水滴石穿

探索，保持渴望，无所畏惧

Word Embedding札记

Dec 18, 2016 | 机器学习 | 1097 Hits

如何利用文本的上下文信息，得到更有意义的向量表达(word embedding)，是NLP领域研究的重点。本篇笔记目的在于整理词向量的发展历程，方便理解什么是词向量，怎么得到词向量。词向量也叫词的分布式表达，主要有三类方法：聚类，矩阵分解，神经网络。

文章目录

基于聚类的分布表示(clusteringbased word representation)

这类方法通过聚类将词和类的标签建立关联，关联关系可以是确定性也可以是概率表示，用这种方式构建词与其上下文之间的关系。相关工作有(Brown, 1992)[10], (Ney, 1993)[35], (Niesler, 1998)[36]。布朗聚类(Brown, 1992)[10]是一种层级聚类方法，聚类结果为每个词的多层类别体系。可以根据两个词的公共类别判断这两个词的语义相似度。

基于矩阵的分布表示(distributional representation)

构建一个“词-上下文”矩阵，从矩阵中获取词的表示。在“词-上下文”矩阵中，每行对应一个词，每列表示一种不同的上下文，矩阵中的每个元素对应相关词和上下文的共现次数。
a. 矩阵构造
b. 矩阵元素值的确定
c. 降维技术将高维稀疏的向量压缩成低维稠密
典型如Latent Semantic Analysis (LSA) 的做法，构造word-doc矩阵，TF-IDF为每个元素的值。使用SVD分解，得到词的低维表达(Deerwester, 1990)[37] (Bellegarda, 1997)[34]。介绍两份比较近的工作：

- GloVe(Pennington, 2014)[27]，GloVe 模型是一种对“词-词”矩阵进行分解从而得到词表示的方法。矩阵第i行第j列的值为词v_i与词v_j 在语料中的共现次数x_{ij} 的对数。在矩阵分解步骤，GloVe 模型借鉴了LSA (Deerwester,1990)[31]，在计算重构误差时，只考虑共现次数非零的矩阵元素，同时对矩阵中的行和列加入了偏移项，根据共现词频对重构误差进行