机器学习 Introduction 作业

姓名： 张泽群 学号： 19049100002 班级：1班

Intro 2.1

答： (a) 手机背后的机器学习运用:

自然语言处理，语音识别，语言翻译，搜索引擎，广告推广，垃圾邮件过滤，偏好推荐系统

(b) 手机以外机器学习的应用:

在许多交叉学科中也常用到了机器学习技术，例如 生物信息学与医疗领域，电子商务和金融。

在应用领域方面，主要有如下多个方面：

1.金融领域：检测信用卡欺诈，证券市场分析等

2.医学领域：医疗诊断

3.自动化与机器人领域：无人驾驶，图像处理，信号处理等

4.生物领域：人体基因序列分析，蛋白质结构预测，DNA序列测序等

5.游戏领域：游戏战略规划

6.刑侦领域：潜在犯罪预测

7.气象领域：天气预测

Intro 2.2

答：**欧氏距离**

在数学中，欧几里得距离或欧几里得度量是欧几里得空间中两点间“普通”（即直线）距离。使用这个距离，欧氏空间成为度量空间。相关联的范数称为欧几里得范数。较早的文献称之为毕达哥拉斯度量。

在欧几里得空间中，点$X=(x_1,x_2,\ldots,x_n)$和$Y=(y_1,y_2,\ldots,y_n)$之间的欧氏距离为：

$$d(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}$$

曼哈顿距离

在欧几里得空间中，点$X=(x_1,x_2,\ldots,x_n)$和$Y=(y_1,y_2,\ldots,y_n)$之间的曼哈顿距离为：

$$d(X,Y)=|x_1-y_1|+|x_2-y_2|+\cdots+|x_n-y_n|$$

切比雪夫距离

点$X=(x_1,x_2,\ldots,x_n)$和$Y=(y_1,y_2,\ldots,y_n)$之间的，切比雪夫距离为：

$$d(X,Y)=\max_i(|x_i-y_i|)=\lim_{p\to\infty}\left(\sum_{i=1}^{n}|X_i-y_i|^p\right)^{1/p}$$

闵可夫斯基距离

闵氏距离又叫做闵可夫斯基距离，是欧氏空间中的一种测度，被看做是欧氏距离和曼哈顿距离的一种推广。闵氏距离不是一种距离，而是一组距离的定义，是对多个距离度量公式的概括性的表述。

在欧几里得空间中，点$X=(x_1,x_2,\ldots,x_n)$和$Y=(y_1,y_2,\ldots,y_n)$之间的，闵氏距离为：

$$d(X,Y)=\left(\sum_{i=1}^{n}|X_i-y_i|^p\right)^{1/p}$$

马氏距离

马氏距离(Mahalanobis Distance)是度量学习中一种常用的距离指标，同欧氏距离、曼哈顿距离、汉明距离等一样被用作评定数据之间的相似度指标。但却**可以应对高维线性分布的数据中各维度间非独立同分布的问题。**

数据点x,y之间的马氏距离：

$$d_M(x,y)=\sqrt{(x-y)^T\sum^{-1}(x-y)}$$

余弦距离

两点$X=(x_1,x_2,...,x_n)$和$Y=(y_1,y_2,...,y_n)$之间的余弦距离：

$$cos\theta=\frac{X\cdot Y}{|X||Y|}$$

即$cos\theta=\frac{\sum_{k=1}^{n}x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n}x_{1k}^2}\sqrt{\sum_{k=1}^{n}x_{2k}^2}}$

汉明距离

在信息论中，两个等长字符串之间的汉明距离，是两个字符串对应位置的不同字符的个数。换句话说，它就是将一个字符串变换成另外一个字符串所需要替换的字符个数。

如 X=0110， Y=0111 则汉明距离为 1

Jaccard距离

杰卡德相似系数(Jaccard similarity coefficient)：两个集合A和B的交集元素在A，B的并集中所占的比例，称为两个集合的杰卡德相似系数，用符号J (A , B) 表示：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}$$

杰卡德距离(Jaccard Distance)：与杰卡德相似系数相反，用两个集合中不同元素占所有元素的比例来衡量两个集合的区分度：

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

相关度量和相关距离

相关系数：是衡量随机变量X与Y相关程度的一种方法，相关系数的取值范围是[-1,1]。相关系数的绝对值越大，则表明X与Y相关度越高。当X与Y线性相关时，相关系数取值为1（正线性相关）或-1（负线性相关）：

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}}$$

相关距离：

$$D_{XY} = 1 - \rho_{XY}$$

信息熵

当给定随机变量X的条件下随机变量Y的熵可定义为条件熵$H(Y|X)$:

$$E(Y|X) = -\sum_{i=1}^{n} p_i E(Y|X = x_i)$$

Intro 2.3

(1) 建立[0,1]范围内均匀分布随机变量的概率密度函数（pdf），并由此导出cdf；

$$f_1(x) = \begin{cases} 0, & x < 0 \ \ or \ \ x > 1 \\ 1, & 0 <= x <= 1 \end{cases}$$

$$C_1(x) = \begin{cases} 0, & x < 0 \\ x, & 0 <= x <= 1 \\ 1, & x > 1 \end{cases}$$

(2) 公式化[5,10]中均匀分布随机变量的pdf，并由此导出cdf。

$$f_1(x) = \begin{cases} 0, & x < 5 \ \ or \ \ x > 10 \\ 0.2, & 5 <= x <= 10 \end{cases}$$

$$C_2(x) = \begin{cases} 0, & x < 5 \\ \frac{x-5}{5}, & 5 <= x <= 10 \\ 1, & x > 10 \end{cases}$$