北风网
IBEIFENG.COM

人工智能之机器学习

聚类算法

上海育创网络科技股份有限公司

主讲人：刘老师(GerryLiu)

课程要求

- 课上课下“九字”真言
  - 认真听，**善摘录，勤思考**
  - **多温故，乐实践**，再发散
- 四不原则
  - 不懒散惰性，不迟到早退
  - 不请假旷课，不拖延作业
- 一点注意事项
  - 违反“四不原则”，不推荐就业

课程内容

- Jaccard相似度、 Pearson相似度
- K-means聚类
- 聚类算法效果评估(准确率、召回率等)
- 层次聚类算法
- 密度聚类算法
- 谱聚类算法

什么是聚类

- 聚类就是对大量未知标注的数据集，按照数据**内部存在的数据特征**将数据集划分为**多个不同的类别**，使**类别内的数据比较相似**，**类别之间的数据相似度比较小**；属于**无监督学习**
- 聚类算法的重点是计算样本项之间的**相似度**，有时候也称为样本间的**距离**
- 和分类算法的区别：
  - 分类算法是有监督学习，基于有标注的历史数据进行算法模型构建
  - 聚类算法是无监督学习，数据集中的数据是没有标注的

相似度/距离公式1

- 闵可夫斯基距离(Minkowski)
  - 当p为1的时候是曼哈顿距离(Manhattan)
  - 当p为2的时候是欧式距离(Euclidean)
  - 当p为无穷大的时候是切比雪夫距离(Chebyshev)

$$dist(X,Y)=\sqrt[p]{\sum_{i=1}^{n}|x_i-y_i|^p}$$

$$M\_dist=\sum_{i=1}^{n}|x_i-y_i| \qquad E\_dist=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \qquad C\_dist=\max_i(|x_i-y_i|)$$

相似度/距离公式2

- 夹角余弦相似度(Cosine)

$$a = (x_{11}, x_{12}, \ldots, x_{1n}), b = (x_{21}, x_{22}, \ldots, x_{2n})$$

$$\cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2} * \sqrt{\sum_{k=1}^{n} x_{2k}^2}} = \frac{a^T \cdot b}{|a||b|}$$

相似度/距离公式3

- 杰卡德相似系数(Jaccard)

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|} \qquad dist(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|}$$

- Pearson相关系数

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{\sum_{i=1}^{n}(X_i-\mu_X)(Y_i-\mu_Y)}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2}*\sqrt{\sum_{i=1}^{n}(Y_i-\mu_Y)^2}}$$

$$dist(X,Y)=1-\rho_{XY}$$


北风网
IBEIFENG.COM