本课时，我将为你介绍数据挖掘操作流程的倒数第二个步骤：模型评估。

在每次训练一个模型之后，尤其是现在的深度模型，通常要消耗大量的时间等待模型的产出，那种心情是可想而知的，谁都希望能够有一个好的结果。模型评估就是对你的模型进行多种维度的评估，来确认你的模型是否可以放到线上去使用。

这一课时，我将介绍一些常用的评估指标，其中会涉及一些比较难理解的名词和计算，不过不用担心，我会带你逐个突破难关。当然，我也准备了一个关于 “训练一个小猪图片分类模型” 的例子，让你能够更加直观地理解如何去评估模型。好了，我们先来看看这个例子。

假设我们训练了一个 “识别图片是不是关于小猪” 的分类模型，这是一个二分类器，当你给它一张图片的时候，它会告诉你这是一张小猪的图片，或者不是一张小猪的图片。我们有 1000 张图片用于测试该模型的效果，并且预先已经进行了人工的标注（这里假设人工标注的数据都是 100% 正确），每张图都会标注是或者不是小猪的图片，假设有 800 张标注“是”，200 张标注“否”。

评估指标

混淆矩阵与准确率指标

准确率相关指标是在模型评估时最受关注的指标，它可以直接反映一个模型对于样本数据的学习情况，是一种标准化的检验。就像老师教给你 10 道计算题，然后又用这 10 道题来出考题，如果你都答对了，说明你已经学会了。准确率相关的指标就反映了这样一种结果，下面看看模型学习的直接效果。

我们把这 1000 张图放进分类器进行分类计算，每张图都会得到一个预测结果，通过对预测结果的统计可以知道，被模型预测为 “是” 的图片有 770 张，被模型预测为 “否” 的图片有 230 张。这个时候每张图上会有两个结果：一个人工标注结果、一个模型预测结果。根据这两个数据的统计，可以得到一个混淆矩阵：

样本 1000 份	**模型预测：是**	**模型预测：否**
人工标注：是	745（TP）	55（FN）
人工标注：否	25（FP）	175（TN）

矩阵中包含以下 4 种数值：

1. 真阳性（True Positive，TP）：小猪图被判定为小猪图。样本的真实类别是正例，并且模型预测的结果也是正例（在本案例中此数值为 745）。

2. 真阴性（True Negative，TN）：不是小猪图被判定为不是小猪图。样本的真实类别是负例，并且模型将其预测成为负例（在本案例中此数值为 175）。

3. 假阳性（False Positive，FP）：不是小猪图被判定为小猪图。样本的真实类别是负例，但是模型将其预测成为正例（在本案例中此数值为 25）。

4. 假阴性（False Negative，FN）：小猪图被判定为不是小猪图。样本的真实类别是正例，但是模型将其预测成为负例（在本案例中此数值为 55）。

根据上述混淆矩阵，我们可以计算一些数值。

- 准确率（Accuracy）：是指所有预测正确的占全部样本的概率，即小猪图被预测成小猪，以及不是小猪被预测成不是小猪的结果，与所有图片的比值，公式为 (TP+TN)/(TP+FP+FN+TN)，在本案例中为 (745+175)/(745+175+25+55)=0.92。
- 精确率（Precision）：指的是预测正确的结果占所有预测成 “是” 的概率，即 TP/(TP+FP)。精确率按照类别来计算，比如说对于 “是小猪图” 这个类别的精确率是 745/(745+25)≈ 0.9675。

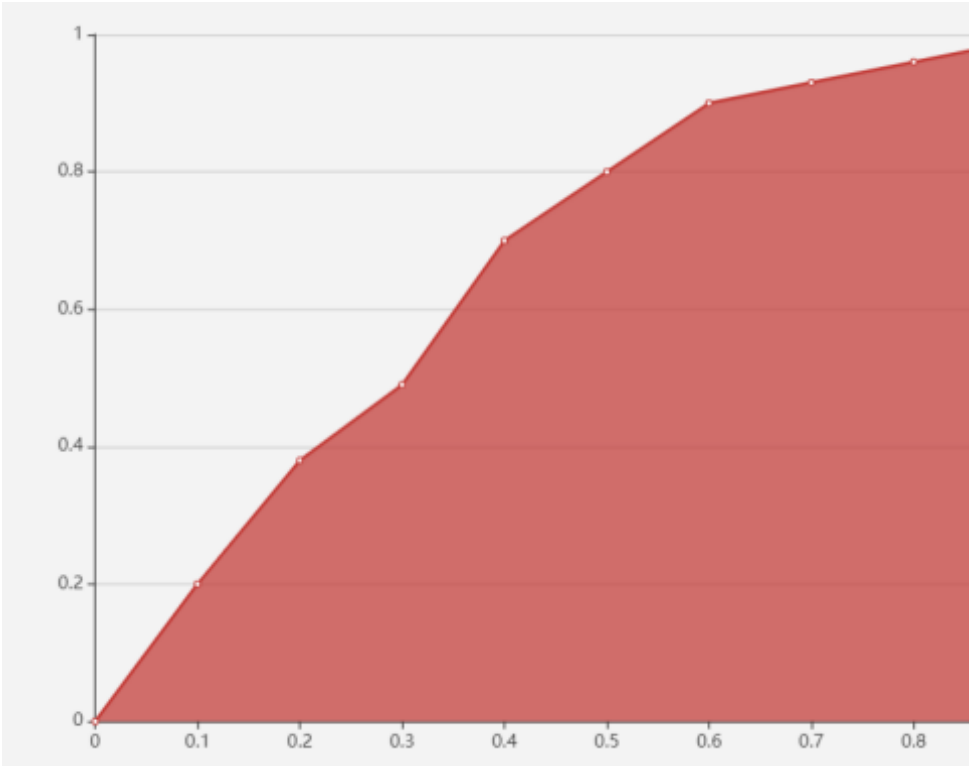
- 召回率（Recall）：按照类别来区分，某个类别结果的召回率即该类别下预测正确的结果占该类别所有数据的概率，即 TP/(TP+FN)，在本案例中 “是” 类别召回率 745/(745+55)≈0.93。
- F 值（F Score）：基于精确率和召回率的一个综合指标，是精确率和召回率的调和平均值。一般的计算方法是 2 *(Precision*Recall）/（Precision+Recall）。如果一个模型的准确率为 0，召回率为 1，那么 F 值仍然为 0。
- ROC 曲线和 AUC 值：这个略微有点复杂，但也是一个非常常用的指标。仍然是基于混淆矩阵，但不同的是这个对指标进行了细化，构建了很多组混淆矩阵。

具体是怎么构建混淆矩阵的呢？仍然以这个小猪图分类为例，在有些模型的产出中，通常给出 “是” 和 “否”的概率值（这两个概率值相加为 1），我们根据概率值来判定最终的结果，那么这时就有个问题了，我们选多少概率值来判定结果？比如可以指定 “是” 的概率为 0.1 及以上时，就判定结果为 “是”；“是” 的概率小于 0.1 的时候，判定结果为“否”。那么，选定若干组判定的概率，就能得到若干组混淆矩阵。

那么在每一组混淆矩阵中，我们获取两个值：真正例率和假正例率。

- 真正例率：TP/（TP+FN）
- 假正例率：FP/（FP+TN）

使用这两个值在坐标系上画出一系列的点，纵坐标是真正例率，横坐标是假正例率，把这些点连起来形成的曲线就是 **ROC 曲线**（Receiver Operating Characteristic，接收者操作特征）。ROC 曲线下方的面积是 AUC 值（Area Under Curve，曲线下面积），ROC 曲线和 AUC 值可以反映一个模型的稳定性，当 ROC 曲线接近对角线时，说明模型输出很不稳定，模型就越不准确。



ROC 曲线和 AUC 值图

十分重要的业务抽样评估

除了上面一系列的指标评估，我们还有一项重要的评估需要进行，那就是业务抽样。因为我们的模型是基于业务制定的，最终的效果还是要回归到业务上。

在理想的状况下，使用上面的指标基本可以判定模型的效果，但是在实际中还存在着一些问题，这通常都是由数据本身并不完美导致的。对于标注数据，人工标注通常也存在一定的错误率，而不是 100% 正确，所以在使用前面的指标进行评估的时候可能会与实际结果存在一些差异。进行业务抽样评估可以减弱这种情况，在多方背靠背评估之后再进行意见的统一，最终得到一个在业务上认可的准确率情况。

泛化能力评估

除了要求模型的准确情况，还有一项重要的能力也是我们非常重视的，那就是模型的泛化能力。泛化能力反映的是模型对未知数据的判断能力，就好像学生具备举一反三的能力，老师教了 1+1=2、1+2=3，考试时出了一道 1+1+1=3 也能够计算正确，这就有良好的泛化能力。因为在数据挖掘中，数据的维度通常有很多，而且数据也都是非标准值，任意记录之间的数据都会存在着差异，所以泛化能力好的模型在数据存在着波动的情况下，仍然能够做出正确的判断。

- 过拟合与欠拟合

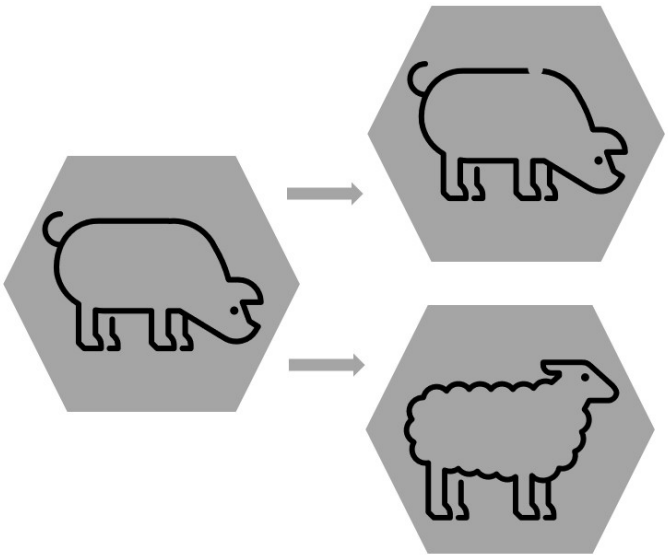
我们通过两个指标可以评估模型的泛化能力是好还是坏，那就是**过拟合（overfitting）**和**欠拟合（underfitting）**。

过拟合：模型在训练集上表现良好，而在测试集或者验证集上表现不佳。这就是说，模型对样本学习有些过度了，已经进入了死记硬背的程度，而不是掌握了普适规律，这个时候可以说泛化能力比较差。

欠拟合：在训练集和测试集上的表现都不好。这就是说模型连最基本的内容都没有学到，比如老师教你 1+1=2、1+2=3，考试也考 1+1=2，结果还是做错了。

下面我们再来看看小猪的例子。

通常情况下，我们的小猪图都像左侧一样，右侧有两张图，上面一张可以看出仍然是小猪图，但是后背上的线条有一个缺口，如果此时模型告诉我们，这个后背上有个缺口，这不是小猪图，那么这时就出现了过拟合（判断条件过于苛刻）。右下侧是一张小羊图，如果模型告诉我们这个也有四条腿，这个是小猪，那就是欠拟合（特征学习不完全）。



小猪例子图

关于泛化性能的评估，主要依赖于在不同的数据集上的准确结果之间的比较。要处理过拟合和欠拟合的问题，通常需要对我们的数据进行重新整理，总结出现过拟合和欠拟合的原因，比如是否数据量太少、数据维度不够丰富、数据本身的准确性较差等，然后调整数据重新进行训练。

其他评估指标

除了上述的两大类指标，还可以从以下几个方面来对模型进行评估。

模型速度：主要评估模型在处理数据上的开销和时间。这个主要是基于在实际生产中的考虑，由于模型的应用在不同的平台、不同的机器会有不同的响应速度，这直接影响了模型是否可以直接上线使用，关于更多模型速度相关的问题，我们将在下一课时模型应用中介绍。

鲁棒性：主要考虑在出现错误数据或者异常数据甚至~是~数据缺失时，模型是否可以给出正确的结果，甚至是否可以给出结果，会不会导致模型运算的崩溃。

可解释性：随着机器 学习算法越来越复杂，尤其是在深度学习中，模型的可解释性越来越成为一个问题。由于在很多场景下（比如金融风控），需要给出一个让人信服的理由，所以可解释性也是算法研究的一大重点。

评估数据的处理

关于数据集的处理，重点目标在于消减评估时可能出现的随机误差。在前面准备数据的课时内容中已经提过，这里我们再对一些方案详细介绍一下。

- **随机抽样**：即最简单的一次性处理，把数据分成训练集与测试集，使用测试集对模型进行测试，得到各种准确率指标。
- **随机多次抽样**：在随机抽样的基础上，进行 n 次随机抽样，这样可以得到 n 组测试集，使用 n 组测试集分别对模型进行测试，那么可以得到 n 组准确率指标，使用这 n 组的平均值作为最终结果。
- **交叉验证**：交叉验证与随机抽样的区别是，交叉验证需要训练多个模型。譬如，k 折交叉验证，即把原始数据分为 k 份，每次选取其中的一份作为测试集，其他的作为训练集训练一个模型，这样会得到 k 个模型，计算这 k 个模型结果作为整体获得的准确率。
- **自助法**：自助法也借助了随机抽样和交叉验证的思想，先随机有放回地抽取样本，构建一个训练集，对比原始样本集和该训练集，把训练集中未出现的内容整理成为测试集。重复这个过程 k 次、构建出 k 组数据、训练 k 个模型，计算这 k 个模型结果作为整体获得的准确率，该方法比较适用于样本较少的情况。

总结

这一课时我们终于进入了模型评估环节，这是检验模型效果的重要阶段，直接决定一个模型是进入下一个环节，还是回到上一个环节回炉重炼。我们主要讲了模型的各种评估指标，从一个混淆矩阵出发，衍生出一系列的准确度评测；然后对模型泛化能力进行评估。在评估指标后面，我们又介绍了如何在数据上进行一些优化从而减少评估时产生误差，这部分是准备数据的延伸。

在这里需要说明的是，这一课时我们所介绍的模型评估方法中，主要适用于分类模型，因为分类模型是一种有监督模型，所以通过指标来进行评测相对容易。对于无监督模型，由于本身没有非常明确的结果标准，所以也比较难找到一个衡量指标。