



浙江工商大学  
ZHEJIANG GONGSHANG UNIVERSITY

人工智能：模型与算法

# 统计机器学习算法应用

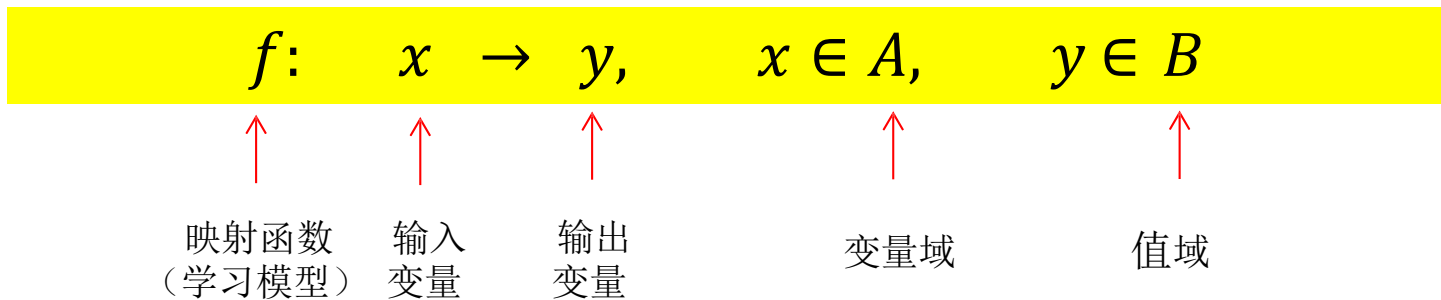


# 目录

- 01** 基于逻辑斯蒂回归模型的分类
- 02** 基于矩阵分解的潜在语义分析
- 03** 线性判别分析及分类

# 回归与分类的区别

- 回归与分类均是挖掘和学习输入变量和输出变量之间潜在关系模型，基于学习所得模型将输入变量映射到输出变量。

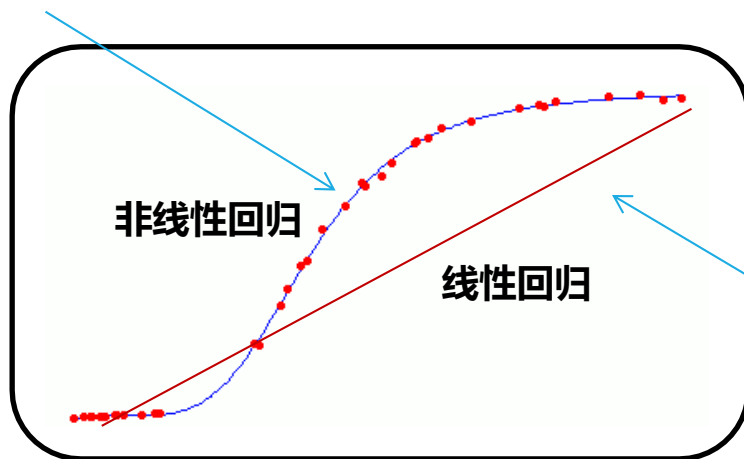


- 在回归分析中，学习得到一个函数将输入变量映射到连续输出空间，如价格和温度等，即值域是连续空间。
- 在分类模型中，学习得到一个函数将输入变量映射到离散输出空间，如人脸和汽车等，即值域是离散空间。

问题：回归与分类可否统一，即用回归模型来完成分类任务？

# 回归分析：从线性到非线性

非线性回归模型  $y = \phi(x)$



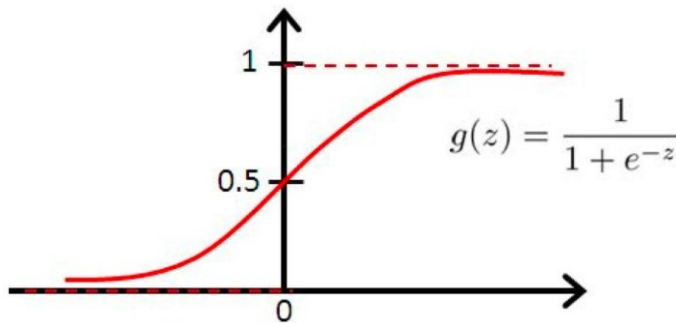
线性回归模型  $y = ax + b$

线性回归模型难以刻画数据的复杂分布，需要寻找非线性回归模型

## 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

逻辑斯蒂回归(logistic regression)就是在回归模型中引入 sigmoid函数的一种非线性回归模型。

Logistic回归模型可如下表示：



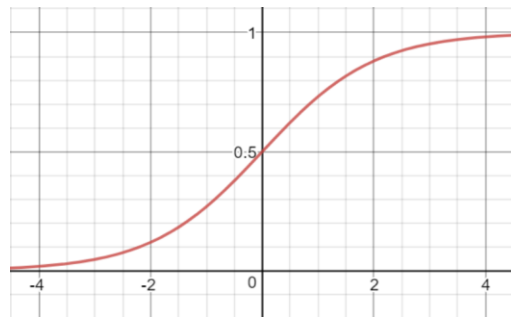
$$y = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}} \quad , \quad \text{其中 } y \in (0,1), z = \mathbf{w}^T \mathbf{x} + b$$

这里  $\frac{1}{1 + e^{-z}}$  是sigmoid函数、 $\mathbf{x} \in \mathcal{R}^d$  是输入数据、 $\mathbf{w} \in \mathcal{R}^d$  和  $b \in \mathcal{R}$  是回归函数的参数

# 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

sigmoid函数 $\frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(w^T x+b)}}$ 的性质

- **概率形式输出。** sigmoid函数是单调递增的，其值域为(0, 1)，因此使sigmoid函数输出可作为概率值。在前面介绍的线性回归中，回归函数的定义域一般为 $(-\infty, +\infty)$
- **数据特征加权累加。** 对输入 $z$ 取值范围没有限制，但当 $z$ 大于一定数值后，函数输出无限趋近于1，而小于一定数值后，函数输出无限趋近于0。特别地，当 $z = 0$ 时，函数输出为0.5。这里 $z$ 是输入数据 $x$ 和回归函数的参数 $w$ 相乘结果（可视为 $x$ 各维度进行加权叠加）
- **非线性变化。**  $x$ 各维度加权叠加之和结果取值在0附近时，函数输出值的变化幅度比较大（函数值变化陡峭），且是非线性变化。但是，各维度加权叠加之和结果取值很大或很小时，函数输出值几乎不变化，这是基于概率的一种认识与需要。



Sigmoid函数的可视化

# 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

从回归到分类：概率输出

$$y = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$$



输入数据 $\mathbf{x}$ 属于正例的概率，  
这里 $y$ 理解为 $\mathbf{x}$ 为正例的概率、  
 $1 - y$ 理解为 $\mathbf{x}$ 为负例的概率，  
即 $p(y = 1|\mathbf{x})$

$$\frac{p}{1 - p}$$



几率(odds)： $\mathbf{x}$ 作为正  
例的相对可能性  
( $p$  为正例的概率)


$$\log\left(\frac{p}{1 - p}\right)$$



对数几率(log  
odds)或logit

# 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

概率输出：从回归到分类


$$y = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$$


- $p(y = 1 | \mathbf{x}) = h_{\theta}(x) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$
- $p(y = 0 | \mathbf{x}) = 1 - h_{\theta}(x) = \frac{e^{-(\mathbf{w}^T \mathbf{x} + b)}}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$

$\theta$ 表示模型参数 ( $\theta = \{\mathbf{w}, b\}$ )



- 对 $x$ 作为正例可能性取对数得到线性回归模型
- $x$ 为正例的概率越大，几率取值就越大
- 线性回归模型输出结果去逼近（拟合）真实标记结果的对数几率
- 逻辑斯蒂回归函数被称为“对数几率回归(log-odds regression)”。

$$\begin{aligned}\text{logit}(p(y = 1 | \mathbf{x})) &= \log \left( \frac{p(y = 1 | \mathbf{x})}{p(y = 0 | \mathbf{x})} \right) \\ &= \log \left( \frac{p}{1 - p} \right) = \mathbf{w}^T \mathbf{x} + b\end{aligned}$$


线性回归方程



# 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

从回归到分类：概率输出

$$\text{logit}(p(y = 1|\mathbf{x})) = \mathbf{w}^T \mathbf{x} + b$$

- 对数几率回归模型的输出 $y$ 可作为将输入数据 $\mathbf{x}$ 分类为某一类别概率的大小。
- 输出值越接近1，说明输入数据 $\mathbf{x}$ 分类为该类别的可能性越大。与此相反，输出值越接近0，输入数据 $\mathbf{x}$ 不属于该类别的概率越大。
- 根据具体应用设置一个阈值，将大于该阈值的输入数据 $\mathbf{x}$ 都归属到某个类别，小于该阈值的输入数据 $\mathbf{x}$ 都归属到另外一个类别。

# 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

从回归到分类：概率输出

$$\text{logit}(p(y = 1|\mathbf{x})) = \mathbf{w}^T \mathbf{x} + b$$

- 如果输入数据  $\mathbf{x}$  属于正例的概率大于其属于负例的概率，即  $p(y = 1|\mathbf{x}) > 0.5$ ，则输入数据 $\mathbf{x}$ 可被判断属于正例。

这一结果等价于  $\frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} > 1$ ，即  $\log \left( \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} \right) > \log 1 = 0$ ，也就是  $\mathbf{w}^T \mathbf{x} + b > 0$  成立（ $\mathbf{x}$ 属于正例时）。

- 从这里可以看出，**logistic回归是一个线性模型**。在预测时，可以通过计算线性函数  $\mathbf{w}^T \mathbf{x} + b$  取值是否大于0来判断输入数据 $\mathbf{x}$ 的类别归属。

# 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

从回归到分类：参数求取

模型参数的似然函数被定义为 $\mathcal{L}(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)$ ，其中 $\mathcal{D} = \{(x_i, y_i) | 1 \leq i \leq n\}$ 表示所有观测数据（或训练数据）， $\theta$ 表示模型参数（ $\theta = \{\mathbf{w}, b\}$ ）



每一个样本数据是独立同分布 (independent and identically distributed, i.i.d) ，于是可得：

$$\mathcal{L}(\theta|\mathcal{D}) = p(\mathcal{D}|\theta) = \prod_{i=1}^n P(y_i|x, \theta) = \prod_{i=1}^n (h_{\theta}(x_i))^{y_i} (1 - h_{\theta}(x_i))^{1-y_i}$$

对上述公式取对数(对数似然)：

$$l(\theta) = \log(\mathcal{L}(\theta|\mathcal{D})) = \sum_{i=1}^n y_i \log(h_{\theta}(x_i)) + \sum_{i=1}^n (1 - y_i) \log(1 - h_{\theta}(x_i))$$

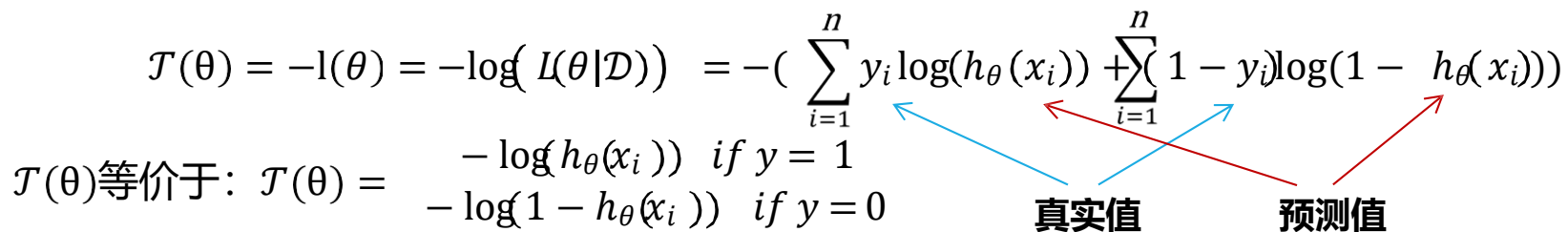
# 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

## 从回归到分类：参数求取

最大似然估计目的是计算似然函数的最大值，而分类过程是需要损失函数最小化。因此，在上式前加一个负号得到损失函数，**这一损失函数就是对数似然的相反数**（Negative log-likelihood log，又叫交叉熵）：

$$\mathcal{J}(\theta) = -l(\theta) = -\log(\mathcal{L}(\theta|\mathcal{D})) = -\left( \sum_{i=1}^n y_i \log(h_{\theta}(x_i)) + \sum_{i=1}^n (1 - y_i) \log(1 - h_{\theta}(x_i)) \right)$$

$\mathcal{J}(\theta)$ 等价于： $\mathcal{J}(\theta) = \begin{cases} -\log(h_{\theta}(x_i)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x_i)) & \text{if } y = 0 \end{cases}$

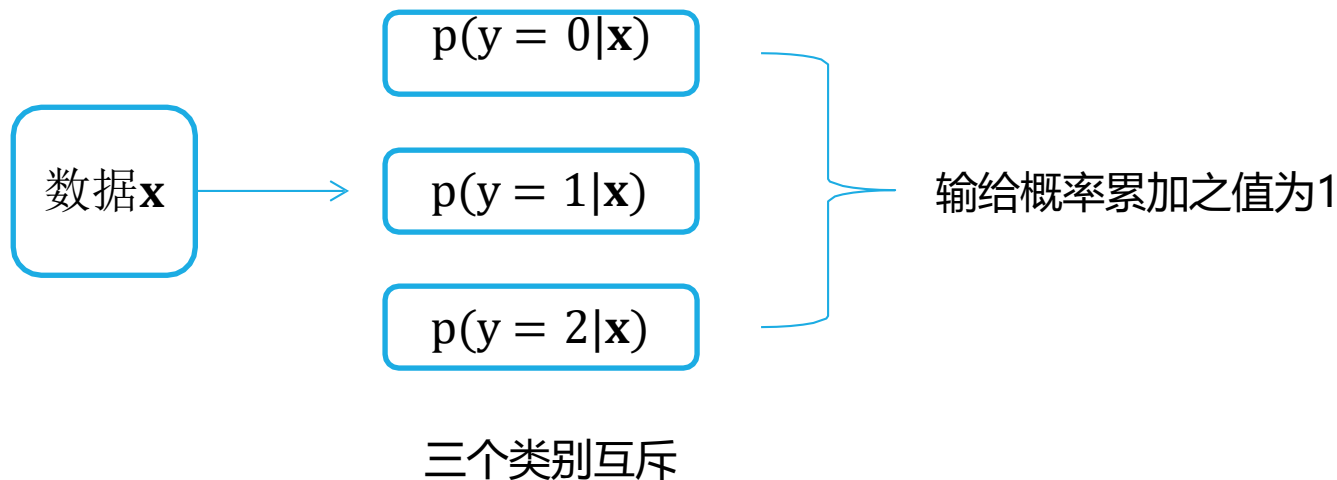


一旦推导得到了逻辑回归的损失函数 $\mathcal{J}(\theta)$ ，需要通过最小化损失函数来求解模型的参数 $\theta$ ，即输入数据 $x$ 各维度的加权系数。对于线性回归模型而言，可以使用**最小二乘法**，但对于逻辑斯蒂回归而言使用传统最小二乘法求解是不合适的，需要考虑使用**迭代算法**进行优化求解，常见的就是“**梯度下降法(gradient descent)**”。

# 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

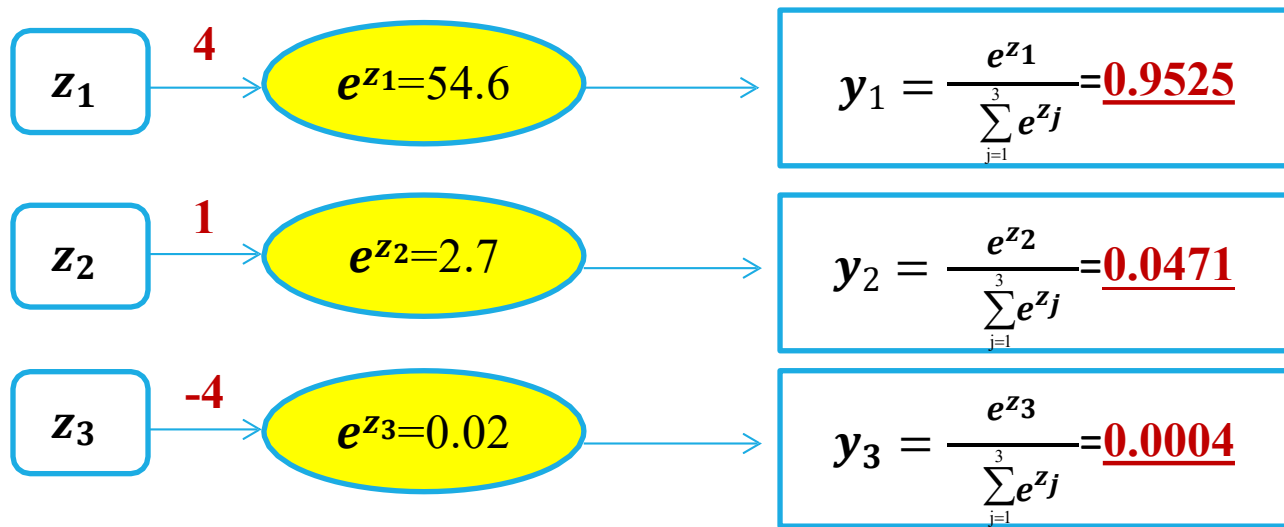
从回归到分类：从两类分类到多类分类

logistic回归只能用于解决二分类问题，将它进行推广为多项逻辑斯蒂回归模型(multi-nominal logistic model, 也即softmax函数)，用于处理多类分类问题，可以得到处理多类分类问题的softmax回归。



# 非线性回归分析模型：逻辑斯蒂回归(logistic regression)

从回归到分类(softmax分类)：从两类分类到多类分类



指数级扩大最后一层输出，每个输出值都会增大，然而值最大的输出相比其他值扩大更多，然后再将所有结果归一化到(0,1)概率区间

$$0 < y_i < 1, y_i = 1$$



- 01** 基于逻辑斯蒂回归模型的分类
- 02** 基于矩阵分解的潜在语义分析
- 03** 线性判别分析及分类

# 基于矩阵分解的潜在语义分析

潜在语义分析 (Latent Semantic Analysis, LSA或者Latent Semantic Indexing, LSI) 是一种从海量文本数据中学习单词-单词、单词-文档以及文档-文档之间隐性关系，进而得到文档和单词表达特征的方法。该方法的基本思想是综合考虑某些单词在哪些文档中同时出现，以此来决定该词语的含义与其他的词语的相似度。

潜在语义分析先构建一个单词-文档 (term-document) 矩阵 $A$ ，进而寻找该矩阵的低秩逼近 (low rank approximation)，来挖掘单词-单词、单词-文档以及文档-文档之间的关联关系。



# 基于矩阵分解的潜在语义分析

## 单词-文档矩阵(term-document): 构造与分解

通过如下九篇文章来解释LSA方法。选取每篇文章的标题

- a1: Efficient Algorithms for Non-convex Isotonic Regression through Submodular Optimization
- a2: Combinatorial Optimization with Graph Convolutional Networks and Guided Tree Search
- a3: An Improved Analysis of Alternating Minimization for Structured Multi-Response Regression
- a4: Analysis of Krylov Subspace Solutions of Regularized Non-Convex Quadratic Problems
- a5: Post: Device Placement with Cross-Entropy Minimization and Proximal Policy Optimization
- b1: CRISPR/Cas9 and TALENs generate heritable mutations for genes involved in small RNA processing of Glycine max and Medicago truncatula
- b2: Generation of D1-1 TALEN isogenic control cell line from Dravet syndrome patient iPSCs using TALEN-mediated editing of the SCN1A gene
- b3: Genome-Scale CRISPR Screening Identifies Novel Human Pluripotent Gene Networks
- b4: CHAMPIONS: A phase 1/2 clinical trial with dose escalation of SB-913 ZFN-mediated in vivo human genome editing for treatment of MPS II (Hunter syndrome)

**机器学习 (Machine Learning) 类别五篇文章**

**基因编辑 (gene editing) 类别四篇文章**

# 基于矩阵分解的潜在语义分析

## 单词-文档矩阵(term-document): 构造与分解

从这九篇论文标题中，筛选有实际意义且至少出现在两篇文章标题中的十个单词，分别是nonconvex, regression, optimization, network, analysis, minimization, gene, syndrome, editing, human。这样，十个单词和九篇文章就可以形成一个 $10 \times 9$ 大小的单词-文档矩阵A。

	a1	a2	a3	a4	a5	b1	b2	b3	b4
nonconvex	1	0	0	1	0	0	0	0	0
regression	1	0	1	0	0	0	0	0	0
optimization	1	1	0	0	1	0	0	0	0
network	0	1	0	0	0	0	0	1	0
analysis	0	0	1	1	0	0	0	0	0
minimization	0	0	1	0	1	0	0	0	0
gene	0	0	0	0	0	1	1	1	0
syndrome	0	0	0	0	0	0	1	0	1
editing	0	0	0	0	0	0	1	0	1
human	0	0	0	0	0	0	0	1	1

单词-文档矩阵中每一行表示某个单词在不同文档标题中所出现次数，比如单词regression分别在文档a1和文档a3的标题中各出现了一次，那么这两处相应位置值为1。

# 基于矩阵分解的潜在语义分析

## 单词-文档矩阵(term-document): 构造与分解

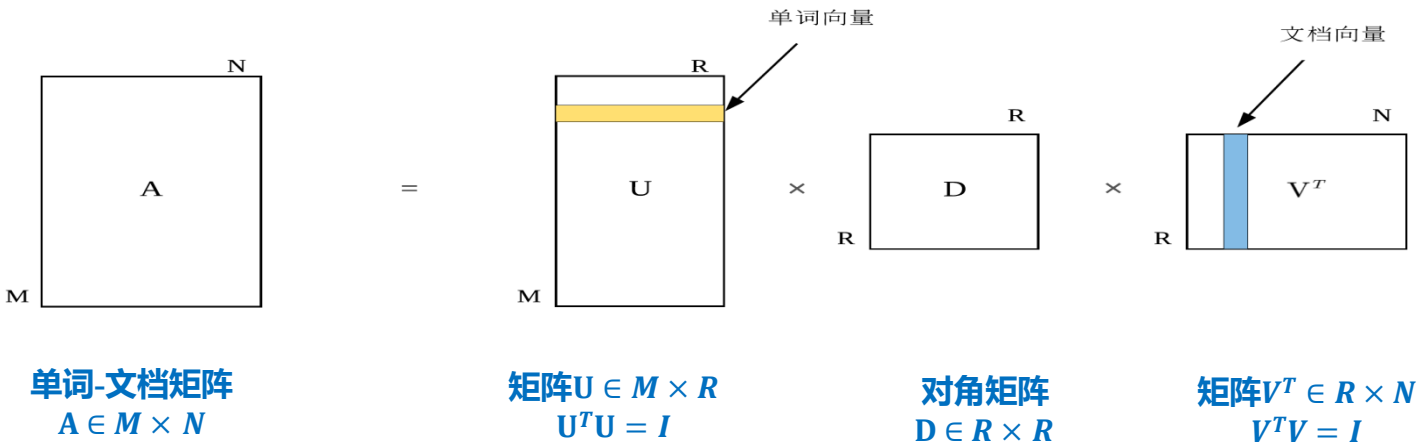
	a1	a2	a3	a4	a5	b1	b2	b3	b4
nonconvex	1	0	0	1	0	0	0	0	0
regression	1	0	1	0	0	0	0	0	0
optimization	1	1	0	0	1	0	0	0	0
network	0	1	0	0	0	0	0	1	0
analysis	0	0	1	1	0	0	0	0	0
minimization	0	0	1	0	1	0	0	0	0
gene	0	0	0	0	0	1	1	1	0
syndrome	0	0	0	0	0	0	1	0	1
editing	0	0	0	0	0	0	1	0	1
human	0	0	0	0	0	0	0	1	1

- 当用户输入“optimization”这一检索请求，由于文档a3标题中不包含这一单词，则文档a3被认为是不相关文档，但实际上文档a3所涉及“minimization”内容与优化问题相关。出现这一问题是因为单词-文档矩阵只是刻画了单词是否在文档中出现与否这一现象，而无法对单词-单词、单词-文档以及文档-文档之间语义关系进行建模。
- 如果用户检索“eat an apple”，则文档“Apple is a great company”会被检索出来，而实际上该文档中单词“Apple”所指苹果公司、而非水果，造成这一结果的原因是一些单词具有“一词多义”。
- 因此需要一种方法能够建模单词-单词、单词-文档以及文档-文档之间语义关系，解决包括“异词同义”和“一词多义”在内的诸多挑战。

# 基于矩阵分解的潜在语义分析

单词-文档矩阵(term-document): 构造与分解  $A = UDV^T$

- 奇异值分解(Singular Value Decomposition, SVD)将一个矩阵分解为两个正交矩阵与一个对角矩阵的乘积
- 单词个数为 $M$ 、文档个数为 $N$



# 基于矩阵分解的潜在语义分析

## 单词-文档矩阵(term-document): 构造与分解

a1	a2	a3	a4	a5	b1	b2	b3	b4
1	0	0	1	0	0	0	0	0
1	0	1	0	0	0	0	0	0
1	1	0	0	1	0	0	0	0
0	1	0	0	0	0	0	1	0
0	0	1	1	0	0	0	0	0
0	0	1	0	1	0	0	0	0
0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	1	1

$$U = \begin{pmatrix} -0.09 & -0.38 & -0.14 & -0.04 & 0.71 & -0.02 & 0.12 & 0.33 & 0.45 & -0. \\ -0.1 & -0.47 & -0.2 & -0.06 & 0. & -0.06 & -0.76 & -0.38 & -0.01 & 0. \\ -0.18 & -0.53 & 0.44 & 0.51 & 0. & -0.14 & 0.16 & 0.09 & -0.42 & -0. \\ -0.28 & -0.08 & 0.57 & -0.23 & -0. & 0.32 & 0.17 & -0.5 & 0.38 & 0. \\ -0.07 & -0.35 & -0.4 & -0.46 & -0. & 0.22 & 0.46 & -0.19 & -0.45 & -0. \\ -0.09 & -0.38 & -0.14 & -0.04 & -0.71 & -0.02 & 0.12 & 0.33 & 0.45 & -0. \\ -0.52 & 0.13 & 0.16 & -0.47 & 0. & -0.65 & -0.04 & 0.16 & -0.11 & -0. \\ -0.45 & 0.16 & -0.32 & 0.34 & -0. & 0.01 & 0.11 & -0.19 & 0.08 & 0.71 \\ -0.45 & 0.16 & -0.32 & 0.34 & -0. & 0.01 & 0.11 & -0.19 & 0.08 & -0.71 \\ -0.42 & 0.1 & 0.09 & -0.11 & -0. & 0.63 & -0.32 & 0.49 & -0.22 & -0. \end{pmatrix}$$

$$D = \begin{pmatrix} 2.46 & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\ 0. & 2.35 & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\ 0. & 0. & 1.79 & 0. & 0. & 0. & 0. & 0. & 0. \\ 0. & 0. & 0. & 1.51 & 0. & 0. & 0. & 0. & 0. \\ 0. & 0. & 0. & 0. & 1.41 & 0. & 0. & 0. & 0. \\ 0. & 0. & 0. & 0. & 0. & 1.23 & 0. & 0. & 0. \\ 0. & 0. & 0. & 0. & 0. & 0. & 0.93 & 0. & 0. \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.73 & 0. \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.15 \end{pmatrix}$$

$$V^T = \begin{pmatrix} -0.15 & -0.18 & -0.1 & -0.06 & -0.11 & -0.21 & -0.58 & -0.49 & -0.54 \\ -0.59 & -0.26 & -0.51 & -0.31 & -0.39 & 0.06 & 0.19 & 0.07 & 0.18 \\ 0.05 & 0.57 & -0.42 & -0.3 & 0.17 & 0.09 & -0.26 & 0.46 & -0.3 \\ 0.27 & 0.18 & -0.37 & -0.33 & 0.32 & -0.31 & 0.13 & -0.54 & 0.38 \\ 0.5 & 0. & -0.5 & 0.5 & -0.5 & 0. & 0. & 0. & -0. \\ -0.19 & 0.15 & 0.11 & 0.16 & -0.14 & -0.53 & -0.52 & 0.25 & 0.53 \\ -0.51 & 0.36 & -0.19 & 0.62 & 0.3 & -0.04 & 0.21 & -0.2 & -0.1 \\ 0.05 & -0.56 & -0.33 & 0.19 & 0.58 & 0.21 & -0.3 & 0.2 & 0.16 \\ 0.07 & -0.27 & -0.07 & 0. & 0.14 & -0.72 & 0.38 & 0.33 & -0.36 \end{pmatrix}$$

# 基于矩阵分解的潜在语义分析

## 单词-文档矩阵(term-document): 构造与分解

选取最大的前两个特征根及其对应的特征向量对矩阵A进行重建。下面给出了选取矩阵U、矩阵D和矩阵V的子部分重建所得矩阵 $A_2$

	a1	a2	a3	a4	a5	b1	b2	b3	b4
nonconvex	1	0	0	1	0	0	0	0	0
regression	1	0	1	0	0	0	0	0	0
optimization	1	1	0	0	1	0	0	0	0
network	0	1	0	0	0	0	0	1	0
analysis	0	0	1	1	0	0	0	0	0
minimization	0	0	1	0	1	0	0	0	0
gene	0	0	0	0	0	1	1	1	0
syndrome	0	0	0	0	0	0	1	0	1
editing	0	0	0	0	0	0	1	0	1
human	0	0	0	0	0	0	0	1	1

$$A_2 = \left\{ \begin{array}{l} \begin{array}{cc} & \begin{array}{cccccc} a1 & a2 & a3 & a4 & a5 & b1 & b2 & b3 & b4 \end{array} \\ \begin{array}{l} nonconvex \\ regression \\ optimization \\ network \\ analysis \\ minimization \\ gene \\ syndrome \\ editing \\ human \end{array} & \begin{array}{cccccccc} 0.56 & 0.27 & 0.49 & 0.3 & 0.37 & -0.01 & -0.05 & 0.05 & -0.05 \\ 0.68 & 0.33 & 0.59 & 0.36 & 0.45 & -0.01 & -0.06 & 0.05 & -0.06 \\ 0.79 & 0.4 & 0.68 & 0.41 & 0.53 & 0.02 & 0.02 & 0.14 & 0.02 \\ 0.22 & 0.18 & 0.17 & 0.1 & 0.15 & 0.13 & 0.36 & 0.32 & 0.33 \\ 0.51 & 0.25 & 0.44 & 0.27 & 0.34 & -0.01 & -0.06 & 0.03 & -0.06 \\ 0.56 & 0.27 & 0.49 & 0.3 & 0.37 & -0.01 & -0.05 & 0.05 & -0.05 \\ 0.01 & 0.16 & -0.03 & -0.02 & 0.02 & 0.29 & 0.81 & 0.66 & 0.75 \\ -0.05 & 0.11 & -0.07 & -0.05 & -0.02 & 0.26 & 0.72 & 0.58 & 0.67 \\ -0.05 & 0.11 & -0.07 & -0.05 & -0.02 & 0.26 & 0.72 & 0.58 & 0.67 \\ 0.01 & 0.13 & -0.02 & -0.01 & 0.02 & 0.23 & 0.65 & 0.53 & 0.6 \end{array} \end{array} \right\}$$

- 回到之前举的一个例子，用户输入“optimization”来检索与之相关的文档。尽管单词“optimization”在文档a3中没有出现，但是在重建矩阵 $A_2$ 中，对应的位置被0.68取代，说明单词“optimization”对表征文档a3所蕴含内容具有重要作用，这也符合文档a3描述的minimization问题是一个optimization问题的事实。
- 在单词-矩阵A中，文档b3所对应network、gene和human三个单词取值为1，在重建矩阵 $A_2$ 中，network、gene和human三个单词取值分别为0.32、0.66和0.53。可见，network在表征文档b3时重要性降低，因为算法认为这一单词在机器学习所相关文档表达中更具有区别性。

# 基于矩阵分解的潜在语义分析

## 单词-文档矩阵(term-document): 构造与分解

选取最大的前两个特征根及其对应的特征向量对矩阵 $A$ 进行重建。下面给出了选取矩阵 $U$ 、矩阵 $D$ 和矩阵 $V$ 的子部分重建所得矩阵 $A_2$

$$A_2 = \left\{ \begin{array}{cccccc} & a1 & a2 & a3 & a4 & a5 & b1 & b2 & b3 & b4 \\ nonconvex & 0.56 & 0.27 & 0.49 & 0.3 & 0.37 & -0.01 & -0.05 & 0.05 & -0.05 \\ regression & 0.68 & 0.33 & 0.59 & 0.36 & 0.45 & -0.01 & -0.06 & 0.05 & -0.06 \\ optimization & 0.79 & 0.4 & 0.68 & 0.41 & 0.53 & 0.02 & 0.02 & 0.14 & 0.02 \\ network & 0.22 & 0.18 & 0.17 & 0.1 & 0.15 & 0.13 & 0.36 & 0.32 & 0.33 \\ analysis & 0.51 & 0.25 & 0.44 & 0.27 & 0.34 & -0.01 & -0.06 & 0.03 & -0.06 \\ minimization & 0.56 & 0.27 & 0.49 & 0.3 & 0.37 & -0.01 & -0.05 & 0.05 & -0.05 \\ gene & 0.01 & 0.16 & -0.03 & -0.02 & 0.02 & 0.29 & 0.81 & 0.66 & 0.75 \\ syndrome & -0.05 & 0.11 & -0.07 & -0.05 & -0.02 & 0.26 & 0.72 & 0.58 & 0.67 \\ editing & -0.05 & 0.11 & -0.07 & -0.05 & -0.02 & 0.26 & 0.72 & 0.58 & 0.67 \\ human & 0.01 & 0.13 & -0.02 & -0.01 & 0.02 & 0.23 & 0.65 & 0.53 & 0.6 \end{array} \right\}$$

- 由于 $A_2$ 是从最大两个特征根及其对应特征向量重建得到, 因此 $A_2$ 与 $A$ 不是完全一样的, 两者存在一定的误差
- $A_2$ 捕获得到了原始单词-文档矩阵 $A$ 中所蕴含的单词与单词之间的关联关系
- 如果两个单词在原始单词-文档矩阵 $A$ 中分布一致, 则其在重建矩阵 $A_2$ 中分布也可能一致的, 如editing和syndrome。
- 对于归属于同一类别文档的单词, 可以发现它们之间的值彼此接近, 而与不是归属于同一个类别中的单词不相似, 如minimization在机器学习类别文档中均为正数、其在基因编辑类别文档中几乎为负数。

# 基于矩阵分解的潜在语义分析

## 单词-文档矩阵(term-document): 构造与分解

基于单词-文档矩阵 $A$ , 对 $A^T A$ 结果归一化可得到如下文档-文档相关系数矩阵:

$$\begin{pmatrix} & a1 & a2 & a3 & a4 & a5 & b1 & b2 & b3 & b4 \\ a1 & 1. & 0.22 & 0.05 & 0.22 & 0.22 & -0.22 & -0.43 & -0.43 & -0.43 \\ a2 & 0.22 & 1. & -0.33 & -0.25 & 0.37 & -0.17 & -0.33 & 0.22 & -0.33 \\ a3 & 0.05 & -0.33 & 1. & 0.22 & 0.22 & -0.22 & -0.43 & -0.43 & -0.43 \\ a4 & 0.22 & -0.25 & 0.22 & 1. & -0.25 & -0.17 & -0.33 & -0.33 & -0.33 \\ a5 & 0.22 & 0.37 & 0.22 & -0.25 & 1. & -0.17 & -0.33 & -0.33 & -0.33 \\ b1 & -0.22 & -0.17 & -0.22 & -0.17 & -0.17 & 1. & 0.51 & 0.51 & -0.22 \\ b2 & -0.43 & -0.33 & -0.43 & -0.33 & -0.33 & 0.51 & 1. & 0.05 & 0.52 \\ b3 & -0.43 & 0.22 & -0.43 & -0.33 & -0.33 & 0.51 & 0.05 & 1. & 0.05 \\ b4 & -0.43 & -0.33 & -0.43 & -0.33 & -0.33 & -0.22 & 0.52 & 0.05 & 1. \end{pmatrix}$$

基于重建单词-文档矩阵 $A_2$ , 对 $A_2^T A_2$ 结果归一化可得到如下文档-文档相关系数矩阵:

$$\begin{pmatrix} & a1 & a2 & a3 & a4 & a5 & b1 & b2 & b3 & b4 \\ a1 & 1. & 0.97 & 1. & 1. & 1. & -0.95 & -0.95 & -0.93 & -0.95 \\ a2 & 0.97 & 1. & 0.97 & 0.97 & 0.98 & -0.85 & -0.86 & -0.83 & -0.86 \\ a3 & 1. & 0.97 & 1. & 1. & 1. & -0.95 & -0.96 & -0.94 & -0.96 \\ a4 & 1. & 0.97 & 1. & 1. & 1. & -0.95 & -0.96 & -0.94 & -0.96 \\ a5 & 1. & 0.98 & 1. & 1. & 1. & -0.94 & -0.95 & -0.93 & -0.95 \\ b1 & -0.95 & -0.85 & -0.95 & -0.95 & -0.94 & 1. & 1. & 1. & 1. \\ b2 & -0.95 & -0.86 & -0.96 & -0.96 & -0.95 & 1. & 1. & 1. & 1. \\ b3 & -0.93 & -0.83 & -0.94 & -0.94 & -0.93 & 1. & 1. & 1. & 1. \\ b4 & -0.95 & -0.86 & -0.96 & -0.96 & -0.95 & 1. & 1. & 1. & 1. \end{pmatrix}$$

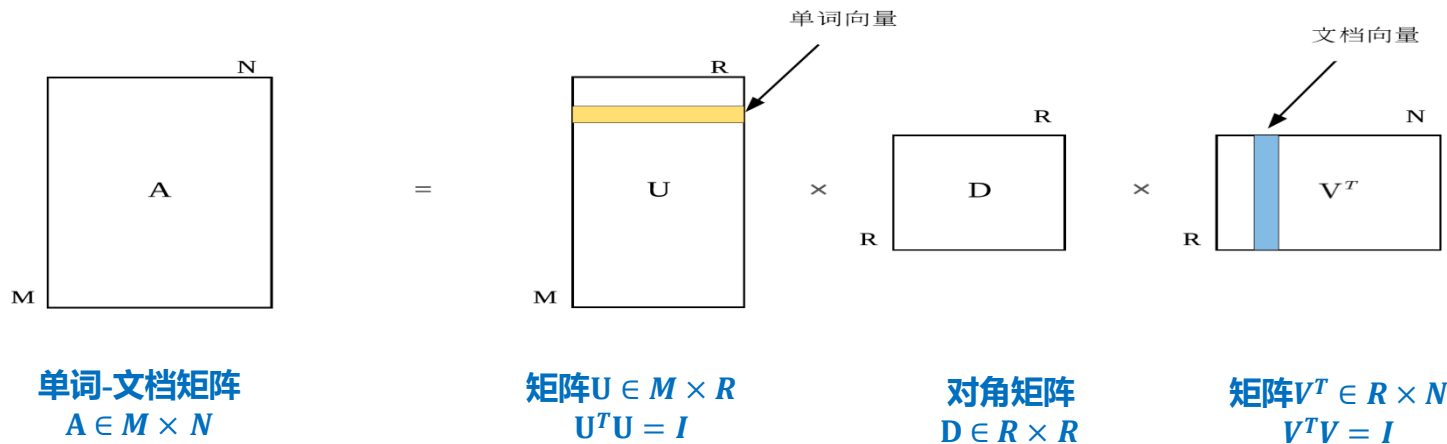
- 在基于单词-文档矩阵 $A$ 所构建的文档-文档相互关系矩阵中, 同一类别文档之间的相关性系数普遍很低, 比如 $a1-a3$ ,  $a2-a3$ ,  $a2-c4$ 等。比如, 机器学习类别所包含五篇文档的平均相关性仅为0.069, 基因编辑类别所包含四篇文档的平均相关性为0.237。
- 在基于重建单词-文档矩阵 $A_2$ 所构建的文档-文档相关系数矩阵中, 由于经过隐性语义分析, 可以看到相关性矩阵中数值已可以相当反映不同文档所蕴含语义关系。机器学习类别所包含五篇文档的平均相关性上升到0.989, 基因编辑类别所包含四篇文档的平均相关性上升到1.00。机器类别所包含文档与基因编辑类别所包含文档之间的平均相关性为-0.927。



# 基于矩阵分解的潜在语义分析

单词-文档矩阵(term-document): 构造与分解  $A = UDV^T$

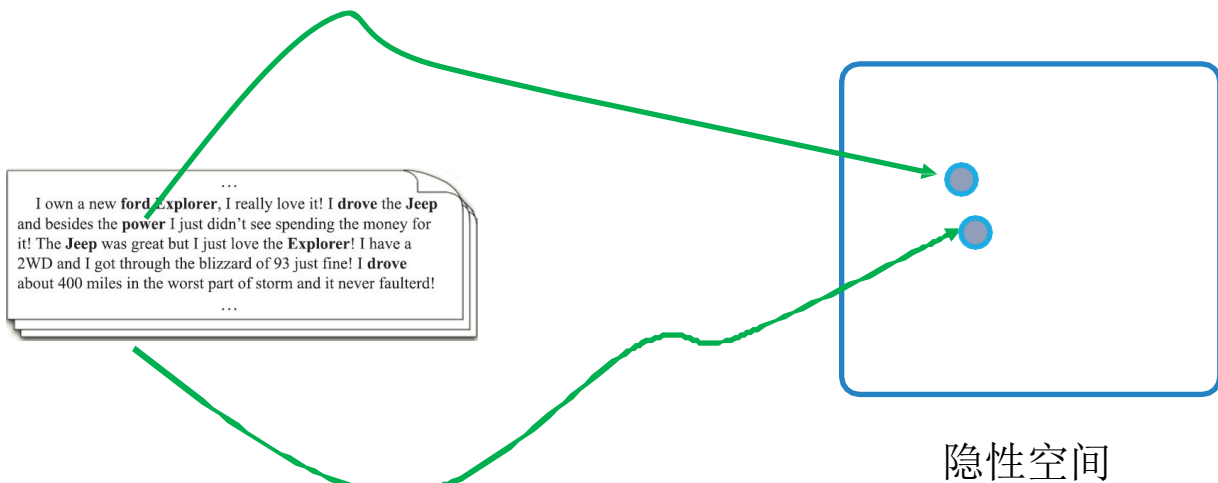
- 单词个数为 $M$ 、文档个数为 $N$
- 将每个单词映射到维度为 $R$ 的隐性空间、将每个文档映射到维度为 $R$ 的隐性空间：统一空间
- 隐性空间可视为“主题空间 (topic)”



# 基于矩阵分解的潜在语义分析

单词-文档矩阵(term-document): 构造与分解  $A = UDV^T$

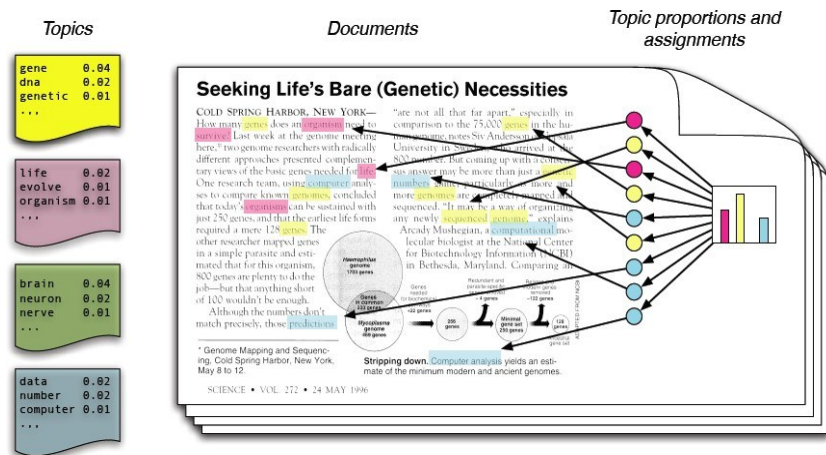
- 单词个数为 $M$ 、文档个数为 $N$
- 将每个单词映射到维度为 $R$ 的隐性空间、将每个文档映射到维度为 $R$ 的隐性空间：统一空间
- 隐性空间可视为 “主题空间 (topic) ”



## 基于矩阵分解的潜在语义分析

## 单词-文档矩阵(term-document): 构造与分解 $A = UDV^T$

- 单词个数为 $M$ 、文档个数为 $N$
- 将每个单词映射到维度为 $R$ 的隐性空间、将每个文档映射到维度为 $R$ 的隐性空间：统一空间
- 隐性空间可视为 “主题空间 (topic) ”



- 文章包含了基因、生命、神经和生物信息计算四个主题。
- 每个单词以不同的概率属于某一主题

## Latent Dirichlet Allocation(隐狄利克雷分配模型)



- 01** 基于逻辑斯蒂回归模型的分类
- 02** 基于矩阵分解的潜在语义分析
- 03** 线性判别分析及分类

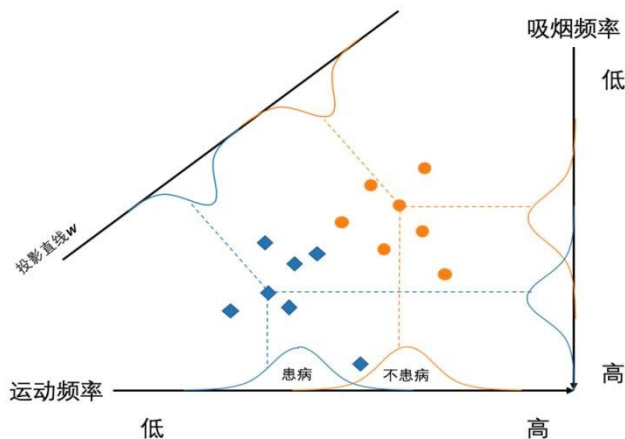
# 线性判别分析及分类

---

线性判别分析(linear discriminant analysis, LDA)是一种基于监督学习的降维方法, 也称为Fisher线性判别分析 (Fisher's Discriminant analysis, FDA)。

对于一组具有标签信息的高维数据样本, LDA利用其类别信息, 将其线性投影到一个低维空间上, 在低维空间中**同一类别样本尽可能靠近, 不同类别样本尽可能彼此远离**。

# 线性判别分析及分类



君子和而不同，小人同而不和  
虽然同一类别中每个数据各有特色，但是它们均具有相同内在模式，因而同一类别数据可被投影映射到一起。

- 左图给出了用矩形和圆圈来表示的患有某一疾病或者不患有某一疾病的两类人群。这两类人群分别用吸烟频率高低和运动频率高低来描述。
- 为了对这两类人群进行区分，需要将其投影到一个低维空间中。从图中可见，将其向 $x$ 轴方向和 $y$ 轴方向投影后，总会存在重叠部分（即若干人群在投影后的空间中不可区分）。
- 将数据向直线 $w$ 所位于方向投影，则可见两类数据已经被完全区分开来。
- 在所得投影空间中，同一类别人群数据聚集在一起、不同类别人群数据具有较大间隔，体现了“**类内方差小、类间间隔大**”的原则。

# 线性判别分析及分类

---

假设样本集为  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), \dots, (\mathbf{x}_N, y_N)\}$ , 样本  $\mathbf{x}_i \in R^n$  的类别标签为  $y_i$ 。其中,  $y_i$  的取值范围是  $\{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K\}$ , 即一共有  $K$  类样本。

定义  $N_i$  为第  $i$  类样本的个数、 $X_i$  为第  $i$  类样本的集合、 $\mathbf{m}$  为所有样本的均值向量、 $\mathbf{m}_i$  为第  $i$  类样本的均值向量。 $\Sigma_i$  为第  $i$  类样本的协方差矩阵, 其定义为:

$$\Sigma_i = \sum_{\mathbf{x} \in X_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$$

# 线性判别分析及分类

先来看 $K = 2$ 的情况，即二分类问题。在二分类问题中，训练样本归属于 $\mathcal{C}_1$ 或 $\mathcal{C}_2$ 两个类别，并通过如下的线性函数投影到一维空间上：

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} \quad (\mathbf{w} \in R^n)$$

投影之后类别 $\mathcal{C}_1$ 的协方差矩阵 $s_1$ 为：

$$s_1 = \sum_{\mathbf{x} \in \mathcal{C}_1} (\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mathbf{m}_1)^2 = \mathbf{w}^T \left[ \sum_{\mathbf{x} \in \mathcal{C}_1} (\mathbf{x} - \mathbf{m}_1) (\mathbf{x} - \mathbf{m}_1)^T \right] \mathbf{w}$$

同理可得到投影之后类别 $\mathcal{C}_2$ 的协方差矩阵 $s_2$ 。



# 线性判别分析及分类

协方差矩阵 $s_1$ 和 $s_2$ 可用来衡量同一类别数据样本之间“分散程度”。为了使得归属于同一类别的样本数据在投影后的空间中尽可能靠近，需要最小化 $s_1+s_2$ 取值。

minimize ( $s_1+s_2$ )

在投影之后的空间中，归属于两个类别的数据样本中心可如下分别计算：

$$m_1 = \mathbf{w}^T \mathbf{m}_1, \quad m_2 = \mathbf{w}^T \mathbf{m}_2$$

可通过 $\|m_2 - m_1\|_2^2$  来衡量不同类别之间的距离。为了使得归属于不同类别的样本数据在投影后空间中尽可能彼此远离，需要最大化 $\|m_2 - m_1\|_2^2$  取值。

maximize  $\|m_2 - m_1\|_2^2$



$$J(\mathbf{w}) = \frac{\|m_2 - m_1\|_2^2}{s_1 + s_2}$$

**投影方向优化目标**

# 线性判别分析及分类

$$J(\mathbf{w}) = \frac{\|m_2 - m_1\|_2^2}{s_1 + s_2}$$

可以把上述式子改写成与 $\mathbf{w}$ 相关的式子：

$$J(\mathbf{w}) = \frac{\|\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1)\|_2^2}{\mathbf{w}^T \Sigma_1 \mathbf{w} + \mathbf{w}^T \Sigma_2 \mathbf{w}} = \frac{\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w}}{\mathbf{w}^T(\Sigma_1 + \Sigma_2) \mathbf{w}} = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

其中， $\mathbf{S}_b$ 称为**类间散度矩阵**(between-class scatter matrix)，其定义如下：

$$\mathbf{S}_b = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$$

$\mathbf{S}_w$ 则称为**类内散度矩阵**(within-class scatter matrix)，其定义如下：

$$\mathbf{S}_w = \Sigma_1 + \Sigma_2$$

# 线性判别分析及分类

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

由于 $J(\mathbf{w})$ 的分子和分母都是关于 $\mathbf{w}$ 的二次项式，因此最后的解只与 $\mathbf{w}$ 的方向有关，与 $\mathbf{w}$ 的长度无关

将分母 $\mathbf{w}^T \mathbf{S}_w \mathbf{w} = 1$ 作为约束条件，将上述优化问题转变为拉格朗日函数：

$$L(\mathbf{w}) = \mathbf{w}^T \mathbf{S}_b \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{S}_w \mathbf{w} - 1)$$

带约束条件（即 $\mathbf{w}^T \mathbf{S}_w \mathbf{w} - 1 = 0$ ）的函数极大值（即 $\mathbf{w}^T \mathbf{S}_b \mathbf{w}$ 取值最大）优化问题所对应拉格朗日函数

对 $\mathbf{w}$ 求偏导并使其求导结果为零，可得：

$$\mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{w} = \lambda \mathbf{w}$$

注：拉格朗日乘子法就是求在某个/某些约束条件下函数极值方法，其主要思想是将约束条件函数与原函数联立，从而求出使原函数取得极值时各个变量的解。

# 线性判别分析及分类

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$



拉格朗日乘子法

$$\mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{w} = \lambda \mathbf{w}$$

Fisher线性判别  
(Fisher linear discrimination)

$\mathbf{w}$ 和 $\lambda$ 是 $\mathbf{S}_w^{-1} \mathbf{S}_b$ 的特征向量和特征根

因为  $\mathbf{S}_b = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$  , 那么  $\mathbf{S}_b \mathbf{w} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} = (\mathbf{m}_2 - \mathbf{m}_1) \times \lambda_w$ , 将其代入 Fisher线性判别, 可得:

$$\mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{w} = \mathbf{S}_w^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \times \lambda_w = \lambda \mathbf{w}$$

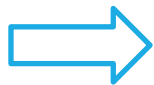
由于对 $\mathbf{w}$ 的扩大和缩小操作不影响结果, 因此可约去上式中的未知数 $\lambda$ 和 $\lambda_w$ , 得到:

$$\mathbf{w} = \mathbf{S}_w^{-1} (\mathbf{m}_2 - \mathbf{m}_1)$$

为了获得“类内汇聚、类间间隔”的最佳投影结果, 只需要分别求出待投影数据的均值和方差, 就可以设计得到最佳投影方向 $\mathbf{w}$ 。LDA模型可从两类问题被拓展到多类问题

# 线性判别分析及分类

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$



$$\mathbf{w}^T \mathbf{x}$$



$$p(\text{class} | \mathbf{w}^T \mathbf{x})$$

寻找“类内汇聚、类间间隔”的  
最佳投影结果

降维：从高维到低维的映射

分类：将降维后结果判断为  
某个类别

主成分分析（PCA）是一种无监督学习的降维方法（无需样本类别标签），线性判别分析（LDA）是一种监督学习的降维方法（需要样本类别标签）。PCA和LDA均是优化寻找一定特征向量 $\mathbf{w}$ 来实现降维，其中PCA寻找投影后数据之间**方差最大**的投影方向、LDA寻找“**类内方差小、类间距离大**”投影方向。

PCA对高维数据降维后的维数是与原始数据特征维度相关（与数据类别标签无关）。假设原始数据维度为  $d$ ，那么PCA所得数据的降维维度可以为小于 $d$ 的任意维度。LDA降维后所得到维度是与数据样本的类别个数 $K$ 有关（与数据本身维度无关）。假设原始数据一共有 $K$ 个类别，那么LDA所得数据的降维维度小于或等于 $K - 1$ 。