第5章　感知机分类与序列标注

第 4 章我们利用隐马尔可夫模型实现了第一个基于序列标注的中文分词器，然而效果并不理想。事实上，隐马尔可夫模型假设人们说的话仅仅取决于一个隐藏的 {B,M,E,S} 序列，这个假设太单纯了，不符合语言规律。语言不是由这么简单的标签序列生成，语言含有更多特征，而隐马尔可夫模型没有捕捉到。隐马尔可夫模型能捕捉的特征仅限于两种：其一，前一个标签是什么；其二，当前字符是什么。

为了利用更多的特征，**线性模型**（linear model）应运而生。线性模型由两部分构成：一系列用来提取特征的特征函数 ϕ，以及相应的权重向量 $\boldsymbol{w}$。也许读者还记得第 1 章中的人名性别预测问题，当时我们利用一系列特征模板抽取了人名的特征向量，这些特征模板就充当了特征函数的角色。而当时我们推导的训练算法，就是感知机算法。

本章将深入讲解感知机算法的原理，以及在分类和序列标注上的应用。在分类问题的应用部分，我们将把性别识别问题写成代码。在序列标注应用部分，我们将实现基于感知机的中文分词器。由于感知机序列标注基于分类，并且分类问题更简单，所以我们先学习分类问题。

5.1　分类问题

5.1.1　定义

分类（classification）指的是预测样本所属类别的一类问题。形式化表述，分类问题的目标就是给定输入样本 $\boldsymbol{x}$，将其分配给 K 种类别 $\mathcal{C}_k$ 中的一种，其中 $k=1,\cdots,K$。如果 $K=2$，则称为**二分类**（binary classification），否则称为**多分类**（multiclass classification）。比如第 1 章所讲的性别预测问题，目标是判断姓名属于 {男,女} 中的哪一种，这是典型的二分类问题。

正如二进制可以表示任意进制一样，二分类也可以解决任意类别数的多分类问题。具体说来，有 one-vs-one 和 one-vs-rest（也称为 one-vs-all）两种方案。两种方案都可以想象为 K 种类别参加球赛。

- one-vs-one：进行多轮二分类，每次区分两种类别 $\mathcal{C}_i$ 和 $\mathcal{C}_j$。一共进行 $\binom{K}{2}$ 次二分类，理想情况下有且仅有一种类别 $\mathcal{C}_k$ 每次都胜出，于是预测结果为 $\mathcal{C}_k$。
- one-vs-rest：依然是多轮二分类，每次区分类别 $\mathcal{C}_i$ 与非 $\mathcal{C}_i$。一共进行 K 次二分类。理想情