第一周

Day1-Day2

注意：回顾题一般为课本上的题目，需要用自己的话进行总结打卡。补充题则为课本上没有另行补充的题目，会提供一些参考答案。

1. 回顾：请参考课本 1.1 节总结，什么是归一化？课本上介绍了哪些归一化的方法？为什么需要进行归一化？标准化和归一化是一个意思吗？

参考答案：参考课本 1.1 节独立进行总结。
标准化和归一化的区别请参考链接(注意回答中的英文翻译)
https://www.zhihu.com/question/20467170/answer/392949674

2. 回顾：请参考课本 1.3，1.4 节总结，什么是组合特征？怎么处理高维组合特征？怎样有效的找到组合特征？

参考答案：参考课本 1.3、1.4 节独立进行总结。

3. 回顾：请参考课本 1.5，1.6 节总结，什么是词袋模型？TF-IDF 怎么计算的？什么是词嵌入模型？如何理解 word2vec？

参考答案：参考课本 1.5、1.6 节独立进行总结。
注意：面试 NLP 的同学需要重点弄懂这个问题！尤其是 word2vec 的理解！请参考：
理解 word2vec：https://zhuanlan.zhihu.com/p/26306795
深入理解 word2vec：请去看斯坦福大学 cs224n 课程

4. 回顾：请参考课本 2.1 节，总结分类问题和回归问题评价指标。

参考答案：参考课本 2.1 独立进行总结。

分类模型常用评估方法：

指标	描述
Accuracy	准确率
Precision	精准度/查准率
Recall	召回率/查全率
P-R曲线	查准率为纵轴，查全率为横轴，作图
F1	F1值
Confusion Matrix	混淆矩阵
ROC	ROC曲线
AUC	ROC曲线下的面积

回归模型常用评估方法：

指标	描述
Mean Square Error (MSE, RMSE)	平均方差
Absolute Error (MAE, RAE)	绝对误差
R-Squared	R平方值

5. 回顾：请参考课本 2.2 节总结，什么是 ROC？如何绘制 ROC？如何计算 AUC？
参考答案：参考课本 2.2 独立进行总结。

6. 回顾：请参考课本 2.7 节，总结什么是过拟合、欠拟合？如何处理过拟合和欠拟合？
参考答案：参考课本 2.7 节独立进行总结。

7. 补充：如何处理数据中的缺省值？
参考答案：

（1） 直接使用含有缺失值的数据：某些算法(如决策树)可以直接使用含有缺失值的情况。优点是直接使用原始数据，排除了人工处理缺失值带来的信息损失。缺点是只有少量的算法支持这种方式。

（2） 删除含有缺失值的数据。如果样本中包含大量的缺失值，只有少量的有效值，则该方法可以接受，否则会损失大量的信息。

（3） 缺失值补全。这是最常用的一种方法，优点是可以保留信息，缺点是会引入额外的误差。最常用的是均值插补/同类均值插补，其他的一些方法如：建模预测、高维映射、多重插补、压缩感知及矩阵补全

8. 补充：为什么类别不平衡问题会影响结果？如何处理类别不平衡问题？
参考答案： 通常在机器学习中都有一个基本假设：不同类别的训练样本数目相当。若假设不成立则使得基于该假设下的方法不能很好的工作。
处理类别不平衡一般有如下的一些方法：

（1） 再缩放

一般的分类问题最终的输出是一个概率分布，我们可以通过将 p 和一个阈值，如 0.5 进行比较，若 $p > 0.5$判断该类是正类。因而 $\frac{p}{1-p}$刻画的是正类可能性与反类可能性的比值。当存在类别不平衡时，假设 N_t表示正类样本数目，N_f表示反类样本数目，则观测几率是$\frac{N_t}{N_f}$，假设训练集是真实样本总体的无偏采样，因此可以用观测几率代替真实几率。于是只要分类器的预测几率高于观测几率就应该判断为正类。即如果$\frac{p}{1-p} > \frac{N_t}{N_f}$，则说明正类预测的概率更高，故而预测为正类。

通常分类器是通过概率进行分类的，因而可以令$\frac{\overline{p}}{1-\overline{p}} = \frac{p}{1-p} \times \frac{N^f}{N^t}$,然后比较$\overline{p}$和设定的阈值的大小进行分类。

再缩放虽然简单，但是由于“**训练集是真实样本总体的无偏采样**”这个假设往往不成立，所以无法基于训练集观测几率来推断出真实几率。

（2） 欠采样

删除一些样本，使得正负样本数量接近。

欠采样若随机抛弃反类，则可能丢失一些重要信息。常用方法是将反类划分成若干个集合供不同学习器使用，这样对每个学习器来看都是欠采样，但是全局来看并不会丢失重要信息。

（3） 过采样

增加一些样本，使得正负样本数量接近，但不能简单对原样本进行重复采样，否则会出现严重的过拟合问题。通常可以使用一些重采样的方法，如 SMOTE，该方法可参考课本 8.7 节。