

集。在训练时，尝试多组超参数的取值，并在验证集上检验哪一组超参数所对应的模型取得了最好的效果。最终，在验证集上效果最好的一组超参数和其对应的模型将被选定，并在测试集上进行测试。

为了衡量LDA模型在验证集和测试集上的效果，需要寻找一个合适的评估指标。一个常用的评估指标是困惑度（perplexity）。在文档集合 D 上，模型的困惑度被定义为

$$\text{perplexity}(D) = \exp \left\{ - \frac{\sum_{d=1}^M \log p(\mathbf{w}_d)}{\sum_{d=1}^M N_d} \right\}, \tag{6.28}$$

其中 M 为文档的总数， $\mathbf{w}_d$ 为文档 d 中单词所组成的词袋向量， $p(\mathbf{w}_d)$ 为模型所预测的文档 d 的生成概率， N_d 为文档 d 中单词的总数。

一开始，随着主题个数的增多，模型在训练集和验证集的困惑度呈下降趋势，但是当主题数目足够大的时候，会出现过拟合，导致困惑度指标在训练集上继续下降但在验证集上反而增长。这时，可以取验证集的困惑度极小值点所对应的主题个数作为超参数。在实践中，困惑度的极小值点可能出现在主题数目非常大的时候，然而实际应用并不能承受如此大的主题数目，这时就需要在实际应用中合理主题数目范围内进行选择，比如选择合理范围内困惑度的下降明显变慢（拐点）的时候。

另外一种方法是在LDA基础之上融入分层狄利克雷过程（Hierarchical Dirichlet Process, HDP），构成一种非参数主题模型HDP-LDA。非参数主题模型的好处是不需要预先指定主题的个数，模型可以随着文档数目的变化而自动对主题个数进行调整；它的缺点是在LDA基础上融入HDP之后使得整个概率图模型更加复杂，训练速度也更加缓慢，因此在实际应用中还是经常采用第一种方法确定合适的主题数目。

问题3 如何用主题模型解决推荐系统中的冷启动问题？

难度：★★★★☆☆

分析与解答

首先对题目做进一步的解释。推荐系统中的冷启动问题是指在有大量用户数据的情况下如何给用户进行个性化推荐，目的是最优化点击率、转化率或用户体验（用户停留时间、留存率等）。冷启动问题一般分为用户冷启动、物品冷启