# 选择题

1. 以下属于生成模型的是（C）

A. 支持向量机 B. 感知机 C. 朴素贝叶斯 D. 逻辑斯谛回归

1. 以下两种描述分别对应哪两种分类算法的评价标准？（A）

(a) 警察抓小偷，描述警察抓的人中有多少个是小偷的标准。

(b) 描述有多少比例的小偷给警察抓了的标准。

A. Precision, Recall B. Recall, Precision C. Precision, ROC D. Recall, ROC

1. 通过聚集多个分类器的预测来提高分类准确率的技术称为（A）

A. 组合(ensemble) B. 聚集(aggregate) C. 合并(combination) D. 投票(voting)

1. 检测一元正太分布中的离群点，属于异常检测中的基于（A）的离群点检测。

A. 统计方法 B. 邻近度 C. 密度 D.聚类技术

1. 熵是为消除不确定性所需要获得的信息量，投掷均匀正六面体骰子的熵是（B）

A. 1bit B. 2.6bit C. 3.2bit D. 3.8bit

6. 进行实体识别时一般不使用的方法是（D）

A. CRF

B. SVM

C. 朴素贝叶斯

D. K近邻

7. 以下不是文本向量表示方法的是（D）

A. One-Hot B. TF C. TF-IDF D. SVM

8. 线性支持向量机的损失函数为（C）

A. B. C. D.

9. 区别于英文，中文信息处理技术中最基础、最重要的问题是（A）

A. 分词 B. 词性标注 C. 未登录词识别 D. 词义消歧

10. 以下哪个应用属于序列标注问题（B）

A. 情感分析 B. 命名实体识别 C. 垃圾邮件检测 D. 新闻文本分类

11. 已知一个词语序列s=w1,w2,…,wn,若P(s)=P(w1)P(w2)…P(wn)，则事件“词语wi出现,i=1,2,…,n”之间的关系为（D）

A. 相互独立 B. 互斥

C. 对立 D. 不一定相互独立

# 简答题

1. 简述构建文本分类系统的流程或主要模块。

**参考解答：**一般文本分类系统的构建分为训练和预测。主要的模块有：分词模块、特征词抽取模块、分类模型模块等。

1. 简述文本向量化表示的方法并写出每种方法的特点。

**参考解答：**文本向量化表示主要有两种一种为VSM向量空间模型，一种为词嵌入模型。其中VSM向量空间模型中又细分为One-Hot表示法、TF表示法、TF-IDF表示法。VSM向量空间维度高，各维度词语明确，忽略词语之间的顺序关系。词嵌入模型维度低，各维度意义不明确，用词语的上下文表达当前词语的向量方向。

1. 文本分类语料如下，请用多项式朴素贝叶斯法计算待预测文本属于各类别的概率。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 文档ID | 文档词列表 | 是否属于China类 |
| 训练集 | 1 | Chinese Beijing Chinese | Yes |
| 2 | Chinese Chinese Shanghai | Yes |
| 3 | Chinese Macao | Yes |
| 4 | Tokyo Japan Chinese | No |
| 测试集 | 5 | Chinese Chinese Chinese Tokyo Japan | ? |

**参考解答：**P(c=China|Chinese Chinese Chinese)=0.000299，P(c!=China|Chinese Chinese Chinese)=0.000135。

1. 简述朴素贝叶斯模型与线性支持向量机模型的区别。

**参考解答：**朴素贝叶斯属于生成模型，模型的训练过程为计算给定类别条件下出现某个词语的概率。线性支持向量机属于判别模型，模型的训练过程是求使P(y|x1,x2,…,xn; )概率最大的参数向量。朴素贝叶斯模型不需要将文本进行向量化表示而线性SVM要求输入为文本的数值化向量。朴素贝叶斯是多分类模型，线性SVM为二分类模型。朴素贝叶斯模型既不是线性模型也不是非线性模型，线性SVM的模型为线性表达式，因此为线性模型。

1. 写出隐马尔可夫模型的三个基本问题并简述利用HMM进行分词的步骤。

**参考解答：**三个基本问题为：概率计算问题，学习问题，预测问题。在利用HMM进行分词训练时，需要统计隐马模型的三要素即初始概率矩阵、概率转移矩阵、观测概率矩阵。在预测阶段求解第3个基本问题即预测问题。

# 英文翻译

1. 请将下边的一段话翻译为中文。

In this chapter, we will introduce the principal approaches to finding collocations: selection of collocations by frequency, selection based on mean and variance of the distance between focal word and collocating word, hypothesis testing, and mutual information. We will then return to the question of what a collocation is and discuss in more depth different definitions that have been proposed and tests for deciding whether a phrase is a collocation or not. The chapter concludes with further readings and pointers to some of the literature that we were not able to include.

1. 请将下边的一段话翻译为中文。

VarianceThreshold is a simple baseline approach to feature selection. It removes all features whose variance doesn’t meet some threshold. By default, it removes all zero-variance features, i.e. features that have the same value in all samples. As an example, suppose that we have a dataset with boolean features, and we want to remove all features that are either one or zero (on or off) in more than 80% of the samples. Boolean features are Bernoulli random variables, and the variance of such variables is given by Var[X] = p(1 - p).