机器学习介绍

目录

CONTENTS

PART 01

机器学习概述

机器学习概述

机器学习是通过一系列计算方法（简称“算法”），使得**计算机具备从大数据中学习的能力**。

机器学习实现的过程是，用户将**既有数据**提供给计算机，计算机基于既有数据，**使用机器学习算法构建模型**，然后将**模型推广泛化到新的样本观测值**，进而可以进行**预测**。

所以，一言以蔽之，**机器学习的内容体现为“算法”，精髓在于预测**。本书介绍的机器学习知识，也围绕各种“算法”展开，而**判断算法或模型的优劣标准则是预测能力的高低。**

既有数据可以是**历史积累的真实数据**，比如电子商务平台商家积累的用户信息及交易数据；也可以是人们**基于经验或规定创造的虚拟数据**，比如商业银行对公信贷中将“受托支付贷款资金回流借款人”的业务判定为存在高合规风险，那么就可以虚拟一些贷款业务形成数据输入计算机，使得计算机能够习得这一规则。

PART 02

机器学习术语

机器学习术语

常用的机器学习术语包括示例、响应变量、特征、属性值、属性空间、特征向量、训练样本、测试样本等。

示例：即统计学或计量经济学中的样本观测值，也被称为“样例”，比如前面所述的每一个客户的财务报表指标及财务表现数据。

响应变量：即被解释、被影响的目标变量，也被称为“目标”，可以理解成统计学或计量经济学中的“因变量”“被解释变量”，在数学公式中常用y来表示。比如前面所述的客户财务表现结果。

特征：即用来解释、影响响应变量的变量，也被称为“预测变量”“属性”等，可以理解成统计学或计量经济学中的“因子（离散型变量）”“协变量（连续性变量）”“解释变量”，在数学公式中常用X来表示。比如前面所述的客户具体财务指标（“经营性现金流量净收支”“营业收入增长率”等）。

机器学习术语

属性值：**特征的取值即为属性值**，比如前面所述的“经营性现金流量净收支”“营业收入增长率”等财务指标的具体取值。

属性空间：**多个属性形成的空间称为属性空间，属性的个数也被称为空间的维数**，比如我们仅考虑“资产负债率”“经营性现金流量净收支”“营业收入增长率”三个属性，将三个属性分别作为x轴、y轴、z轴，则三个属性就构成了一个三维属性空间，每个企业每期的财务指标都可以在三维属性空间找到对应的点。

特征向量：属性空间的每个点都会对应一个特征向量，“特征向量”的名称正来自于此。比如某示例特征向量为（75.6%，3600,10.23%）表示其“资产负债率”“经营性现金流量净收支”“营业收入增长率”分别为“75.6%”“3600”“10.23%”。

机器学习术语

训练样本：即计算机用来**应用算法构建模型**时使用的样本。

测试样本：即计算机用来**检验机器学习效果，检验外推泛化应用能力**时使用的样本。有的模型可能在基于训练样本的预测方面有着卓越表现，但在测试样本方面表现差强人意，反映出泛化能力不足（关于“泛化”的概念在下文详细介绍）。

PART 03

机器学习分类

机器学习分类

根据输入数据是否具有**“响应变量”**信息，机器学习被分为**“监督学习”和“无监督学习”**。

“监督学习”即输入数据中**既有X变量，也有y变量**，特色在于**使用“特征（X变量）”来预测“响应变量（y变量）”**。前面介绍的客户财务报表指标及财务表现甄别学习，因为是基于客户财务报表指标预测财务表现类别，即为典型的“监督学习”。

如果响应变量（y变量）为分类变量，比如信贷资产五级分类（正常、关注、次级、可疑、损失）或客户信用评级（AAA，AA……，C,D），则又可进一步成为**“分类问题监督学习”**；**如果响应变量（y变量）为连续变量**，比如计算客户最大债务承受额，则又可进一步称为**“回归问题监督学习”**。

“无监督学习”即算法在训练模型时期不对结果进行标记，而直接在数据点之间找有意义的关系，或者说**输入数据中仅有x变量而没有y变量**，特色在于**针对x进行降维或者聚类，以挖掘特征变量自身特征**。

机器学习分类

“监督学习”和“无监督学习”只是常见的机器学习分类，除了这两种之外，还有**半监督学习、强化学习**等学习方式。

猜一猜：“阿尔法狗”（AlphaGo）是什么学习方式？

“监督学习”的模型优劣及应用场景很好理解。模型优劣评价方面，把x变量值输入后，通过机器学习算法构建模型得到y变量拟合值，将其与y变量实际值进行对比，即可检验模型的优劣。其中针对“回归问题监督学习”，比较的是**y变量拟合值与实际值的数量差异**；针对“分类问题监督学习”，则比较的是**y变量拟合的分类和实际的分类**。应用场景方面，比如根据目标客户的基本信息、交易信息等特征，预测客户的价值贡献度（回归问题）、是否购买新产品（分类问题），进而制定针对性的**市场营销**策略；又比如根据目标客户的基本信息、财务信息、负债及对外担保信息等预测违约概率（回归问题）、进行信用评级（分类问题），进而制定针对性的**风险防控**策略。

机器学习分类

“无监督学习”由于目标不明确，所以其效果很难评估，其价值在于发现模式以及相关性。如果从特征（变量）的角度来看，价值体现在对变量进行降维，从而可以有助于解释变量之间的关系，或降低模型的复杂程度；如果从样本的角度来看，价值体现在可以研究个体之间的关系，将相近的个体划分在一起。比如可以用于商业银行的反洗钱领域或员工行为管理，通过“无监督学习”把行为或个体快速进行分类，即使我们可能无法清楚的知晓分类意味着什么，但是可以快速区分出正常、异常的行为或个体组，从而为深入分析做好准备，显著提升分析效率。又比如在搜索引擎中，我们基于用户特征把用户快速聚类，可精准实施广告投放或偏好信息推送。再比如在电商平台中，系统针对具有相似购买行为的用户，推荐合适的产品，A用户和B用户为1类，若A用户购买了某产品，B用户大概率也会购买该产品，可将该产品推送给B产品，实现精准推荐等等。

PART 04

误差、泛化、过拟合与欠拟合

误差、泛化、过拟合与欠拟合

机器学习中样本的预测值和实际值之间的差异被称作“误差”，其中基于训练样本的误差又被称为“训练误差”或“经验误差”，基于新样本的误差又被称为“泛化误差”。

其中“训练误差”或“经验误差”反映的是机器学习对既有数据的学习能力。基于训练样本得到的机器学习模型向新样本推广应用的能力，或者说模型的预测能力，称为模型的“泛化”能力。所以从致力于实现“泛化”能力最强的角度考虑，经验误差并未越小越好。如果经验误差过小，说明机器学习能力有可能“过强”，也就很可能意味着计算机不仅学习了训练样本的一般性、规律性特征，在很大程度上也学习了训练样本的个性化特征，而这些个性化特征往往并不能很好的泛化到新的样本，不仅白白增加了模型复杂度和冗余度，也无法很好的开展预测，甚至模型是否可用都待商榷，这一现象也被称为“过拟合”。

误差、泛化、过拟合与欠拟合

当然，经验误差也不能很大，如果经验误差很大，说明机器学习能力不够，意味着**没有充分利用训练样本信息，没有充分挖掘出训练样本的一般性、规律性特征，从而也不能很好的泛化到新的样本，这一现象也被称为“欠拟合”**。

“泛化误差”反映的模型的“泛化”能力，“泛化误差”越小，模型“泛化”能力越强。我们之所以开展机器学习，目的是为了基于既有数据来预测未知，以期进一步改善未来商业表现，所以从应用的角度，**我们主要关注的是泛化误差而不是经验误差**，如果某种机器学习模型比另一种具有更小的泛化误差，那么这种模型就相对更加有效。

PART 05

偏差、方差与噪声

偏差

偏差度量的是，学习算法的**期望预测与真实结果**的偏离程度，反映的是学习算法的拟合能力。

$$\mathrm{Bias}(\hat{f}(x)) = E\hat{f}(x) - f(x)$$

高偏差意味着期望预测与真实结果的偏离度大，也就是学习算法的拟合能力差；相应的，低偏差意味着期望预测与真实结果的偏离度小，也就是学习算法的拟合能力强。

偏差产生的原因通常包括：**一是选择了错误的学习算法**，比如真实为非线性关系，但模型为线性算法；**二是模型的复杂度不够**，比如真实为二次线性关系，但模型为一次线性算法。

一般来说，线性回归，线性判别分析和逻辑回归等线性机器学习算法因为局限于线性，会导致无法从数据集中学习足够多的知识，针对复杂问题预测性能较低，偏差相对较高，而具有较大灵活性的非线性机器学习算法如决策树，KNN和支持向量机等机器学习算法的偏差相对较低。

方差

方差度量的是，在大量重复抽样过程中，**同样大小的训练样本的变动导致的学习性能的变化**，反映的是**数据扰动所造成的影响**，也就是模型的**稳定性**。

$$\mathrm{Var}(\hat{f}(x)) = E[\hat{f}(x) - E\hat{f}(x)]^2$$

方差越小意味着模型越为稳定、在未知数据上的泛化能力越强，但由于目标函数是由机器学习算法从训练样本中得出，所以算法具有一定方差的事实不可避免。高方差意味着同样大小的训练样本的变化对目标函数的估计值会造成较大的变动，容易受到训练样本细节的强烈影响；相应的，低方差意味着同样大小的训练样本的变化对目标函数的估计值会造成较小的变动。

方差的出现原因往往是由于**模型的复杂度过高**，比如比如真实为一次线性关系，但模型为二次线性算法。与前面介绍的偏差恰好相反，一般来说，线性回归，线性判别分析和逻辑回归等线性机器学习算法方差相对较低，而人工神经网络、决策树，KNN和支持向量机等非线性机器学习算法的方差相对较高。

噪声

噪声度量的是针对既定学习任务，使用任何学习算法所能达到的期望泛化误差的最小值，属于不可约减误差，反映的是学习问题本身的难度，或者说是无法用机器学习算法解决的问题。噪声大小取决于数据本身质量，当数据给定时，机器学习所有达到的泛化能力的上限也就确定。

$$\mathrm{Noise}(\hat{f}(x)) = E(\varepsilon^2)$$

误差与偏差、方差、噪声的关系

从数学的角度理解，“误差”就是学习得到的模型的期望风险，对于有监督学习，“误差”使用MSE（均方误差）任何学习算法的“均方误差”都可分解为“偏差”“方差”“噪声”三者之和，这一点已有严格的数学公式所证明。

偏差与方差的权衡

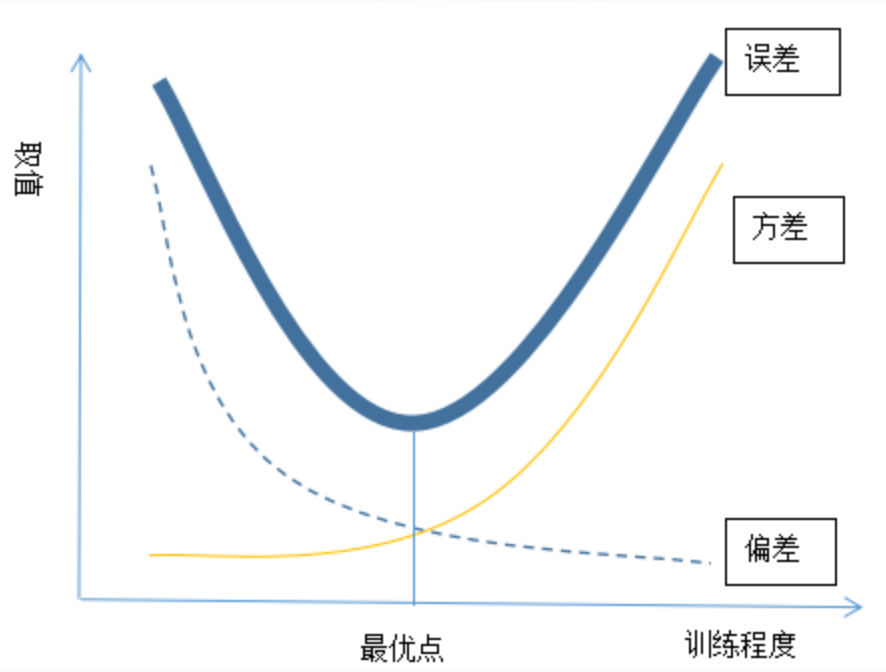
从上面介绍也可以看出，偏差和方差之间存在选择两难。

低的偏差意味着高的方差，或者说模型的灵活性增强就会导致稳定性下降；

同时低的方差意味着高的偏差，或者说模型的稳定性增强就会导致灵活性下降。

偏差与方差的权衡

如图所示，横轴代表训练程度，纵轴代表取值，对于既定的机器学习任务，假定用户可以控制其训练程度，当训练程度比较低时，意味着学习不够充分，导致偏差会比较大，然而正是因为学习不够充分，训练样本的变化对于模型的扰动影响也不构成显著性，方差会比较小；而随着训练程度的不断增加，学习越来越充分，偏差会越来越小，然而也正是因为学习饱和度的上升，训练样本的变化对于模型的扰动影响也逐渐增加，越来越显著。在中间位置会有一个泛化误差最小的最优点（注意不一定是方差线和偏差线的交点），在最优点上，模型的泛化误差最小，泛化能力最强。



PART 06

性能度量

性能度量

性能度量是指衡量机器学习算法模型的评价标准。

针对“监督学习”，性能度量的评估方式是机器学习的预测能力，也就是基于机器学习获得的拟合值与真实值之间的差异，“回归问题监督学习”的差异反映在数值的大小，“分类问题监督学习”的差异反映在分类的表现。

针对“无监督学习”，以聚类分析为例，性能度量则为聚类结果中同一类别内部各个示例的相似度，以及不同类别示例间的不相似度。**好的“无监督学习”算法应该是组内相似度高，而组间相似度低。**

“回归问题监督学习”的性能度量

针对“回归问题监督学习”，最常用的性能度量指标为“均方误差”（mean squared error）。

假设示例集为 $D=\{(x_1,y_1),(x_2,y_2)\ldots\ldots(x_k,y_k)\}$，其中$(x_i,y_i)$为各个示例，$x_i$为属性值，$y_i$为响应变量的真实值。“均方误差”的数学公式为：

$$E(f;D)=\frac{1}{k}\sum_{i=1}^{k}(f(x_i)-y_i)^2$$

“分类问题监督学习”的性能度量　一、错误率和精度

针对“分类问题监督学习”，性能度量最简单的就是观察其预测的错误率和正确率，用到的性能度量指标即为错误率和精度。

其中错误率即为预测错误的比率，也就是预测类别和实际类别不同的示例数在全部示例中的占比；精度即为预测正确的比率，也就是预测类别和实际类别相同的示例数在全部示例中的占比。

基于上述定义，不难看出错误率和精度之和等于1，或者说错误率=1−精度。

“分类问题监督学习”的性能度量 二、查准率、查全率（召回率）、F1

在“分类问题监督学习”中，我们除了观察预测的正确率、错误率，很多情形下，我们需要特别关心特定类别被查找是否准确，比如针对员工行为管理中的异常行为人员界定，可能需要特别审慎，非常忌讳以莫须有的定性伤害员工的工作积极性，分类为异常行为的准确性非常重要，这时候就需要用到“查准率”的概念；也有很多情形下，我们需要特别关心特定类别被查找是否完整，比如针对某种传染性极强的病毒，核酸检测密切接触者其是否为阳性，对阳性病例的查找的完整性就显得尤为重要，这时候就需要用到“查全率”的概念。

如果“分类问题监督学习”为二分类问题，我们会很容易得到如下图所示的分类结果矩阵，也称为“混淆矩阵”（confusion matrix）。其中的“正例”通常表示研究者所关注的分类结果，比如授信业务发生违约，所以并不像字面意思那样必然代表正向分类结果；“反例”则是与“正例”所对应的分类，比如前述的授信业务不发生违约。

“分类问题监督学习”的性能度量 二、查准率、查全率（召回率）、F1

在“混淆矩阵”中：

样本示例		机器学习预测分类	
		正例	反例
真实分类	正例	TP 真正例 (true positive)	FN 假反例 (false negative)
	反例	FP 假正例 (false positive)	TN 真反例 (true negative)

当样本示例真实的分类为正例，且机器学习预测分类也为正例时，说明机器学习预测正确，分类结果即为TP（真正例）。

当样本示例真实的分类为反例，且机器学习预测分类也为反例时，同样说明机器学习预测正确，分类结果即为TN（真反例）。

当样本示例真实的分类为正例，且机器学习预测分类为反例时，说明机器学习预测错误，分类结果即为FN（假反例）。

当样本示例真实的分类为反例，且机器学习预测分类为正例时，说明机器学习预测错误，分类结果即为FP（假正例）。

$$查准率=\frac{TP}{TP+FP}$$

$$查全率=\frac{TP}{TP+FN}$$

“分类问题监督学习”的性能度量 二、查准率、查全率（召回率）、F1

类似于统计学中的“第一类错误”（拒绝为真）和“第二类错误（接受伪值）”，查准率和查全率之间也存在两难选择问题，如果我们要获得较高的查准率，减少“接受伪值”错误的发生，往往就需要在正例的判定上更加审慎一些，仅选取最有把握的正例，也就意味着会将更多实际为正例的样本“错杀误判”为反例样本，造成“拒绝为真”错误的增加，也就是查全率会降低。按照同样的逻辑，如果我们要获得较高的查全率，减少“拒绝为真”错误的发生，往往就需要在正例的判定上更加包容一些，也就意味着会将更多实际为反例的样本“轻信误判”为正例样本，造成“接受伪值”错误的增加，也就是查准率会降低。

注意：查全率也称“召回率”“灵敏度”“敏感度”。

为了平衡查准率与查全率，使用F1值。F1值为精确率和召回率的调和平均值。F1的取值在0-1之间，数值越大表示模型效果越好。

$$F1=\frac{查准率*查全率*2}{查准率+查全率}$$

“分类问题监督学习”的性能度量　　三、累积增益图

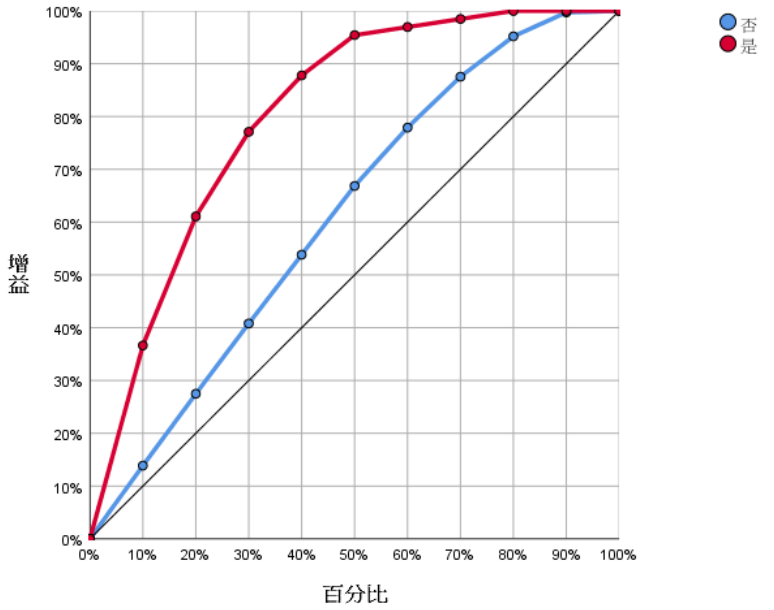
前面提到“分类问题监督学习”中查准率、查全率存在两难选择的问题，用户可以使用累积增益图，进行辅助决策。在针对二分类问题的很多机器学习算法中，系统对每个示例都会预测1个针对目标类别的概率值p，如果p大于0.5，则将会判定为目标类别，如果p小于0.5，则判定为非目标类别。

根据概率值p，用户将所有示例进行降序排列，拥有大的p值的示例将会被排在前面。累积增益图会在给定的类别中显示通过把个案总数的百分比作为目标而“增益”的个案总数的百分比。

下面以某商业银行授信业务违约预测累积增益图为例进行解释，其中“是”表示“违约”，“否”表示“不违约”。

“分类问题监督学习”的性能度量　　三、累积增益图

“是” 类别曲线上的第一点在（10%，37%），即如果用户使用机器学习模型对数据集进行预测，并通过 “是” 预测拟概率p值对所有示例进行排序，将会期望预测拟概率p值排名前 10%的个案中，含有实际上类别真实为“是”（违约）的所有个案的37%。同样，“是” 类别曲线上的第二点在（20%，61%），即预测拟概率p值排名前 20%的个案包括约 61% 的违约者，“是” 类别曲线上的第三点在（30%，77%），即预测拟概率排名前 30% 个案包括 77% 违约者。依此类推，“是” 类别曲线上的最后一点在（100%，100%），如果用户选择数据集的100%，肯定会获得数据集中的所有违约者。

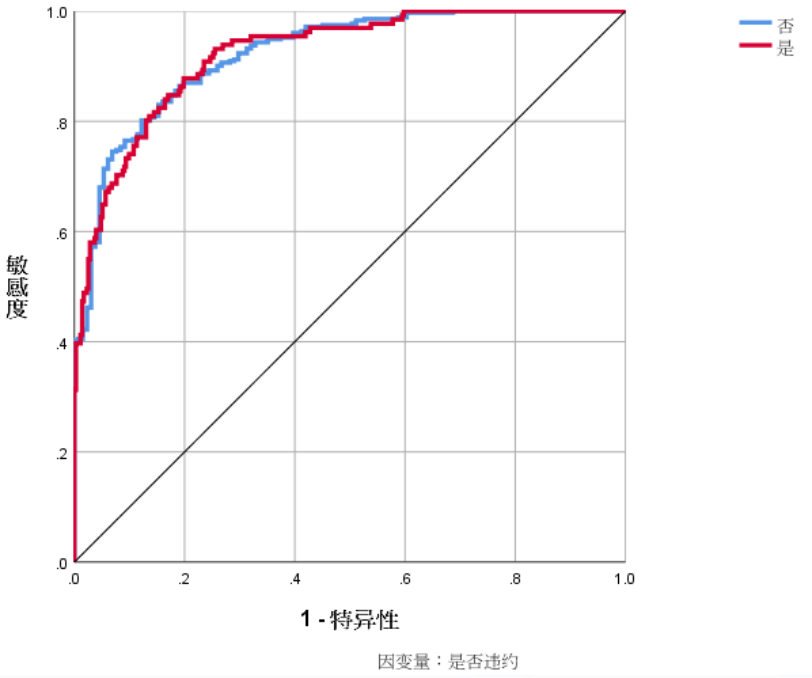


对角线为“基线”，也就是随机选择线；如果用户从评分数据集随机选择 10% 个案，那么从这里面期望“获取”的违约个案，在全部违约个案中占比也肯定是大约10%。所以从这种意义上讲，曲线离基线的上方越远，增益越大。

“分类问题监督学习”的性能度量 四、ROC曲线与AUC值

ROC曲线，又称“接受者操作特征曲线”、“等感受性曲线”，ROC曲线主要是用于预测准确率情况。最初ROC曲线是运用在军事上，现在广泛应用在各个领域，比如判断某种因素对于某种疾病的诊断是否有诊断价值。曲线上各点反映着相同的感受性，它们都是对同一信号刺激的反应，只不过是在几种不同的判定标准下所得的结果而已。

ROC曲线以虚惊概率（又被成为假阳性率、误报率，图中为1-特异性）为横轴，击中概率（又被称为敏感度、真阳性率，图中为敏感度）为纵轴所组成的坐标图，和被试在特定刺激条件下由于采用不同的判断标准得出的不同结果画出的曲线。虚惊概率X轴越接近零，击中概率Y轴越接近1代表准确率越好。



$$敏感度=\frac{TP}{TP+FN}$$ （与前面介绍的查全率一致）

$$特异度=\frac{TN}{TN+FP}$$

“分类问题监督学习”的性能度量 四、ROC曲线与AUC值

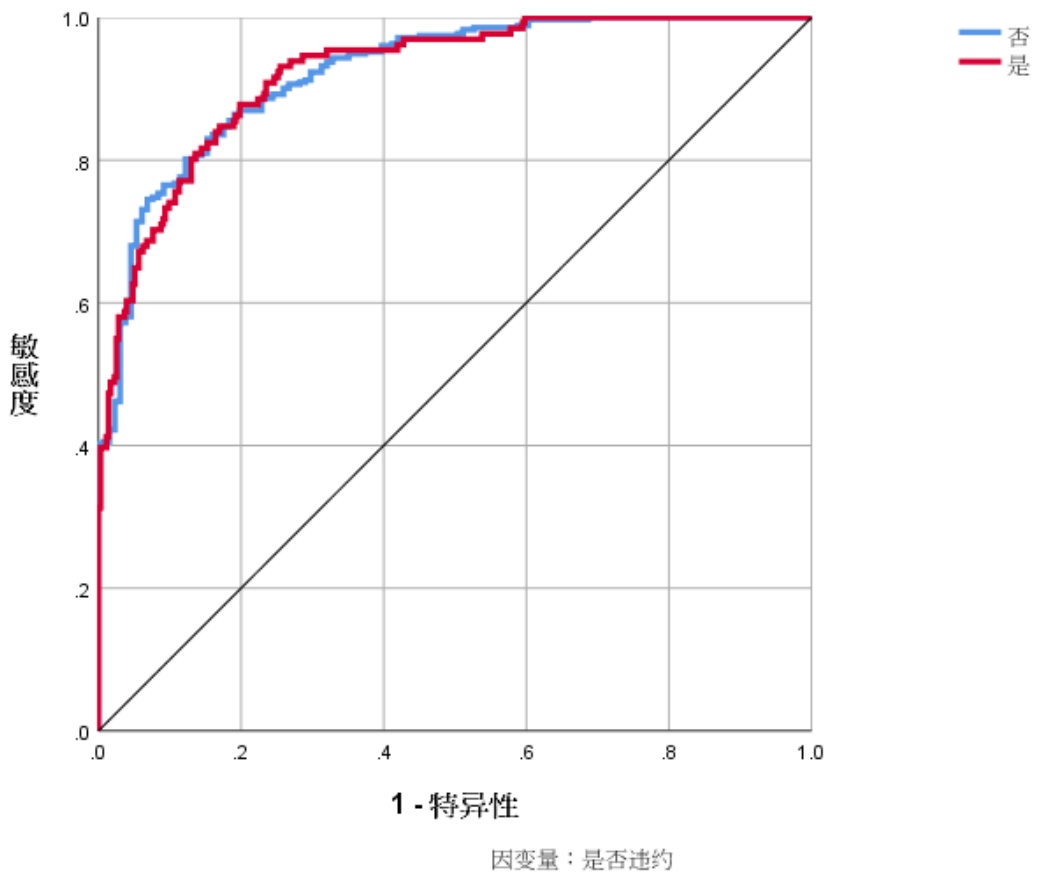
ROC曲线下方区域又被称为AUC值，是 ROC 曲线的数字摘要，取值范围一般在0.5和1之间。

当AUC = 1，是完美模型，采用这个预测模型时，存在至少一个阈值能得出完美预测。绝大多数预测的场合，不存在完美模型。

当0.5 < AUC < 1，优于随机猜测。这个模型妥善设定阈值的话，能有预测价值。

当AUC = 0.5，跟随机猜测一样，模型没有预测价值。

当AUC < 0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。



“分类问题监督学习”的性能度量　五、科恩kappa得分

1960年科恩等提出用Kappa值作为评价判断的一致性程度的指标。科恩kappa得分既可用于统计中的一致性检验，也可以用于机器学习中来衡量分类精度。

科恩kappa得分的基本思想是：将样本示例的预测值和实际值视为两个不同的评分者，观察两个评分者之间的一致性。由于样本分类一致性的大小不完全取决于特定机器学习算法的性能，还可能是由于随机因素的作用，致使随机猜测与特定机器学习算法得出相同的分类结论。或者说，没有采用特定机器学习算法的随机猜测对样本进行分类也可能会得出与特定机器学习算法一样的结论，而这种一致性结论完全是由于随机因素导致的。所以在评价机器学习的真正性能时，需要剔除掉随机因素这种虚高的水分。

PART 07

模型评估

模型评估

前面我们提到，评估机器学习模型优劣的标准是模型的泛化能力，所以关注的应该是泛化误差而不是经验误差，一味追求经验误差的降低，就会导致模型“过拟合”现象。但是泛化误差是难以直接观察到的，我们应该如何选择泛化能力比较好的模型呢？

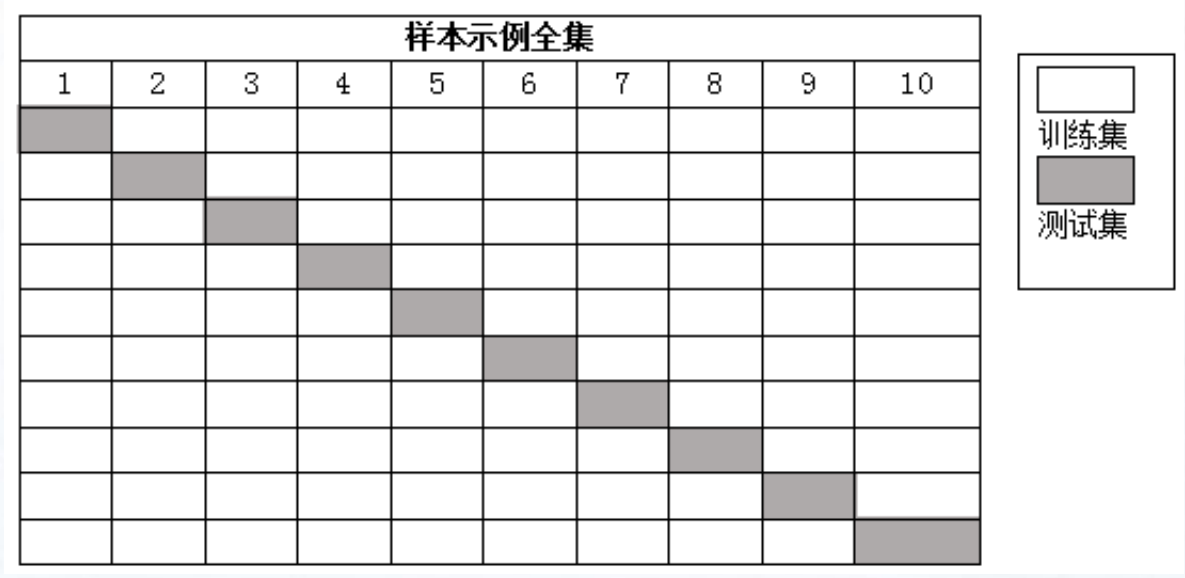
机器学习中常用的度量模型的泛化能力的方法包括验证集法、K折交叉验证、自助法。

验证集法

验证集法又被称为“留出法”，基本思路是将样本示例数据集划分为两个互斥的集合：训练集和测试集。其中训练集占比一般为2/3~4/5，常用70%；测试集占比一般为1/5~1/3，常用30%。训练集用来构建机器学习模型；测试集也称“验证集”“保留集”用来进行样本外预测，并计算测试集误差，估计模型预测能力。

K折交叉验证

K折交叉验证是针对验证集法的另外一种改进方式，也广泛用于机器学习实践。具体操作方式，就是我们首先把样本示例全集采用分层抽样的方式，随机划分为大致相等K个子集，每个子集包含约1/K的样本，K的取值通常为5或者10，其中10最为常见。然后，我们每次都把K-1个子集的并集，也就是约（K-1）/K的样本作为训练集，把1/K的样本作为测试集，基于训练集训练获得模型，基于测试集进行评价，计算测试集均方误差。最后，将K次获得的K个验证集的均方误差进行平均，即为对测试误差的估计结果。



K折交叉验证

假定样本示例全集中有n个样本，如果采取n折交叉验证，那么只有1种划分方式，即每个样本都构成1个测试集，其他n-1个样本构成训练集，这种方法被称为“留一法”，属于K折交叉验证方法的特例。在“留一法”情形下，由于训练集相对于样本示例全集只减少了1个样本，所以其高度接近使用样本示例全集进行训练的结果，其评估结果也是相对准确的。但是，“留一法”的缺陷也很明显，那就是计算量非常大，有多少个样本就需要训练多少次模型，然后求平均值，如果是针对大数据样本，那么其计算时间开销将是非常大的。而且正如前面所指出的，Wolpert(1996)等提出的“没有免费的午餐定理（No Free Lunch Theorem）”依旧适用，或者说，“留一法”未必也不可能在所有情形下都优于其它算法

K折交叉验证

关于确定K值的问题，实质上也涉及偏差和方差权衡的问题。如果K的值非常大，比如前述的“留一法”，那么其偏差会比较小，但是由于“留一法”每次训练集的样本变化比较小，只有1个样本发生变动，其结果之间存在很高的正相关性，基于这些高度相关的结果进行平均，会导致其方差比较大。而如果K的值非常小，那么训练集在样本示例全集中的占比比较少，会产生相对比较大的偏差，但是由于训练集较少，所以每次样本变化比较大，结果之间的相关性相对较小，将结果进行平均得到的方差也会相对比较小。

与验证集法类似，我们可以重复使用K折交叉验证法，这也是所谓的“重复K折交叉验证法”。具体操作方式，就是我们把K折交叉验证法重复K次，最后将每次得到的结果误差进行平均。

自助法

自助法本质上是一种有放回的再抽样，其实现过程是这样的：假设样本示例全集容量为k，在样本示例全集中我们首先抽取一个示例记下其编号，再将其放回全集使得该样本在下次抽取时仍有可能被抽到，然后再重新抽取一个示例、记下后放回全集，如此重复k次，就会得到由k个示例构成的“自主抽样样本集”。不难明确的一个事实就是，在抽样的过程中，很可能有的示例被多次重复抽到，而也有的示例一次也没有被抽到过，示例一次也没有被抽到过的概率计算如下：

$$\lim_{k\to\infty}\left(1-\frac{1}{k}\right)^{k}=e^{-1}\approx 0.368$$

当k趋近于正无穷大时，示例在k次抽样中，一次也没有被抽到过的概率约为36.8%，那么这些没有被抽到过的示例就可以被当做测试集，也被称作“包外测试集”，基于其计算的测试集误差也被称为“袋外误差”，而不包含在测试集之内的示例即作为训练集。

PART 08

机器学习项目流程

机器学习项目流程

一个完整机器学习项目的流程包括如下几个步骤：明确需解决的业务问题、获取与问题相关的数据、特征选择与数据清洗、训练模型与优化模型、模型融合、上线运行。各个步骤之间既环环相扣又相互融合。

明确需解决的业务问题

机器学习项目的第一步就是明确需解决的业务问题。如果业务问题不够明确，那么就无法在此基础上选择恰当的机器学习算法模型，一方面由于机器学习的计算时间一般都比较长，会导致大量的时间、计算量浪费，另一方面，根据“没有免费的午餐定理（No Free Lunch Theorem）”，算法的优劣评价仅能针对特定业务问题，而不可能在解决所有问题方面都具有相对优势。明确需解决的业务问题，就是要明确是监督学习还是无监督学习；进一步的，如果是监督学习，那么是分类问题还是回归问题？如果是分类问题，那么应该更除了关注错误率和精度之外，应该更关注查准率还是查全率等等。

获取与问题相关的数据

在明确需解决业务问题的前提下，前面我们提到，受噪声的影响，当数据给定时，机器学习所有达到的泛化能力的上限也就确定，再优秀的机器学习算法所能做的只是尽可能逼近泛化能力上限，所以获取与问题相关数据的质量至关重要。

数据质量体现在数据的完整性、准确性、相关性等多个方面，首先数据应该尽可能保持字段完整，且样本示例的各个特征（属性）及响应值尽可能有较少的缺失值；其次数据应该是准确的，无论是积累的历史数据，还是人工审核后的虚拟数据，用户都应该实施必要的审查，尽可能降低数据本身的错误或偏差；然后数据与所需解决的业务问题之间还应该具备较强的相关性，或者说数据应该对解决业务问题能够有所帮助，而不是毫无关联。

获取与问题相关的数据

除了数据质量之外，还应该对数据的样本示例全集容量有着恰当的评估，至少应该知晓数据的量级，结合拟确定的特征，估算数据集对于内存的消耗程度，如果计算量超出承受能力，那么就需要针对性的选择一些更加简单的算法，或者在选择特征方面更加审慎，必要时进行降维处理。

特征选择与数据清洗

获取数据后，我们需要进行特征选择与数据清洗。在实务中，这一步通常与上一步“获取与问题相关的数据”同时进行，因为在很多情形下，我们只有进行了特征选择，才能依据选择的具体特征，去获取相应的数据。

特征选择即选择能够影响响应变量的变量，或者说预期哪些因素能够对响应变量产生影响，比如商业银行个人授信业务中，如果响应变量为授信业务是否违约，那么可能需要选择客户的性别、职业、年收入水平、历史征信记录等系列变量作为模型的特征变量，而客户的爱好、身高、体重等变量可能不是很好的特征变量，因为这些对于客户是否违约的影响预期不够显著。所以，特征选择是基于对业务、对拟解决问题的深刻理解与洞察，或者说，机器学习成功的关键，很大程度上并不在于技术人员或数据分析时，更多的取决于实施机器学习的相关业务领域专家的能力、水平和经验。

特征选择与数据清洗

数据清洗是一个提高数据质量的过程，良好的数据清洗能够使得机器学习算法的效果和性能得到显著提高。针对数据的清洗，包括归一化、标准化、离散化、因子化、去除共线性、缺失值处理等多种方式。

训练模型与优化模型

训练模型时主要关注模型的“泛化”能力，尽可能减少我们在前面提出的“过拟合”“欠拟合”问题，在“偏差”“方差”之间找到平衡，并基于“性能度量”和“模型评估”中的注意事项，客观的对模型优劣进行评价。

针对训练模型不及预期，或认为还有更多优化空间，那么就需要对模型进行进一步的优化，当模型出现过拟合时，基本调优思路是增加数据量，降低模型复杂度；当模型出现欠拟合时，基本调优思路是提高特征数量和质量，增加模型复杂度。在训练过程中，还需注意观察误差样本，全面分析误差产生原因，究竟是参数的问题还是算法选择的问题，是特征的问题还是数据本身的问题等等。

训练模型与优化模型

常用的优化方法包括更换其他机器学习算法（比如将决策树算法换成人工神经网络算法）、调整特定算法中的参数（比如针对人工神经网络算法变换函数）等。用户可对比不同情形下的各种结果，找出最为可用的模型。

优化后的新模型需要重新进行训练和评价，而训练和评价后的模型很可能需要再次优化，如此反复，达到用户满意为止。

模型融合

实务中，提升算法准确度的方法主要就是前面所述的“特征选择与数据清洗”以及本节介绍的“模型融合”。模型融合的基本思想是在各种不同的机器学习任务中使结果获得提升，实现方式是训练多个模型（个体学习器），然后按照一定的方法集成在一起（强学习器）。当个体学习器准确性越高（之间的性能表现不能差距太大），多样性越大（之间的相关性要尽可能的小），则融合越好。一般来说，随着集成中个体学习器数目的增大，集成的错误率将指数级下降，最终趋向于零，这一点已有严格的数学公式所证明。

模型融合　一、Boosting方法（提升法）

如果个体学习器之间存在强依赖关系、必须串行生成的序列化方法，则应选择Boosting方法，主要优化bias（偏差，或称模型的精确性），当然对于降低方差也有一定作用。Boosting方法的操作实现过程是：

（1）首先从训练集用初始权重训练出一个弱学习器x1，得到该学习器的经验误差；然后对经验误差进行分析，基于分析结果调高学习误差率高的训练示例的权重，使得这些误差率高的训练示例在后面的学习器中能够受到更多的关注；再后基于更新权重后的训练集训练出一个新的弱学习器x2。

（2）不断重复这一过程，直到满足训练停止条件（比如弱学习器达到指定数目），生成最终的强学习器。

（3）将弱分类器预测结果进行加权融合并输出，比如AdaBoost通过加权多数表决的方式，即增大错误率小的分类器的权值，同时减小错误率较大的分类器的权值。

需要注意的是：Boosting算法在训练的每一轮要检查当前生成的基学习器是否满足基本条件。

模型融合　二、Bagging方法（袋装法、随机森林）

如果个体学习器之间不存在强依赖关系、可同时生成的并行化方法，则应选择Bagging（Bootstrap Aggregating）方法，bagging算法的思想是通过对样本采样的方式，使我们的个体学习器存在较大的差异，主要优化variance（方差，或称模型的鲁棒性）。具体操作步骤是：

1、采用前面介绍的自助法从样本示例全集中抽取x个示例，进行y轮抽取，最终形成y个相互独立的训练集。

2、使用y个相互独立的训练集，根据具体问题采用不同的机器学习方法，分别训练得到y个模型。

3、针对回归问题，将y个模型的均值作为最终结果；针对分类问题，将y个模型采用投票的方式（多数者胜出）得到分类结果。

模型融合　二、Bagging方法（袋装法、随机森林）

在Bagging方法的基础上，生成了随机森林法，有效解决了决策树方法容易过拟合的问题（随机森林法、决策树方法将在后文详细介绍）。随机森林法较Bagging方法的进步体现在除了Bagging自助法之外，还每次随机抽取一定数量的特征（通常为sqr(n)，n为全部特征个数），也就是说自变量（解释变量）也是随机的。最终针对回归问题，将每颗决策树结果的平均值作为最终结果；针对分类问题，将y个模型采用投票的方式（多数者胜出）得到分类结果。

上线运行

当模型融合工作完成，往往意味着在用户认知范围内以及可用资源条件限制下实现了最优解，或者虽不是最优解但至少是用户可接受的可用解，而整个机器学习算法阶段也基本宣告结束。下一步就是最终的上线运行，将之应用于实践，并且随着环境的不断变化，及时、动态持续修正。需要特别说明的是，上线阶段也非常重要，因为可能会出现实际运行中运行速度、资源消耗、稳定程度与项目预估时存在较大差异的情况，如果这一事项发生且难以通过小范围优化所调整，可行性较差，那么可能就会重新建模，重复上述过程，直至可行为止。

感谢聆听

THANKS