lec02

Main Data Mining Problems

There are four fundamental problems:

- Association pattern mining
- Classification
- Clustering
- Outlier detection

Association pattern mining

The special case for it is **Frequent Pattern Mining** (binary data sets)

Given the $n \times d$ data matrix, identify all subsets of columns (features) such that at least a fraction s of rows (objects) in the matrix have all the features enabled.

Assume that $s = 0.65$, `{Milk, Butter, Bread}` are frequently bought together.

Transaction	Milk	Butter	Bread	Mushrooms	Onion	Carrot
1234	1	1	1	0	1	0
324	0	1	0	1	1	1
234	1	1	1	0	1	0
2125	1	1	1	1	0	1
113	1	0	0	1	1	0
5653	1	1	1	1	1	0

Classification

The goal is to use **training data** to learn relationships between a fixed feature called **class label** and the remaining features in the data (目标是使用训练数据来学习称为类标签的固定特征与数据中其余特征之间的关系)Classification is **supervised learning**

The resulting learned model may then be used to estimate/predict values of the class label for records, where the value is not known (然后，所得的学习模型可用于估计/预测记录的类标签值). The objects whose class label is unknown are test objects/data

Examples:

- Targeted marketing
- Text recognition

Clustering

Given a data set/matrix, partition (划分) its objects/rows into sets/clusters $C_1, C_2, \ldots, C_k$ such that the objects in each cluster are "similar" to one another. Specific definitions depend on how the notion of similarity is defined (具体定义取决于相似性概念的定义方式).

Clustering can be seen as an unsupervised version of classification

Examples:

- Customer segmentation (identify similar customers for targeted product promotion)
- Data summarisation (cluster can be used to create a summary of the data)

Outlier detection

Given a data set, determine the outliers, i.e. the objects that are **significantly different** from the remaining objects

Examples:

- Credit card fraud
- Detecting sensor events
- Medical diagnosis
- Earth science

Math Notions

Outer product

For $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, the outer product is

$$u \otimes v = \begin{pmatrix} u_1v_1 & u_1v_2 & \cdots & u_1v_n \\ u_2v_1 & u_2v_2 & \cdots & u_2v_n \\ \vdots & \vdots & \ddots & \vdots \\ u_mv_1 & u_mv_2 & \cdots & u_mv_n \end{pmatrix}$$

Linear independence (线性无关)

$a_1, a_2, \ldots, a_m$ 是 m 个向量, 对于方程 $\lambda_1 a_1 + \lambda_2 a_2 + \ldots + \lambda_m a_m = 0$, 如果其有非零解 $(\lambda_1, \lambda_2, \ldots, \lambda_m) \neq 0$, 则称 $a_1, a_2, \ldots, a_m$ 线性相关; 如果其有唯一解 $(\lambda_1, \lambda_2, \ldots, \lambda_m) = 0$, 则称 $a_1, a_2, \ldots, a_m$ 线性无关

直观的解释:
首先我们定义线性组合：$\vec{u}$ 与 $\vec{v}$ 的线性组合为 $a\vec{u} + b\vec{v}, (a, b \in \mathbb{R})$
但是什么是线性呢?

在基础数学中，我们很容易理解线性关系与非线性关系：$y = kx$，$y = kx^2$
线性，即**可加**与**数乘**。线性代数中，我们处理的问题全部都是线性的（非线性也太难了）
而我们对多个向量进行数乘与加法，这个结果就是一个线性组合
需要特别注意的是，**仿射变换**经常会被误认为线性变换（虽然我不认为这节课会涉及）

回到线性无关/相关
现在，我们只有一个向量 $\vec{u}$，对其进行数乘得到 $a\vec{u}$，由于 a 可以取全体实数，$a\vec{u}$ 的落点就是一条直线。我们可以说，这条直线就是 $a\vec{u}$ 张成的空间

- 对于两个**不共线**的向量 $\vec{u}$ 与 $\vec{v}$，他们线性组合 $a\vec{u} + b\vec{v}, (a, b \in \mathbb{R})$ 的所有落点可以构成 $\vec{u}, \vec{v}$ 所在平面，这个平面就是这个线性组合的**张成空间**，也可以说这个线性组合可以表示在这个平面内的所有向量
- 而当他们**共线**时，这两个向量的张成空间就只剩下他们的所在直线了；这其实意味着对于任意一个向量，我们都可以对另一个向量进行数乘来得到它，这两个向量是**线性相关**的。

推广到 m 维便是线性无关的定义了，我们用三维空间进行直观的解释
对于向量 $\vec{u}, \vec{v}, \vec{w}$

- 当他们线性无关时，这三个向量的张成空间就是三维空间，此时对于方程 $\lambda_1\vec{u} + \lambda_2\vec{v} + \lambda_3\vec{w} = 0$，有且只有一组解 $(\lambda_1, \lambda_2, \lambda_3) = 0$
- 当他们线性相关时，这三个向量的张成空间不可能为三维空间，这说明其中一个向量与另外两个向量共面，或者这三个向量共线；且对于上述方程，具有非 0 解

Rank (秩)

矩阵的列秩是矩阵列向量生成的最大线性无关组的向量个数，行秩类似
矩阵的行秩总是等于列秩

Trace (迹)

对角线元素的和

Eigenvalues & Eigenvectors (特征值与特征向量)

对于矩阵 A 与非零向量 x 以及实数 λ，满足：

$$Ax = \lambda x$$

对于矩阵 A，一定存在一个向量 x，使得 x 在经过矩阵 A 的线性变换后方向没有发生改变；那么这个变换其实就相当于 x 与一个常数相乘，x 就是矩阵 A 的**特征向量**，λ 就是**特征值**

如何求解 $Ax = \lambda x$
易得：$(A - \lambda I)x = 0$
对于 $|A - \lambda I|$，将其展开

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix}$$

这是一个以 λ 为未知数的一元 n 次方程组。观察发现，λ 只出现在对角线上，显然 A 的特征值就是方程组的解。因为 n 次方程组在复数集内有 n 个解，因此矩阵 A 在复数集内有 n 个特征值，有

- $\lambda_1 + \lambda_2 + \cdots + \lambda_n = a_{11} + a_{22} + a_{nn}$
- $\lambda_1 \lambda_2 \cdots \lambda_n = |A|$

其中，λ_i 为一元 n 次方程组的第 i 个根

Mathematical Preliminaries - Continuous Optimisation

- Unconstrained optimisation
- Constrained optimisation

Gradient Descent Method (梯度下降法)

Uncon

$X = (x_1, x_2, \ldots, x_d)$，$f(X) = f(x_1, \ldots, x_d)$

$$\nabla_X f = \frac{\partial f}{\partial X} = \begin{pmatrix} \frac{\partial f(X)}{\partial x_1} \\ \frac{\partial f(X)}{\partial x_2} \\ \vdots \\ \frac{\partial f(X)}{\partial x_d} \end{pmatrix}$$

对于初始点 X_0，函数 f 沿着 X_0 的负梯度方向下降最快，即，$-(\nabla_X f)(X_0)$
对于一小步 $\gamma \geq 0$，$f\big(X_0 - \gamma \cdot (\nabla_X f)(X_0)\big) \leq f(X_0)$

而对于步骤 i，有

$$X_{i+1} = X_i - \gamma_i \cdot \big((\nabla_X f)(X_i)\big)$$

Partial Derivative

对于 $f(x,y) = 5x + y^2$，$\nabla_{(x,y)} f = (5, 2y)^T$

Lagrange Multipliers (拉格朗日乘数)

Con

问题描述：对于 $X = (x_1, x_2, \dots, x_d)$，约束条件 $g(X) = 0$，我们需要找到 $f(X)$ 在其下的极值

1. 构造拉格朗日函数 $\mathscr{L}(X, \lambda) = f(X) - \lambda \cdot g(X)$
2. 找到 $\mathscr{L}(X, \lambda)$ 所有 stationary points (驻点) (X_0, λ_0)，即 $\mathscr{L}(X, \lambda)$ 的所有偏导都等于 0 的点，$\nabla_{(X,\lambda)} \mathscr{L} = 0$
3. 检查这些点，寻找方案

考虑到 n 元函数 $f(x_1, x_2, \dots, x_n)$ 在 m 个约束条件

$$\begin{cases} g_1(x_1, x_2, \dots, x_n) = 0 \\ g_2(x_1, x_2, \dots, x_n) = 0 \\ \qquad\qquad \dots \\ g_m(x_1, x_2, \dots, x_n) = 0 \end{cases}$$

下的极值

此时拉格朗日函数为

$$\begin{aligned} &\mathscr{L}(x_1, x_2, \dots, x_n, \lambda_1, \lambda_2, \dots, \lambda_m) \\ =&f(x_1, x_2, \dots, x_n) + \lambda_1 g_1(x_1, x_2, \dots, x_n) \\ &+ \lambda_2 g_2(x_1, x_2, \dots, x_n) + \dots + \lambda_m g_m(x_1, x_2, \dots, x_n) \\ =&f(x_1, x_2, \dots, x_n) + \sum_{k=1}^{m} \lambda_k g_k(x_1, x_2, \dots, x_n) \end{aligned}$$