机器学习基石篇

1.机器学习常见应用和适用模型（文本篇）

1.自然语言处理技术概述

- 研究如何让计算机能够像人一样理解人类世界中的自然语言，包括

  （1）自然语言的表示模型

  （2）文本处理技术的应用：分词、搜索、主题等

- 自然语言处理=文本处理+机器学习

2.中英文分词模型与常见工具包

- 常见的中文分词模型

  （1）基于词典的机械分词模型

  最大匹配法、全切分路径选择方法

  （2）基于统计模型的序列标注分词模型

  HMM(HiddenMarkoy Model）隐马尔科夫模基

  CRF(Conditionalrandom field，条件随机场）

- 常见的中文分词工具包

  Word分词器

  Ansj分词器

  Stanford分词器

  Fudannlp分词器

  Jcseg分词器

  smarten分词器

  IKAnalyzer分词器

  MMSeg4j分词器

  Jieba分词器

  Paoding分词器

3.查询扩展模型及实例

查询扩展：词相似性挖掘（交换机--路由器）

- 词向量表示模型

  文档-词频矩阵

  共生词矩阵

  word2Vec（三层神经网络构建词向量矩阵）

- 词向量相似性

  余弦距离

图模型

- 查询扩展实例：word2vec工具包

  github上下载开源工具包

  用maven建立工程进行包管理

4.主题模型及实例

- 主题模型：

  一种概率主题模型：隐含狄利克雷分布

  一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成

  主要用在推荐系统中

- 主题模型实例分析

  github上下载*LDAGibbsSampling*

  同样使用maven工程进行包管理

5.自动文本摘要模型及实例

- 自动文本摘要模型

  抽取型（主要方法）：直接从文章中抽取一定量的句子构成摘要

  摘取型：基于对文章的深刻理解及形式化表达的基础上，重新生成文章摘要

- 核心步骤

  （1）（抽取型）句子打分与文摘句选择

  （2）（摘取型）文摘句排序

  句子压缩、句子融合、句子生成（核心词、词性分析、句子规则建模）

- 自动文本摘要工具：HanLP工具包

2.机器学习常见应用和适用模型（图像篇）

1.计算机视觉技术概述

- 计算机视觉技术 = 图像数据 + 机器学习

  图像信息表示模型

  应用场景：分类、检索、分割、降噪

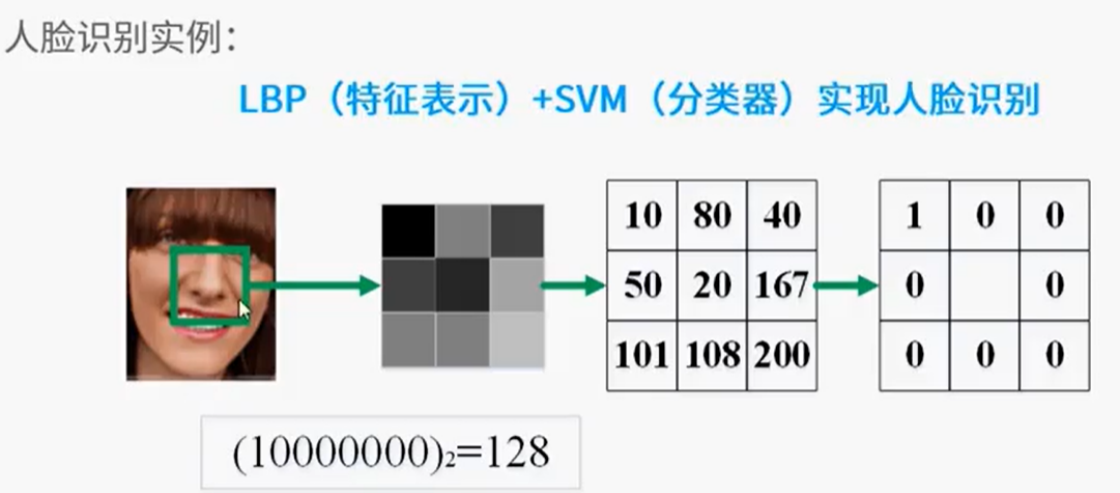
2.人脸识别模型及实例

- 人脸识别问题

  人脸特征表示模型（颜色、纹理、形状等）

  人脸识别分类器（线性判别、SVM、深度学习等）

- 人脸识别实例（MATLAB实现）



LBP 常用于灰度图像

3.图像检索模型及实例

- 图像检索（基于内容的图像检索技术，CBIR）

  （1）图像特征表示模型

  颜色、纹理、形状、哈希等

  （2）图像相似性度量模型

  欧氏距离、堪培拉距离、余弦距离、马氏距离等

- 图像检索实例（开源工具）：

  LIRE: Lucene Image Retrieval

  该工具使用gradle作为包管理工具。

4.行人检测模型及实例

- 行人检测模型

  （1）行人图像特征表示模型

  灰度、边缘、纹理、颜色、梯度直方图等

  （2）行人图像分类器

  神经网络、SVM、adaboost以及深度学习等

- 行人检测实例

  HOG（特征表示） + SVM（分类器）实例

5.图像分类模型及实例

- 图像分类模型

  （1）图像特征表示模型

  颜色、纹理、形状、哈希等

  （2）图像分类器

  最邻近、K临近、SVM、Deep Learning等

- 图像分类实例：

  CIFIR-10图像分类问题（KNN）

  MATLAB作为编程工具。

3.机器学习常见应用和适用模型（语音篇）

1.语音处理技术概述

- 识别和理解语音信号实现同相应文本或命令的相互转化技术，包括

（1）语音识别技术（解决设备只能通过按键操作）

（2）语音合成技术（解决只能看不能听）

- 语音识别技术=语音数据+机器学习

2.语音识别模型及应用

- 语音识别模型包含以下几个模块

  语音信号预处理和特征提取

  声学模型建立

  语音模型建立

$$W^* = arg\max_W P(W|X)$$
$$= arg\max_W \frac{P(X|W)P(W)}{P(X)}$$
$$= arg\max_W P(X|W)P(W)$$

W表示文字序列
X表示语音信号

- 语音信号预处理和特征提取包括

  VAD

  分帧

  梅尔频率倒谱系数（MFCC）

- 语音识别模型包括

  声学模型建立：P（X|W）

  - 字典
  - 隐马尔可夫（HMM）
  - viterbi算法

  语言模型建立：P（W）

  - n-gram

  设W是w1，w2，...，wn组成的，则P（W）可以拆成：

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \dots P(w_n|w_1 \dots w_{n-1})$$

- 语音识别开源工具

  CMU- Sphinx、HTK、 Julius、 RWTH ASR、 Kaldi、 simon、 iATROS- speech、SHOUT、Zanzibar OpenIVR、百度语音识别等等。

- 百度语音识别实例

  具体通过百度开发者账号调用百度语音API实现，编写的是Maven项目。

3.语音合成模型及应用

- 语音合成模型（TTS技术又称文语转换技术）

  文本分析模块

  韵律生成模块

  声学模块

- 文本分析模块

  文本规整、词的切分、语法分析和语义分析

  常用方法：规则、二元、三元文法，隐马尔可夫、神经网络等

- 韵律生成模块

  为合成语音规划出音段特征

  基于规则、基于神经网络、基于统计模型等方法

- 声学模块

  语音合成

  基于参数合成的方法，基于波形拼接的方法（PsoLA算法）等

- 语音合成开源工具：

  MARY

  SpeakRight Framework

  Festival

  FreeTTS

  eSpeak

  Flite

- 百度语音合成实例