模式识别
（Pattern Recognition）

广东工业大学集成电路学院　邢延

第七讲 特征提取

（07 Feature Extraction）

提醒

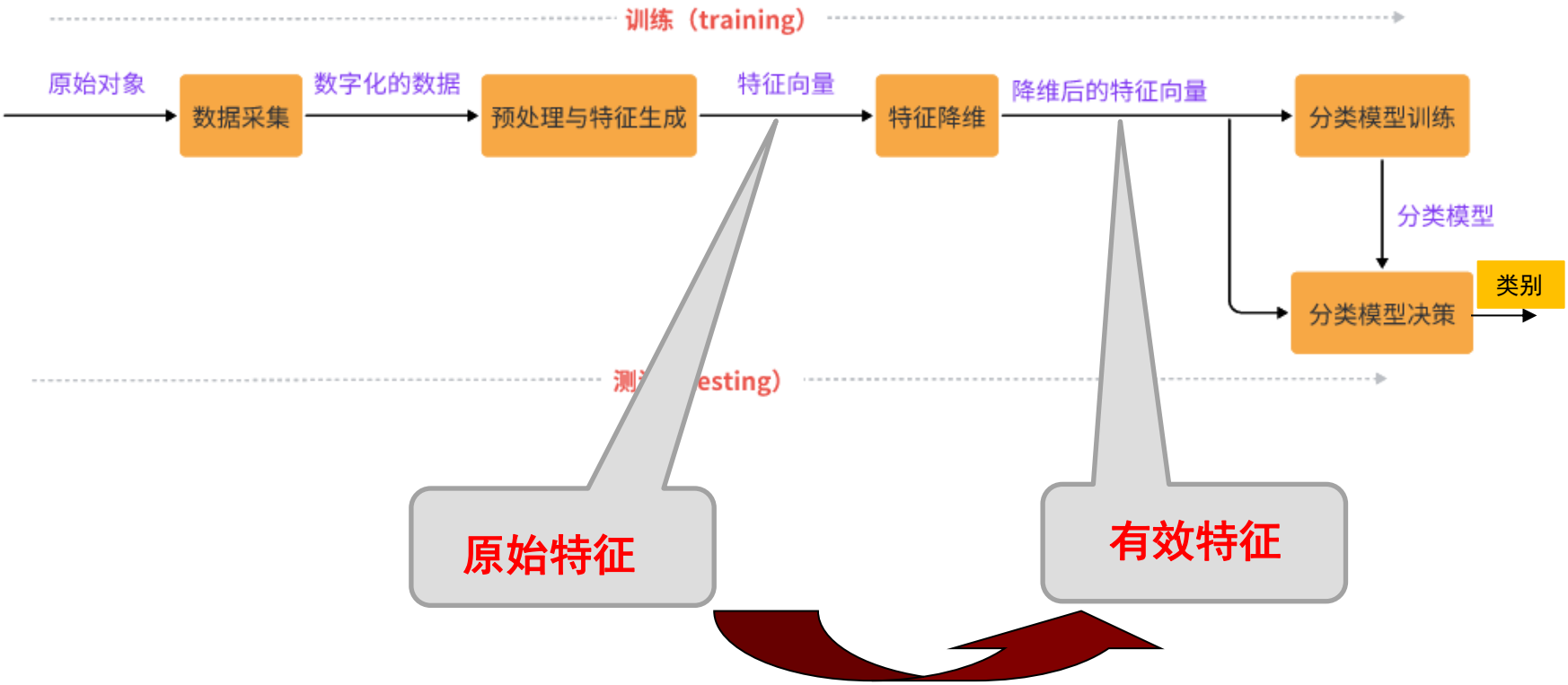
- 课程任务
  - 任务2：自学Python编程、复习数学基础
    - ◆安装Anaconda集成开发工具包
    - ◆安装Pycharm开发工具
    - ◆自学华为云免费课程
      - Python入门篇 - AI基础课程-数学基础知识
      - Python进阶篇
    - ◆学习时间为4-9周，并于10月28日前将以下文档打包（命名：学号+姓名）交给学委，由学委总打包（命名：班级+PR慕课学习）发给课程助教
      - 课程完成情况截图（见后页例子）
      - 学习笔记、习题、练习、思考题、编程练习等

内容提要

- 数据降维的必要性
- 数据降维的方法
  - 特征提取
    - 图像的特征提取
    - 文本的特征提取
  - 特征选择

数据降维的必要性

- 模式识别系统



数据降纬(Data Reduction)

特征工程(Feature Engineering)

数据降维的必要性

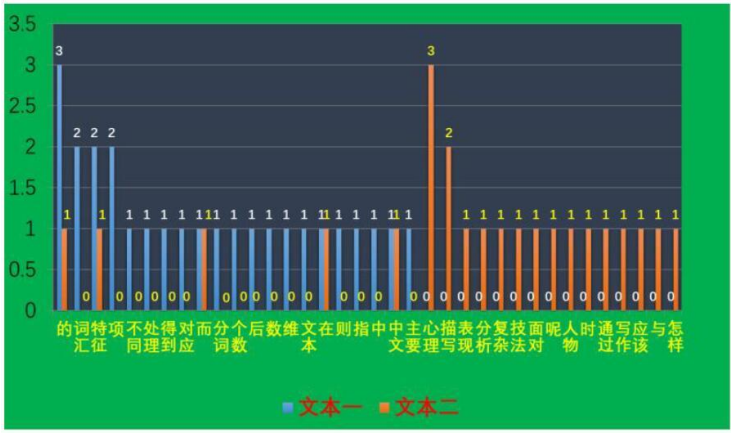
- **数据降维(Data Reduction)**
  - **人脸识别为例：**
    - **原始特征：在人脸图像库中，每幅图像的分辨率为128*128，即高达16384维。**
    - **如果是256个灰度级的图像，即1个字节可以存储1个像素，则每幅图像大小为16384byte，即16KB。**
    - **通过PCA算法提取有效特征，则每幅图像的有效特征为99维。**
    - **如果用双精度浮点数代表每一维数值，而每个数值需要占用8byte，则每幅图像的有效特征大小为792byte，即小于1KB。**

数据降维的必要性

- 数据降维(Data Reduction)
  - 以文本分类为例：
    - ◆词频
      - 即不同的词在不同的文档中出现的频率.
    - ◆作为文本分类的特征向量
      - 不同的词的词频组合起来
      - 维度非常高
      - 把可能出现的每个词都作为一个特征维度，来统计它们的词频，而这样得到的特征空间，将囊括一本厚厚词典里的所有词，甚至还会更多（要考虑各种专有名词、姓名等）

“特征项”在中文文本中主要指分词处理后得到的词汇，而特征项的维数则对应不同词汇的个数。

而面对复杂的心理描写与心理分析技法，在中文写作时应该怎样通过心理描写表现人物特征呢



数据降维的必要性

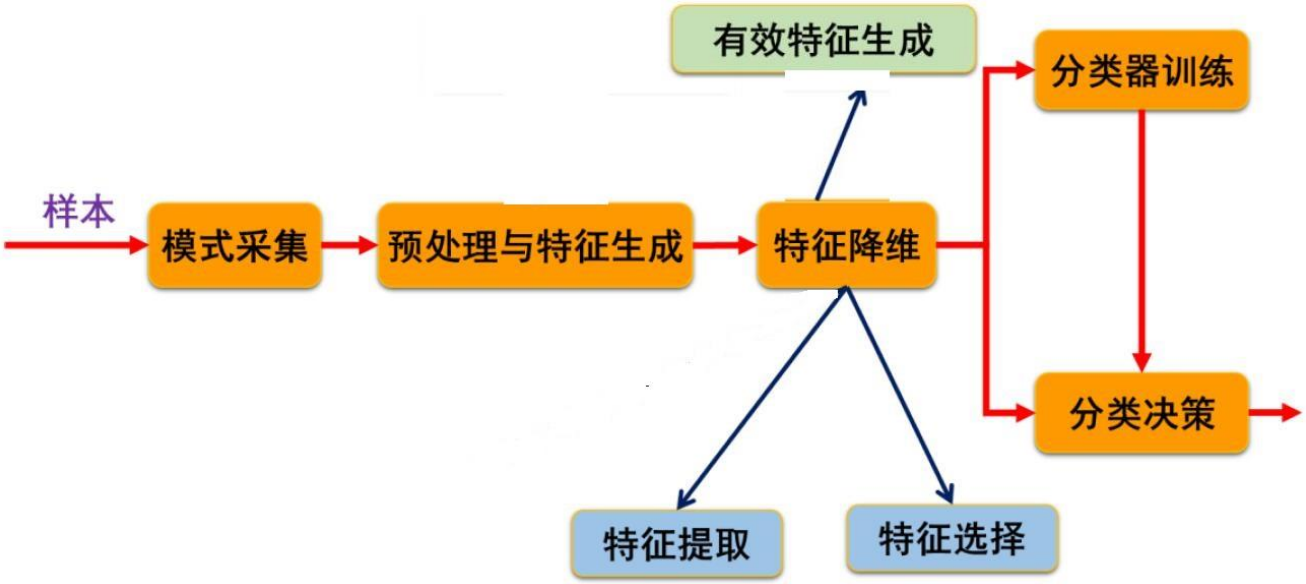
- 数据降维，或者特征降维
  - 可以大大地降低一个模式识别任务的计算复杂度，
  - 有可能提升分类决策的正确率
  - 使用更少的代价，设计出一个更加优秀的模式识别系统

数据降维的必要性

- **原始特征 VS 有效特征**
  - **原始特征**
    - **通过直接测量得到的特征称为原始特征。**
      - **例如：人体的各种生理指标（描述其健康状况）**
      - **数字图像中的每点灰度值（以描述图像内容）**
      - **对商品评价的文本等**
  - **有效特征**
    - **采用数据降维方法获得的有代表性、分类性能好的特征**

数据降维的必要性

- 模式识别系统
  - 特征降维是一个必不可少的重要环节
  - 特征降维的主要目标
    - 获得对分类最有效的特征
    - 同时尽最大可能减少特征维数



特征提取

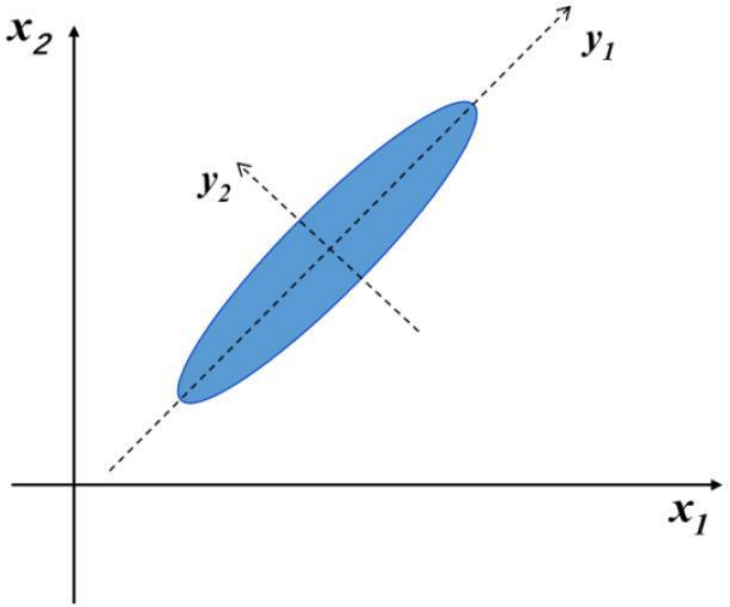
- **定义**
  - **通过映射(或变换)的方法获取最有效的特征，实现特征空间的维数从高维到低维的变换；**
  - **经过映射后的特征是原始特征的某种组合，最常用的是线性组合。**

图像的特征提取

- **主成分分析法(PCA, principle component analysis)**
  - **也称主分量分析;**
  - **对样本集整体进行的降维操作;**
  - **来源于统计学;**
  - **在统计样本中找到影响结果的最关键的那些变量;**
  - **属于无监督的方法。**

图像的特征提取

- **PCA的核心思想**
  - 样本集在各个不同的方向上进行投影，其方差是不同的，方差越大的方向，包含的信息量也就越大，就越是整个样本集分布特性的“主成分”。



$$Y = W^T X$$

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{D1} & \cdots & x_{Dm} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{d1} & \cdots & y_{dm} \end{bmatrix}$$

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1d} \\ \vdots & \ddots & \vdots \\ w_{D1} & \cdots & w_{Dd} \end{bmatrix}$$

要求：Y的协方差矩阵为对角阵

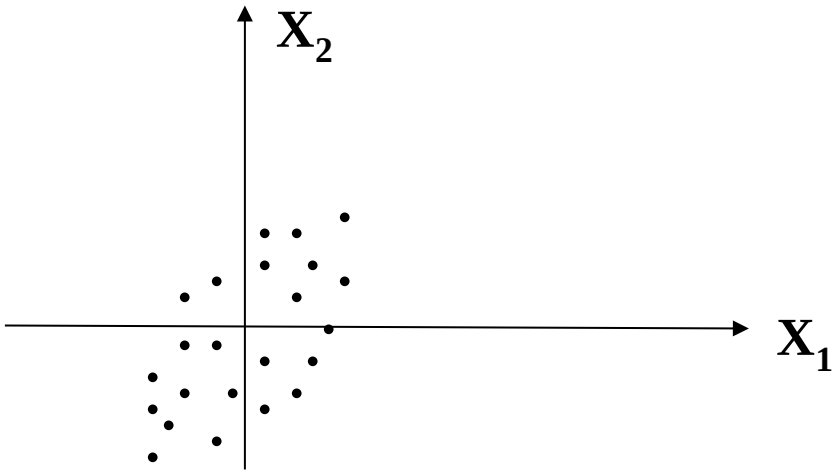
图像的特征提取

- PCA的核心思想
  - 是一个线性变换
  - 把数据变换到一个新的坐标系统中
  - 新坐标系是正交坐标系
  - 新坐标系中的特征是互相独立的
  - 新坐标系的维度不高于原始坐标系维度

图像的特征提取

- **PCA的核心思想**
  - **例子**

有n个样本，每个样本有两个变量值X_1和X_2，这n个样本的散点图如下：

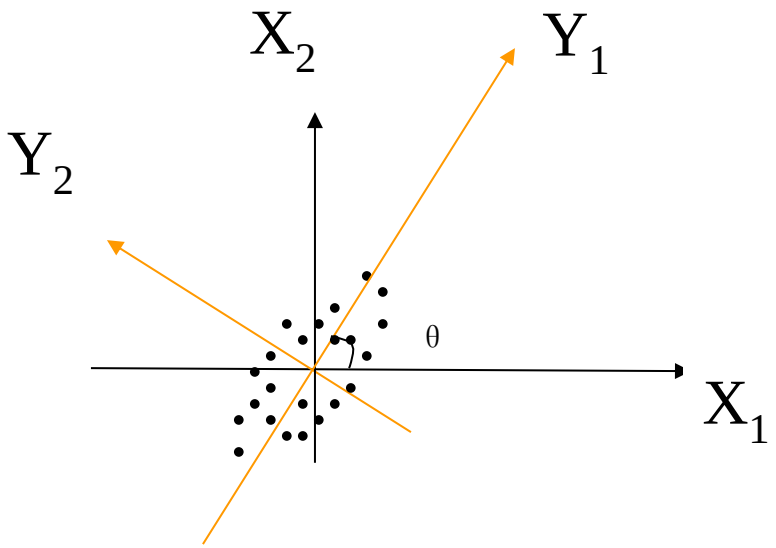


这n个样本点沿着X_1轴方向或X_2轴方向都具有较大的离散性，

其离散的程度可以分别用观测变量X_1的方差和X_1的方差定量地表示.

图像的特征提取

- **PCA的核心思想**
  - **例子**



将X_1轴和X_2轴同时按逆时针方向旋转θ角度，得到新坐标轴Y_1 和Y_2

Y_1 和Y_2是两个新变量，且Y_1 上的方差最大，

数据在Y_1 上的投影为第一主成分，

数据在Y_2上的投影为第二主成分.

图像的特征提取

- **PCA的求解方法**

求解主成分分析问题，可以先将原始数据集做平移变换，将坐标原点移到样本集均值点，使得协方差矩阵便于计算。可得

$$cov\ X' = \frac{1}{m} X' X'^{T}$$

则映射后的 Y 空间中

$$cov\ Y' = \frac{1}{m} Y' Y'^{T} = W(cov\ X')W^{T}$$

即要求使得 Cov Y'对角化的变换阵 W，并且映射后各维度按方差从大到小排列。主成分分析的解为：将 X '的协方差矩阵的特征根从大到小排列，对应的前 d 个特征向量构成的变换阵 W，即可以按主成分分析的要求得到降维的样本集变换结果。

图像的特征提取

- **PCA的求解方法**
  - **例子1**

10位同学的身高、胸围和体重数据如下，要求对数据进行主成分分析

身高x_1(cm)	胸围x_2(cm)	体重x_3(kg)
149.5	69.5	38.5
162.5	77.0	55.5
162.7	78.5	50.8
162.2	87.5	65.5
156.5	74.5	49.0
156.1	74.5	45.5
172.0	76.5	51.0
173.2	81.5	59.5
159.5	74.5	43.5
157.7	79.0	53.5

图像的特征提取

- **PCA的求解方法**
  - **例子1**
    - **解：1）求样本均值和样本协方差矩阵**

$$\begin{pmatrix} \overline{x_1} \\ \overline{x_2} \\ \overline{x_3} \end{pmatrix} = \begin{pmatrix} 161.2 \\ 77.3 \\ 51.2 \end{pmatrix}$$

$$S = \begin{pmatrix} 46.67 & & \\ 17.12 & 21.11 & \\ 30.00 & 32.58 & 55.53 \end{pmatrix}$$

图像的特征提取

- **PCA的求解方法**
  - **例子1**
    - ◆ **解：2）求解协方差矩阵的特征方程** $|S-\lambda I|=0$

$$\begin{vmatrix} 46.67-\lambda & 17.12 & 30.00 \\ 17.12 & 21.11-\lambda & 32.58 \\ 30.00 & 32.58 & 55.53-\lambda \end{vmatrix}=0$$

得 3个特征值和对应的特征向量：

$$\lambda_1 = 98.15 \quad (a_{11}, a_{21}, a_{31}) = (0.56, 0.42, 0.71)$$
$$\lambda_2 = 23.60 \quad (a_{12}, a_{22}, a_{32}) = (0.81, -0.33, -0.48)$$
$$\lambda_3 = 1.56 \quad (a_{13}, a_{23}, a_{33}) = (0.03, 0.85, -0.53)$$

图像的特征提取

- **PCA的求解方法**
  - **例子1**
    - **◆解：3）写出三个主成分的表达式**

$$F_1 = 0.56(x_1 - 161.2) + 0.42(x_2 - 77.3) + 0.71(x_3 - 51.2)$$

$$F_2 = 0.81(x_1 - 161.2) - 0.33(x_2 - 77.3) - 0.48(x_3 - 51.2)$$

$$F_3 = 0.03(x_1 - 161.2) + 0.85(x_2 - 77.3) - 0.53(x_3 - 51.2)$$

- **◆主成分的含义**
  - **F_1表示学生身材大小**
  - **F_2反映学生的体形特征**
  - **F_3反映学生的体重特征**

- **PCA的求解方法**
  - **例子1**
    - **三个主成分的方差贡献率分别为：**

$$\frac{\lambda_1}{\sum_{i=1}^{3}\lambda_i}=\frac{98.15}{98.15+23.60+1.56}=\frac{98.15}{123.31}=79.6\%$$

$$\frac{\lambda_2}{\sum_{i=1}^{3}\lambda_i}=\frac{23.60}{123.31}=19.1\% \qquad \frac{\lambda_3}{\sum_{i=1}^{3}\lambda_i}=\frac{1.56}{123.31}=1.3\%$$

    - **前两个主成分的累积方差贡献率为：**

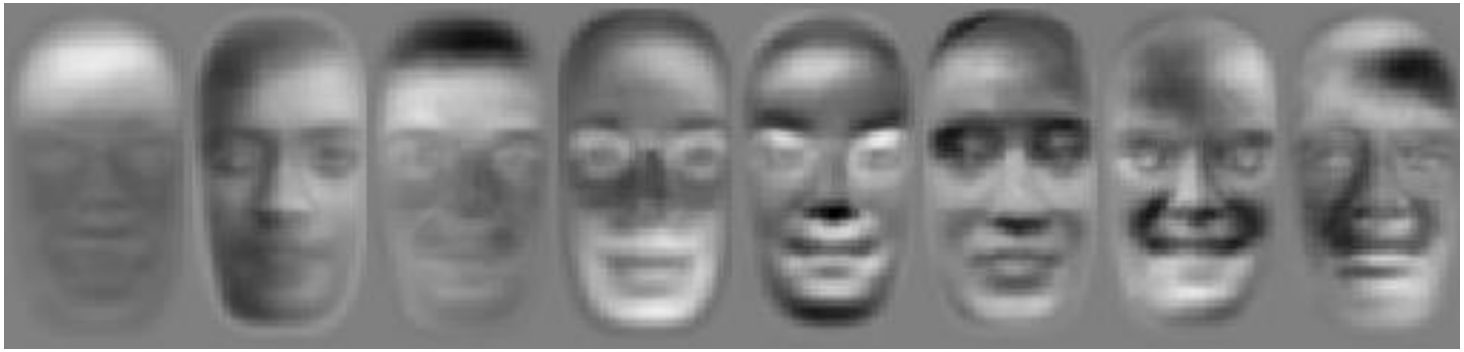
$$\frac{\lambda_1+\lambda_2}{\sum_{i=1}^{3}\lambda_i}=\frac{121.75}{123.31}=98.7\%$$

图像的特征提取

- **PCA的注意事项**
  - **原始变量的标准化**
    - **主成分是根据变量的离散度也即方差的大小来确定主成分的，这样当不同指标的量纲不同时，不同指标的方差大小差别很大，主成分会受到影响。**
    - **例如：X1表年收入，从万元到百万元变化，X2表净收入与总资产之比，从0.01到0.60变化,那么X1的方差的绝对量将远远大于X2的方差,主成分会过于照顾方差大的变量。**
    - **为使主成分能均等地对待每一个原变量,应将原变量作标准化处理。**
  - **对于高维度的协方差矩阵，实际采用奇异值分解的方法求取特征值**

图像的特征提取

- **PCA的求解方法**
  - **例子2**
    - **特征脸（eigenface）方法**
      - **是人脸识别的基准技术，并已成为工业标准**
      - **该方法基于主成分分析（PCA），求解过程相同**
      - **如果将特征向量恢复成图像，这些图像很像人脸，因此称为“特征脸”**



文本的特征提取

- **文本的特征提取**
  - **难点**
    - **人类语言（自然语言）是用来传递信息、表达意图的特有系统**
    - **不是由任何物理表现产生的**
    - **不同于视觉以及其他任何机器学习任务**

计算机

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} = \begin{bmatrix} a_1b_1 & a_1b_2 & a_1b_3 \\ a_2b_1 & a_2b_2 & a_2b_3 \\ a_3b_1 & a_3b_2 & a_3b_3 \end{bmatrix}$$

VS

人类

老外来访被请吃饭。落座后，一中国人说："我先去方便一下。"老外不解，被告知"方便"是"上厕所"之意。席间主宾大悦。道别时，另一中国人对老外发出邀请："我想在你方便的时候也请你吃饭。"老外愣了，那人接着说："如果你最近不方便的话，咱找个你我都方便的时候一起吃。

词、句子、段落、文档

上下文

文本的特征提取

- 词向量(Word Vectors)
  - 自然语言最基本的单位是：词
  - 用数值向量表示词 --- 词向量
  - 词向量的表示方法
    - One-hot representation（独热表示）
    - Distributional representation（分布式表示）

文本的特征提取

- **One-hot表示**
  - **例子**

语料库1: {男、女}
语料库2：{Jauuary, February, …, December}

One-hot1: 男=[1,0], 女=[0,1]

One-hot2：Jauuary= [1,0,0,0,0,0,0,0,0,0,0,0]
February=[0,1,0,0,0,0,0,0,0,0,0,0]
……
December=[0,0,0,0,0,0,0,0,0,0,0,1]

英语的词汇量非常庞大（总计990,000个），英文母语者常用的2-3万，受过教育的3-5万。

现代汉语的单字大概有30000个左右，又大约每个字有10个左右的词汇，所以汉语的词汇总量大约是30万个，常用的有10万个左右。

文本的特征提取

- One-hot表示
  - 优点
    - 简单、直观
    - 解决了分类器不好处理离散数据的问题
    - 在一定程度上起到了扩充特征的作用
  - 缺点
    - 是词袋模型
    - 不考虑词与词之间的顺序（文本中词的顺序信息非常重要）
    - 假设词与词相互独立（在大多数情况下，词与词是相互影响的）
    - 得到的特征是离散稀疏的（语料库较大的情况下可能产生维数灾难）

文本的特征提取

- 分布式表示
  - 词表示为如[0.792,-1.177,-0.107,0.109,...]这种稠密的向量形式
  - 常见的维度为50或者100
  - 解决词汇鸿沟问题
    - 通过计算向量之间的距离来体现词与词的相似性
  - 需要通过“语言模型(language model)”进行训练得到

文本的特征提取

- **分布式表示**
  - **语言模型的概念**
    - **判断一句话是否符合自然语言法则**
    - **通俗地说，就是判断一句话是不是人话**

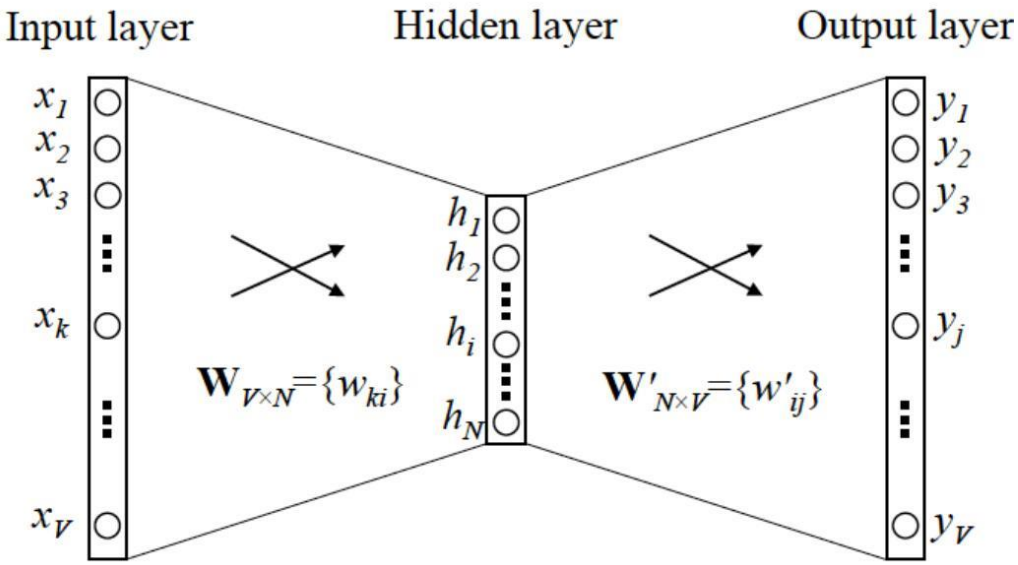
假设：x是 一个句子里的一个词语，y 是这个词语的上下文词语，
有：f(x) -> y
则：f 就是语言模型

文本的特征提取

- 分布式表示
  - 语言模型的种类
    - 矩阵
    - 聚类
    - 神经网络
  - 基于神经网络的分布式表示
    - 又名：词嵌入（word embedding）
    - 算法
      - Word2vector
      - SEENA
      - FastText
      - ……

文本的特征提取

- **Word2vector**
  - **采用简化的浅层神经网络，用误差反馈方法训练；**
  - **输入是One-Hot Vector；**
  - **隐含层是线性单元，没有激活函数；**
  - **输出层维度跟输入层的维度一样，用的是Softmax回归；**
  - **模型训练是为了得到隐含层的权重矩阵；**
  - **隐含层维度远小于其它层。**

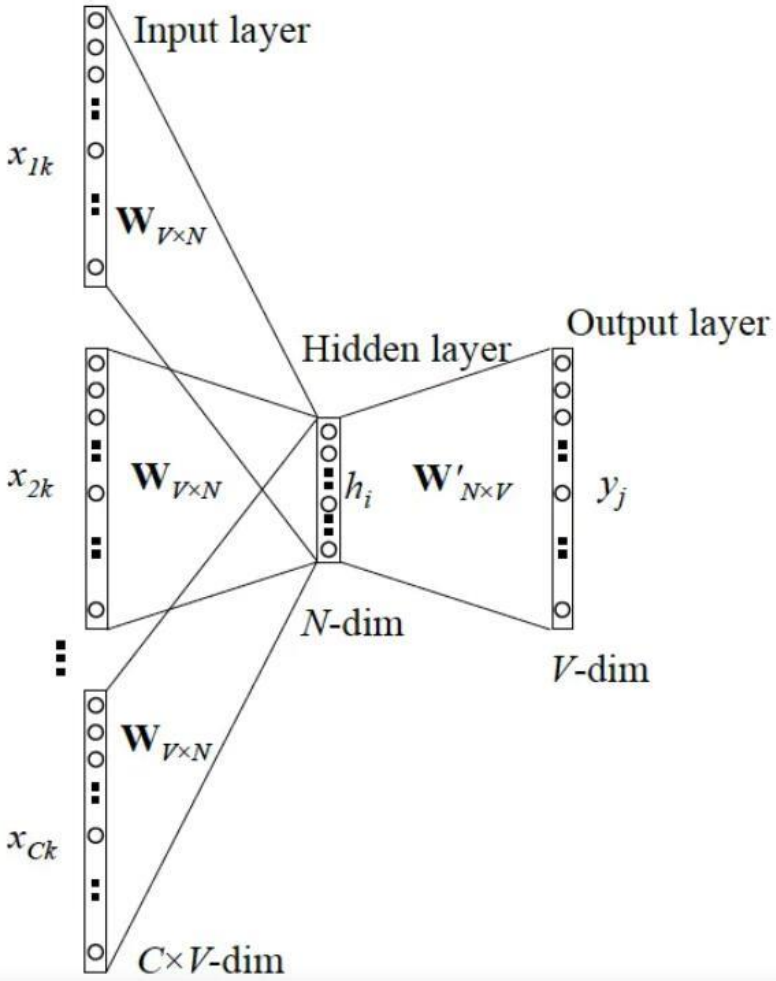


文本的特征提取

- **Word2vector**
  - **神经网络模型**
    - **CBOW模型(continuous Bag of Word model)**
      - **拿一个词语的上下文作为输入，来预测这个词语本身**
    - **Skip-gram模型(Continuous Skip-Gram model)**
      - **用一个词语作为输入，来预测它周围的上下文**

文本的特征提取

- **CBOW模型**



模型训练：

1. 正向传播

1.1 输入层：上下文单词的onehot. (设单词向量空间维度为V，上下文单词个数为C);

1.2 所有onehot分别乘以共享的输入权重矩阵W(V*N矩阵，N预先设定);

1.3 所得的向量相加求平均作为隐层向量(1*N)，乘以输出权重矩阵W’ (N*V);

1.4 得到输出向量 (1*V),经激活函数处理得到V维概率分布，且概率最大的那一维对应的词汇为预测输出。

2. 误差反向传播

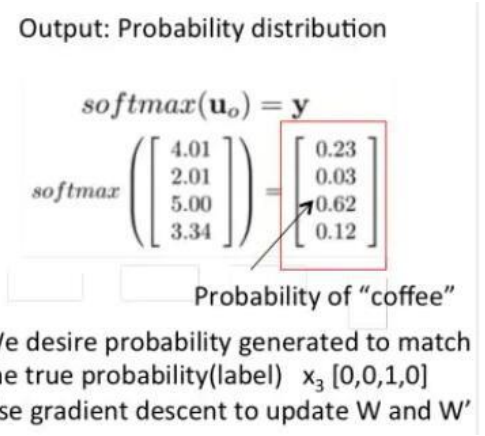
计算预测输出的onehot与真实单词之间的误差，根据误差采用梯度下降法更新权重矩阵W和W’。

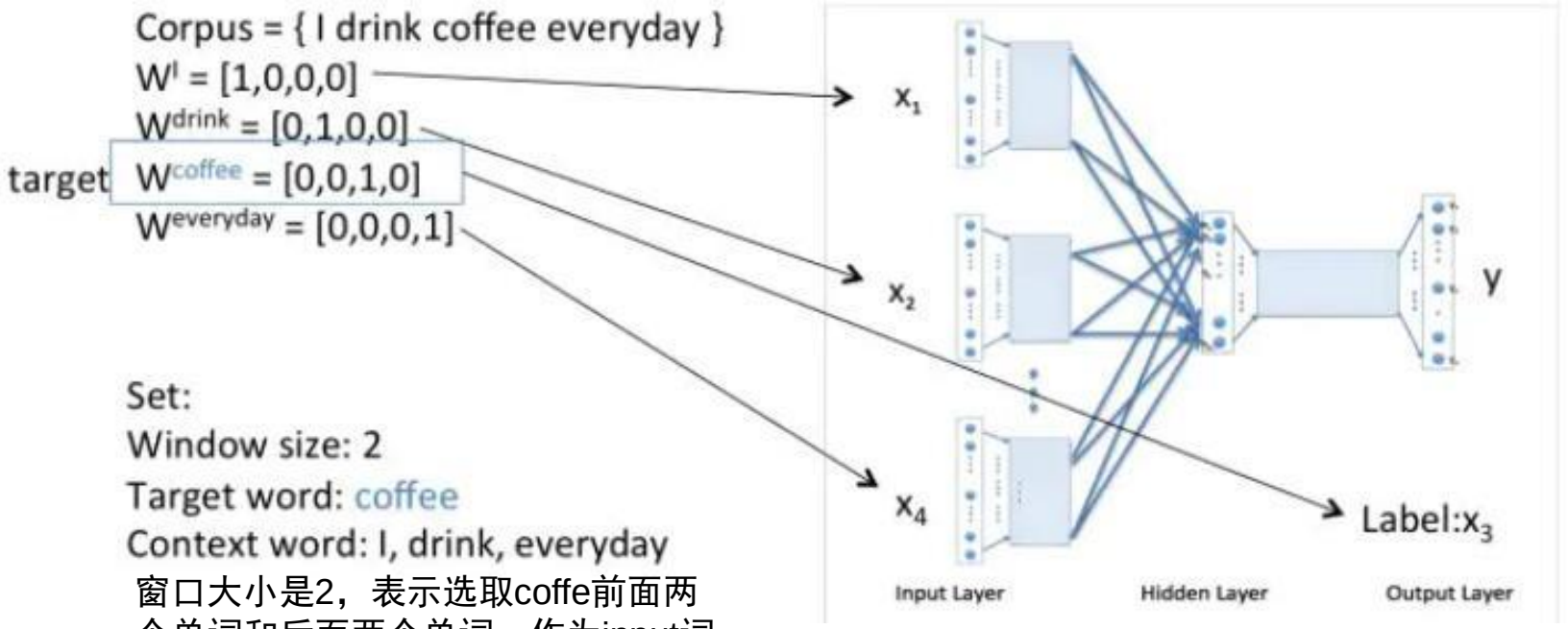
模型训练完成后，**输入层的每个单词与矩阵W相乘得到的向量的就是词向量**。

文本的特征提取

- **CBOW模型的例子**

训练出来的权向量矩阵W；
一个单词的one-hot乘以W得到自己的词向量。

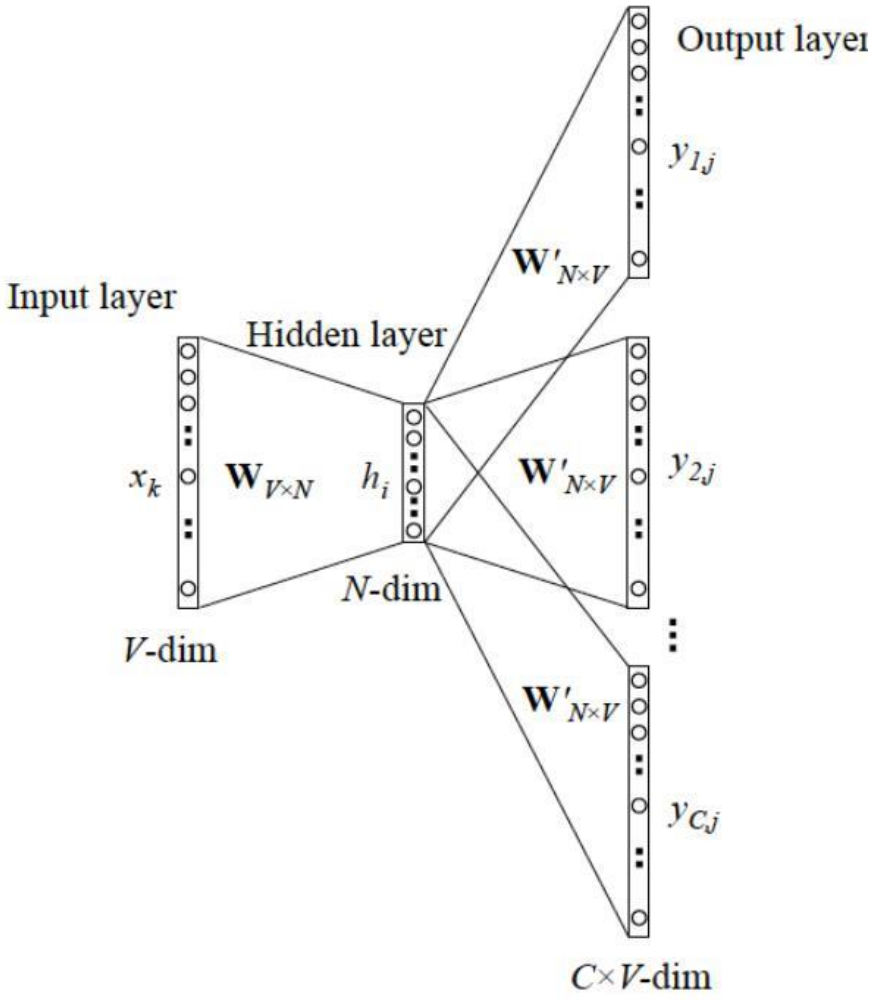




窗口大小是2，表示选取coffe前面两个单词和后面两个单词，作为input词

文本的特征提取

- **Skip-Gram模型**



模型训练：

1. 正向传播

1.1 输入层：关键单词的onehot. (设单词向量空间维度为V，上下文单词个数为C);

1.2 该onehot乘以输入权重矩阵W(V*N矩阵，N预先设定);

1.3 所得的向量相加求平均作为隐层向量(1*N)，乘以输出权重矩阵W' (N*V);

1.4 得到输出矩阵 (C*V),经激活函数处理得到C*V维概率分布。这些概率代表着上下文中每个词有多大可能性跟输入的单词同时出现。

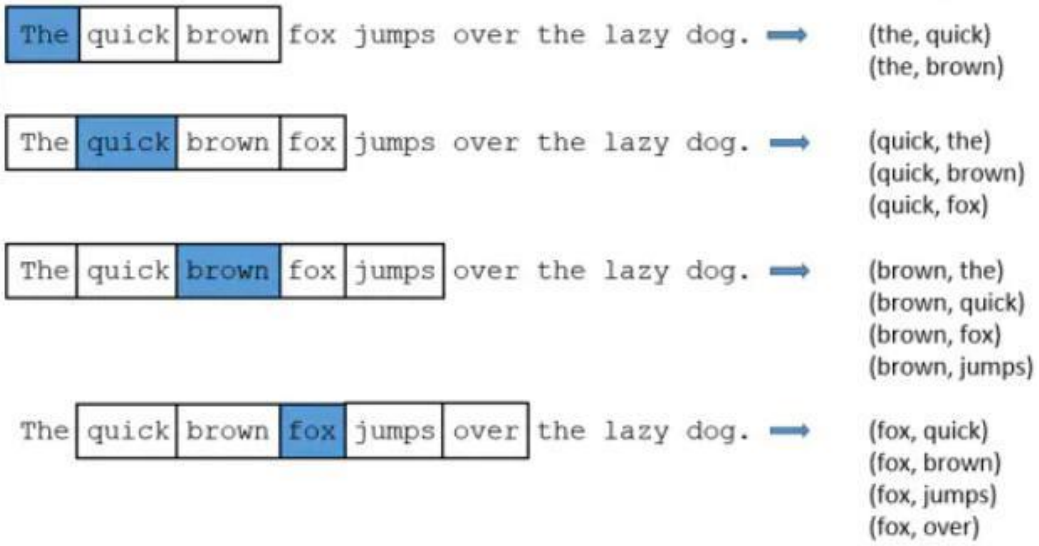
2. 误差反向传播

计算预测输出的onehot与真实上下文单词之间的误差，根据误差采用梯度下降法更新权重矩阵W和W'。

模型训练完成后，**输入层的单词与矩阵W相乘得到的向量的就是词向量**。

文本的特征提取

- **Skip-Gram模型的训练集**



选定句子：
The quick brown fox jumps over lazy dog

设定窗口大小（window_size=2）

蓝色代表input word

方框内代表位于窗口内的单词

Training Samples（输入， 输出）

模式识别算法编程实例演示

- **实例7：特征提取**
  - **参见Jupyter Notebook文档目录：lesson07**