课程概述

您好！欢迎来到数据分析中级课程的第四个案例：寻找划分公司所属行业的最佳模型。从本案例开始，进入招聘数据分析阶段。案例数据集来源于数据分析相关岗位招聘薪资的爬取结果，本次案例主要涉及以下内容：

1. 内容简介

1）问题界定

基于业务理解和数据理解构造主业务问题。

2）数据准备

除了简单的数据检视与清理工作，重点在于中文长文本数据的处理--文本向量化，这是本案例的第一个大难点。

3）数据建模与数据可视化

本案例注重于模型与模型之间的对比，即使用决策树算法、朴素贝叶斯算法和 SVM 算法探讨同一个问题，然后探讨不同的数据分割和数据重构方法对模型的影响，同时也会基于模型的表现，寻找分类效果最佳的模型。

2. 学习目标

学习完本次案例，你能够达到以下目标：

1）能够掌握中文长文本数据的处理方法；

2）了解决策树剪枝、卡方检验和 SVM 算法的基本原理；

3）了解两种不同的数据分割方式：随机分割、分层分割；

4）了解三种不同的数据重构方式：SMTOE、ADASYN、RamdomUnderSampler；

5）掌握决策树算法和 SVM 算法的调参；

6）能够结合模型评估方法筛选出最佳的分类模型。

业务理解

1. 业务现状描述

青青招聘网站是一家提供招聘信息的网站，用人公司在该网站发布招聘信息时，除了岗位、薪资、技能要求等外，还需要填写公司概述和公司所属行业，其中行业是由网站受控词表控制的，不能使用自然语言的填写。那么就需要公司从网站的行业分类体系中，找到自己公司的所属行业。

这是一个比较麻烦的过程，为此青青招聘网站准备推出一个新功能：

网站可以根据公司的概述自动匹配公司所属行业

现有一批关于“数据分析”岗位相关的招聘信息，主要涉及众多公司的概述，以及该公司所属的行业数据。接下来数据分析人员需要从这些数据中进行分析，构建公司所属行业的分类模型，以实现该功能成功进行测试，为系统研发部分提供参考。

2. 识别利益相关群体

作为受数据分析项目影响的利益相关者：发布招聘信息的公司。他们的需求是根据他们提供的公司概述能够准确分类出公司所属行业；

而对于能够影响项目的利益相关者：系统研发部门人员。他们的需求是获得最适合用于分类公司所属行业的算法。

因此，数据分析要在分类公司所属行业这个问题上比较不同分类模型的准确率，得到在大多数情况下准确率最高的模型，最好还能提供这个模型的优化方法作为系统研发的参考。

3. 业务问题构造

现有案例中，网站希望通过分析已有的公司概述和公司所属行业数据，寻找一个最合适的分类模型。当前的主业务问题是：

哪个分类模型在划分公司所属行业上的表现最好？

在这个分类公司所属行业的问题中，模型的好坏主要通过模型的**准确率**来判断。

在给定数据集和模型的情况下，可能影响准确率的因素包括：

①不平衡数据集的处理方法；

②数据分割的方式（我们将其设定为随机划分和根据标签分布划分两种情况）；

③模型超参数的设定。

在后面模型的比较中，我们需要首先确定不同分类模型最高准确率的条件，然后再选择使用哪个分类模型。

数据理解

本案例中涉及到的数据较为简单，包括公司概述和公司所属行业两个文本型字段。如表中所示：

数据字段	字段含义
company_overview	公司概述
company_industry	公司所属行业

而业务问题需要解决的就是：

通过“公司概述”的文本来分类“公司所属行业”。