**文本向量化**

将一个句子处理为一个向量。

1. 词袋模型

首先统计数据中出现的所有单词，然后将每个句子记录为一个数值向量的形式——a[i]表示第i个单词出现的频率。

1. N元特征

长度为N的滑动窗口，将长度为L的句子划分为L-N+1个词组。划分完成后，再利用词袋模型。

在实际使用中，可以同时取N=1 2 3，防止出现过拟合。

**文本分类**

对超平面中的点划分聚类。

对于多分类问题，设类数为C，先用C个边界把属于c和不属于c的部分划分开，记为fc，其中fc(x)>0表示x属于c这个类。

然后对于每个测试集中的元素，取使fc(x)最大的c。（argmax方式）

1. Logistic回归
2. Softmax回归