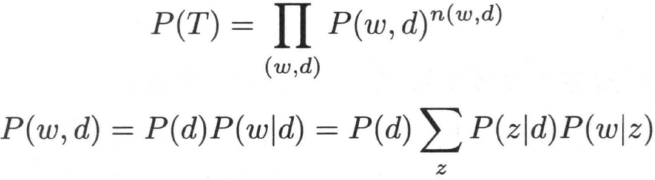
读书报告

1. （必填）读书计划

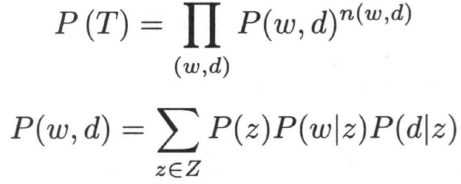
1、本周完成的内容章节：《统计学习方法》十八章

2、下周计划：《统计学习方法》十九章

1. 读书摘要
   * 1. 概率潜在语义分析，也称概率潜在语义检索，是一种利用概率生成模型对文本集合进行话题分析的无监督学习方法。模型的最大特点是用隐变量表示话题；整个模型表示文本生成话题，话题生成单词，从而得到单词-文本共现数据的过程；假设每个文本由一个话题分布决定，每个话题由一个单词分布决定。
     2. 基本思想：给定一个文本集合，每个文本讨论若干个话题，每个话题由若干个单词表示。对文本集合进行概率潜在语义分析，就能发现每个文本的话题，以及每个话题的单词。话题是不能从数据中直接观察到的，是潜在的。文本集合转换为文本——单词共现数据，具体表现为单词-文本矩阵。每一行对应一个单词，每一列对应一个文本，每一杠元素表示单词在文本中出现的次数。一个话题表示一个语义内容。文本数据基于如下的概率模型产生（共现模型）；首先有话题的概率分布，然后有话题给定条件下的条件概率分布，以及话题给定条件下单词的条件概率分布。概率潜在语义分析就是发现由隐变量表示的话题，即潜在语义。
     3. 生成模型表示文本生成话题，话题生成单词，从而得到单词-文本共现数据的过程:假设每个文本由一个话题分布决定，每个话题由一个单词分布决定。单词变量ω与文本变量d 是观测变量话题变量z 是隐变量。生成模型的定义如下：



* + 1. 共现模型描述文本单词共现数据拥有的模式。共现模型的定义如下：



* + 1. 概率潜在语义分析的学习通常采用EM算法。通过迭代学习模型的参数，和，而P(d)可直接统计得到。EM算法是一种迭代算法，每次迭代包括交替的两步：E步，求期望；M步，求极大。E步时计算Q函数，即完全数据的对数似然函数对不完全数据的条件分布的期望。M步是对Q函数极大化，更新模型参数。