第八周

1. 无监督学习-聚类算法

在无监督学习中，面对的是一组无标签的数据，数据之间不具有任何关联的标记。通过将这些无标记的数据通过特定算法进行训练，要求输出数据的结构，将数据分成有紧密关系的子集或簇。

K-means聚类算法的步骤：

1. 簇分配

选择聚类中心，通过遍历所有的数据点，依据每个数据点距离不同聚类中心的远近，将数据点分配给不同的聚类中心。

1. 移动聚类中心

计算分配给聚类中心的数据点的均值，然后将该聚类中心移动到平均位置。直至算法收敛，即聚类中心不再改变时，聚类完成。

k-means聚类算法问题：

1. 如果某个聚类中心没有分配数据点，通常做法是移除该聚类中心，但如果是n维，可以重新随机找一个聚类中心。
2. K-means聚类算法的输入参数是x样本数据和聚类量K，在算法实现过程中使用的参数是某数据点所分配的聚类中心索引c(i)和聚类中心位置。
3. K-means聚类算法的代价函数为：



失真代价函数表示样本数据点到聚类中心的距离的平方的均值，需要得到能使最小的c和。

1. 随机选择初始聚类中心，一般方法是随机选择k个样本数据作为聚类中心，但是可能会导致局部最优解，可以通过多次随机选择聚类中心，计算其失真代价函数最小的作为最优解。
2. 选择聚类簇的数量，一般是通过可视化数据或者输出结果来手工的确定。方法一：依据肘部原理，描绘随k增大cost的曲线，cost急剧变化的点的k值即为选择的簇的数量，但是有时肘点不清晰。方法二：依据算法所使用的应用环境来适当设置。
3. 维数约减
4. 用来进行数据压缩：通过降低维数，减少占用空间，提高算法效率。
5. 用来进行可视化数据：高维度数据可视化困难，不利于分析问题。通过降低维数，使得可视化为3D或2D都是利用分析问题的方法。
6. 维数约减算法-PCA

PCA算法是通过线性投影将高维数据映射到低维空间中表示，而要保证投影的数据的方差最大，同时保留原数据较多的数据特性。

PCA与LR的区别：

1. PCA是计算原数据与投影之间的误差，而LR是计算预测结果与样本数据结果的误差。
2. PCA中只有无标签的数据集x,LR是有标签的数据集（x，y）

PCA算法过程：

1. 数据预处理：对原数据集进行均值归一化（使每个特征的均值为0）和特征缩放（使数据具有可比性）。
2. 计算法协方差矩阵：



1. 利用SVD奇异值分解分解协方差矩阵：



1. 降维：

特征向量U为n\*n矩阵，只取前k个向量进行转化

，U为n\*k举证，得到新的数据样本z。

1. PCA应用
2. 恢复降维数据

降维使用：

恢复使用：

1. 选择降维k值

平均平方映射误差：

总变差：

选择方法为：选择不同的K值，运行PCA计算，代入到



选择k值一般要求比值小于等于0.01，选择满足此要求的k值。

而通过计算的奇异值S，有以下：



计算不同K值对应的累积贡献率，可以选择主成分的数量。

例如我要保留原始数据99%的差异性，则有：

1. 利用PCA加速监督学习算法

首先通过PCA将原始训练样本x映射到降维后新数据样本z，将z与标签数据重新组合进行学习求解最优参数。当加入新样本时，只需要将新样本进行PCA映射，加入到z中。PCA只运行在训练样本中，不用在交叉数据和验证数据中，当求得映射后，可直接用在交叉数据和验证数据中。

4）不建议使用PCA来防止过拟合。由于PCA是实现降维来加快算法的，降维将不容易过拟合，是一种方法，但是组好的方法是通过正则化项来防止过拟合。

5）PCA使用：一般情况是先不使用PCA算法，如果未能达到预期效果，再考虑进行PCA降维。

学习总结：

本周学习主要学习了k-means聚类算法和PCA算法以及其他应用。K-means算法是一种无监督学习算法，主要是对无标签的数据集进行分类，通过迭代簇分配和移动聚类中心两步骤，得到参数c和。PCA算法是不失去原有数据特性而实现降维的线性降维方式，其基本原理使用最大方差理论、最小误差理论，选出降维的k个方向向量，使得方差最大和误差最小。同时PCA也可以可视化数据和加快算法效率。