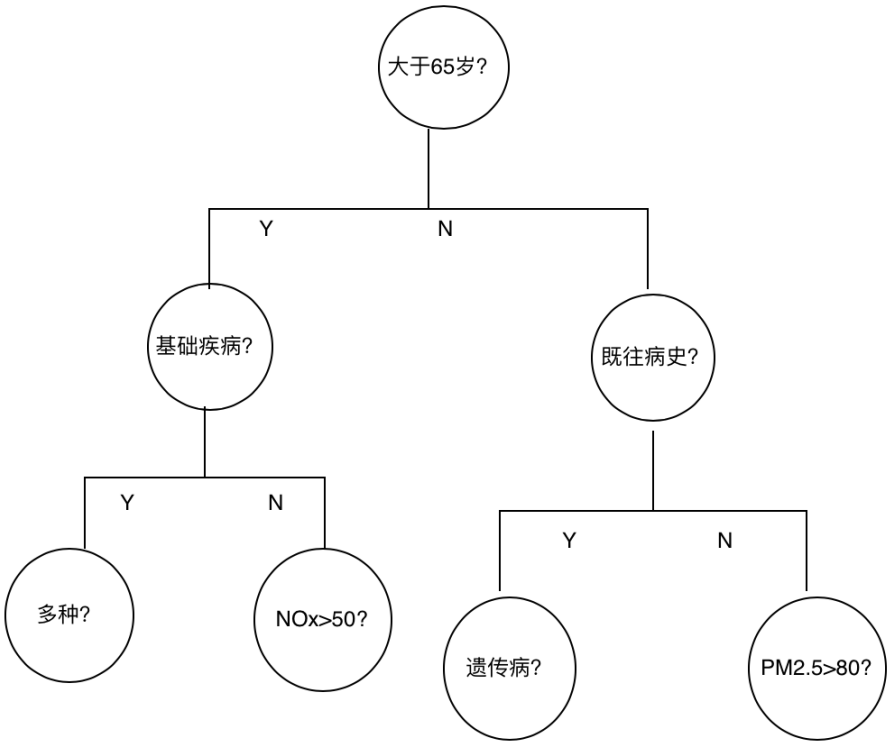
$R_1(j,s)=\{x|x_j \leq s\}$和$R_2(j,s)=\{x|x_j > s\}$是被划分后的两个区域，$c_1$、$c_2$是两区域的输出值。

如何判断某类人得呼吸道或心血管类疾病的风险（由生病或死亡人数综合计算得出）？举例如下：

性别	年龄	常住地址所在区域	其他基础疾病	患某疾病的预测人数	因某疾病死亡的预测人数
女	>65	区域1	有	x_1	y_1
男	30-40	区域2	无	x_2	y_2
男	<5	区域2	无	x_3	y_3
女	<5	区域1	有	x_4	y_4

$X=\{x_1,x_2,\dots x_n\}$ ，　$Y=\{y_1,y_2,\dots y_n\}$

我们可以通过X和Y的两组数据附加权重综合计算，某类人对于某种疾病的综合风险。



上图是对二维平面划分的决策树，右边为对应的划分示意图。划分的过程也就是建立树的过程，每划分一次，随即确定划分单元对应的输出，也就多了一个结点。当根据停止条件划分终止的时候，最终每个单元的输出也就确定了，也就是判断的结果。在本目标中，模型输出的结果将是生病和死亡的人数，以及占全市人口的比例。

模型四：集成学习法

对于稳定预测将某类人患呼吸道或者心血管疾病的概率这一目标，在机器学习中，简单模型的预测效果可能达不到预期的效果，比如在线性回归中，实际问题大多不是线性的，很容易发生过拟合或者欠拟合；在决策树模型中，如果让决策树无限制的增长，会达到训练样本的100%正确率，但这完全过拟合了，没有实际意义，虽然可以加入剪枝或者树深度的的限制，但是还是会比较容易造成过拟合。

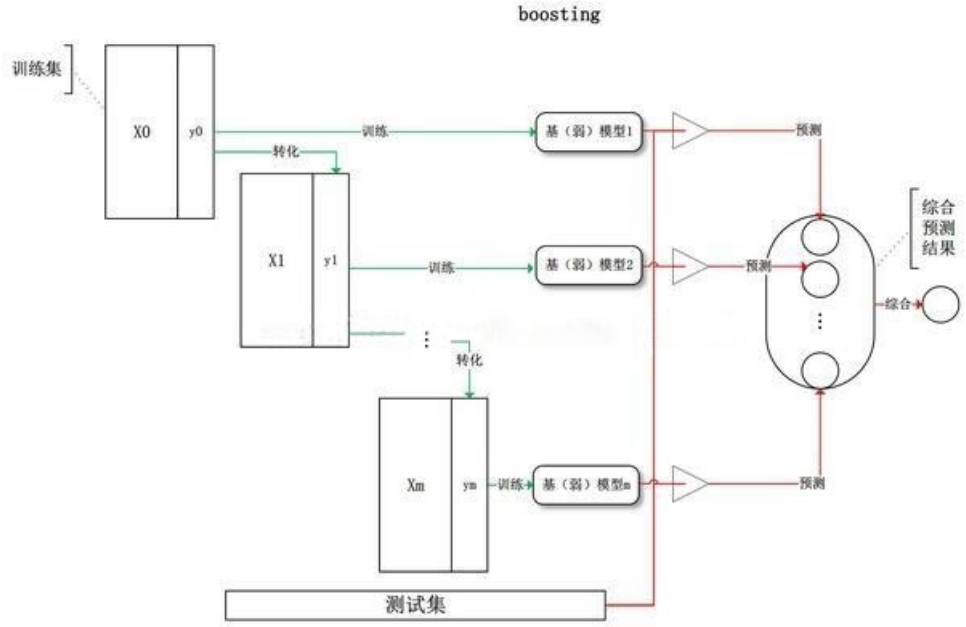
对于一个复杂任务来说，将多个专家的判断进行适当的综合所得出的判断，要比其中任何一个专家单独的判断要好。集成学习方法主要有两种，Bagging和Boosting：

- Bagging：可以理解为用好多个模型都预测然后进行投票，票数多的预测结果作为正式的结果；
- Boosting则可以理解为一个弱模型学习之后用另一个弱模型学习上一个弱模型没有学习到的部分，然后合成一个强模型。

在树模型中，使用Bagging方法的代表是随机森林算法，该算法就是使用多个决策树的结果进行投票；而用Boosting方法的树模型框架是梯度增强决策树（Gradient Boosting Decision Tree，GBDT），该模型使用树模型学习前一棵树模型的泛化误差，最终将所有树的结论累加起来作为结果的算法框架，该模型泛化能力（generalization)较强的算法且不容易发生过拟合。

梯度增强决策树(GBDT)

假设实际观测值为y_i，模型的预测值为$\hat{y}_i$，那下一个模型就会尽量减少残差$y_i - \hat{y}_i$。假设在残差减少的方向上建立了k个模型，GBDT算法可以看作是这k个模型的加法模型：$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$。



4.2 模型评估

k折交叉验证（k-fold Cross Validation）

交叉验证（Cross Validation），是一种统计学上将数据样本切割成较小子集的实用方法。在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。这个过程一直进行，直到所有的样本都被预报了一次而且仅被预报一次。把每个样本的预报误差平方加和，称为PRESS(predicted Error Sum of Squares)。

$$P = \sum_{i=1}^{N} (\bar{p} - p_i)^2$$

假定根据某模型的混淆矩阵(confusion matrix)如下

		预测值(Predicted Value)		计数
		Postive	Negtive	
观测值 (Observed Value)	Positive	TP (True Positive)	FN (False Negtive)	P
	Negtive	FP (False Positive)	TN (True Negtive)	N

对于分类器或者说分类算法，评价指标主要有正确率(precision)，召回率(recall)，F1 score等，以及ROC和AUC。以下为相关专有名词的介绍。

正确率(Precision)：

$$Precision = \frac{TP}{TP + FP}$$

召回率(Recall)，又名真阳性率(True Positive Rate，TPR)，灵敏度(Sensitivity)：

$$Sensitivity = Recall = TPR = \frac{TP}{TP + FN}$$

特异度(Specificity)，又名真阴性率(True Negative Rate，TNR)：

$$Specificity = TNR = \frac{TN}{FP + TN}$$

漏诊率(= 1 - 灵敏度)，又名假阴性率(False Negatice Rate，FNR)：

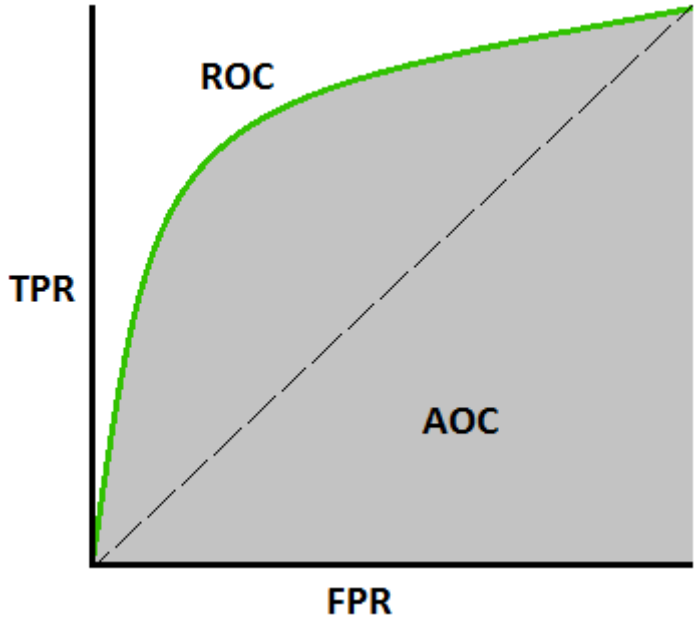
$$FNR = \frac{FN}{TP + FN}$$

误诊率(= 1 - 特异度)，又名假阳性率(False Positice Rate，FPR)：

$$FPR = \frac{FP}{FP + TN}$$

ROC曲线(Receiver Operating Characteristic Curve)：假阳性率(FPR)为横坐标，真阳性率(TPR)为纵坐标构成的曲线。

AUC(Area Under Curve)被定义为ROC曲线下的面积，显然这个面积的数值不会大于1。当某分类器的AUC越接近于1的时候，这个分类器的效果越好。ROC和AUC如下图所示。



例子：第二次世界大战时用雷达技术解析雷达的信号，有时候会把敌军轰炸机解析成飞鸟，有时候会把大鸟当作敌军轰炸机。假设有10个雷达信号报警，其中8个市真的轰炸机（P），2个是大鸟（N）。某个分析员解析出9个轰炸机和1个大鸟。其中有1个被判定为轰炸机的其实是大鸟（FP=1），剩下的确实是轰炸机（TP=8）。因此FPR为0.5，TPR为1，(0.5, 1)就对应ROC曲线上的一个点。目标是能把所有的地方轰炸机都预测出来，但也同时不希望将大鸟预测成轰炸机。所以目标是FPR越低越好并且TPR越高越好。

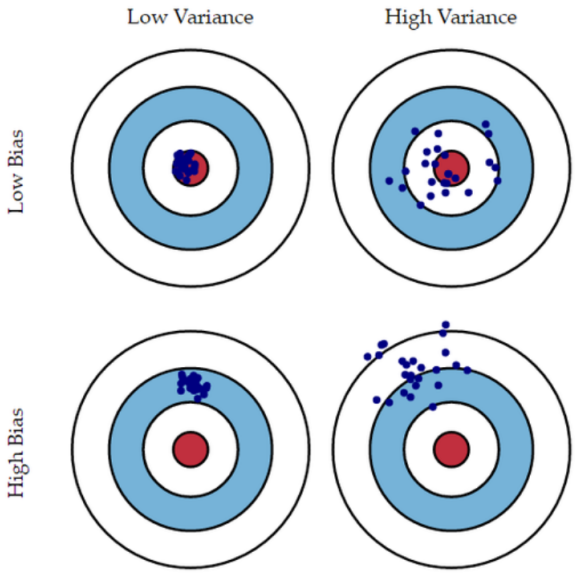
过拟合(overfitting)与欠拟合(underfitting)

- 过拟合

  过拟合成因是对于给定数据集，模型过于复杂、对于已有数据拟合能力过强，常会导致方差(variance)过大并且对新数据预测能力较低。

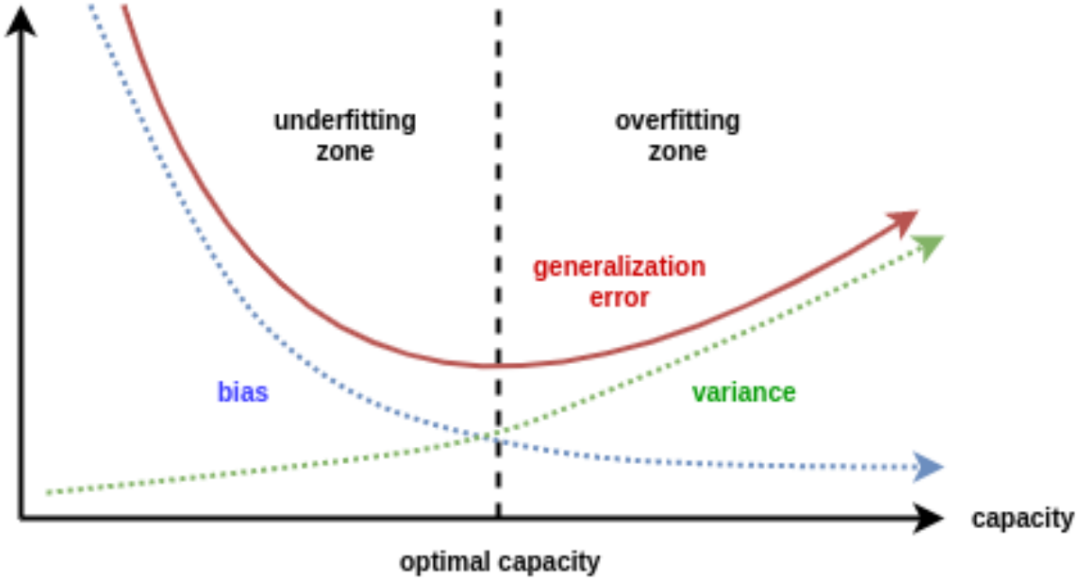
- 欠拟合

  对于给定数据集，欠拟合的成因大多是模型不够复杂，拟合函数的能力不够。这常会导致偏差(bias)过大。



判断方法：

从训练集中随机选一部分作为一个**验证集**。在缺少有效预防欠拟合和过拟合措施的情况下，随着模型拟合能力的增强，错误率在训练集上逐渐减小，而在验证集上先减小后增大；当两者的误差率都较大时，处于欠拟合状态(high bias, low variance)；当验证集误差率达到**最低点**时，说明拟合效果最好，由最低点增大时，处与过拟合状态(high variance, low bias)。



参考文献

1. Salvi SS, Barnes PJ. Chronic obstructive pulmonary disease in non-smokers. Lancet.2009;374:733–43
2. Role of PM2.5 in the development and progression of COPD and its mechanisms
3. Leo Breiman, "Bagging Predictors", Technical Report 421, September 1994, Department of Statistics, University of California Berkeley, CA Also at anonymous ftp site: ftp.stat.berkeley.edulpub/tech-reports/421.ps.Z.
4. Yanyan Chen, Yuanyuan Song, Yi-Jie Chen, etc. Contamination profiles and potential health risks of organophosphate flame T retardants in PM2.5 from Guangzhou and Taiyuan, Chin，Environment International 134 (2020) 105343
5. Sourangsu Chowdhury, Sagnik Dey, Cause-specific premature death from ambient PM2.5 exposure in India: Estimate adjusted for baseline mortality, Environment International 91 (2016) 283–290
6. Xingcheng Lu , Changqing Lin , Ying Li , etc. Assessment of health burden caused by particulate matter in southern China using high-resolution satellite observation, Environment International 98 (2017) 160–170
7. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (PDF). Pattern Recognition Letters. 27 (8): 861–874.
8. Ting, Kai Ming (2011). Encyclopedia of machine learning. Springer. ISBN 978-0-387-30164-8.
9. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). Journal of Machine Learning Technologies. 2 (1): 37–63.