产生模型与判别模型的比较研究

内容提要

- 摘要
- 简介
- 产生式模型（Generative Model）
- 判别式模型（Discriminative Model ）
- 两者之间的关系
- 实验分析

摘要

- 产生式模型：无穷样本==》概率密度模型 ＝＝》 产生模型==》预测
- 判别式模型：有限样本==》判别函数 ＝ 预测模型==》预测

简介

- 简单的说，假设o是观察值，q是模型。
  如果对P(o|q)建模，就是Generative模型。
- 其基本思想是首先建立样本的概率密度模型，再利用模型进行推理预测。要求已知样本无穷或尽可能的大。
  这种方法一般建立在bayes理论的基础之上。

简介

- 如果对条件概率 P(q|o)建模，就是Discrminative模型。基本思想是有限样本条件下建立判别函数，不考虑样本的产生模型，直接研究预测模型。代表性理论为统计学习理论。

- 这两种方法目前交叉较多。

产生式模型

- 估计的是联合概率分布（joint probability distribution），p(class, context)=p(class|context)*p(context)。
  ＝p(context|class)*p(class)
- 用于随机生成的观察值建模，特别是在给定某些隐藏参数情况下。
- 在机器学习中，用于直接对数据建模,或作为生成条件概率密度函数的中间步骤。通过使用贝叶斯规则可以从生成模型中得到条件分布。

产生式模型

- 特点：
  主要是对后验概率建模，从统计的角度表示数据的分布情况，能够反映同类数据本身的相似度。
- 优点：
  由于产生式方法可以在联合分布空间插入变量、不变量、独立性、先验分布等关系的知识。因此，在联合分布空间，通用性（或称多面性）是其本质。这包括了系统中的未知的、观察到的、输入或输出变量，这就使得产生式概率分布成为一个非常灵活的建模工具。

产生式模型

- 缺点
- 产生式分类器需产生的所有变量的联合概率分布仅仅是分类任务的中间目标，对该中间目标优化的过程，牺牲了最终分类判别任务上的资源和性能，影响了最终的分类性能。

产生式模型

- 常用方法

  Gaussians, Naive Bayes,

  Mixtures of multinomials
  Mixtures of Gaussians,

  HMMs
  Bayesian networks
  Markov random fields

判别式模型

- 又可以称为条件模型，或条件概率模型。估计的是条件概率分布(conditional distribution)

- 判别式方法并不对系统中变量和特征的基本分布建模，仅仅对输入到输出之间映射的最优化感兴趣。因此，仅需调整由此产生的分类边界，没有形成可对系统中变量建模的生成器的中间目标，可以得到准确率更高的分类器。

判别式模型

- 主要特点：
  寻找不同类别之间的最优分类面，反映的是异类数据之间的差异。
- 优点：
  相比纯概率方法或产生式模型，分类边界更灵活；
  能清晰的分辨出多类或某一类与其他类之间的差异特征，适用于较多类别的识别
  判别模型的性能比产生模型要简单，比较容易学习

判别式模型

- 缺点：
- 不能反映训练数据本身的特性。可以告诉你的是1还是2，但没有办法把整个场景描述出来；
- 判别式方法在训练时需要考虑所有的数据元组，当数据量很大时，该方法的效率并不高；
- 缺乏灵活的建模工具和插入先验知识的方法。因此，判别式技术就像一个黑匣子，变量之间的关系不像在产生式模型中那样清晰可见。

判别式模型

- 常见的主要有：
  logistic regression
  SVMs
  neural networks
  Nearest neighbor
  Conditional random fields(CRF): 目前最新提出的热门模型，从NLP领域产生的，正在向ASR和CV上发展。

判别式模型

- 主要应用：
  Image and document classification
  Biosequence analysis
  Time series prediction

两者之间的关系

- 由生成模型可以得到判别模型，但由判别模型得不到生成模型。
- 例如当样本的各属性之间相互独立的并且满足高斯概率密度分布时，可以由Naïve Bayes分类算法得到Logistic Regression分类算法

实验分析

- 实验内容
- 对于UCI的Adult 数据集、Breast Cancer数据集、Ionosphere数据集以及Optical Recognition of Hand Written Digits 数据集，分别用Naïve Bayes算法与Logistic Regression算法对其进行分类，并对这两种算法进行比较分析。

实验分析

- 实验结果

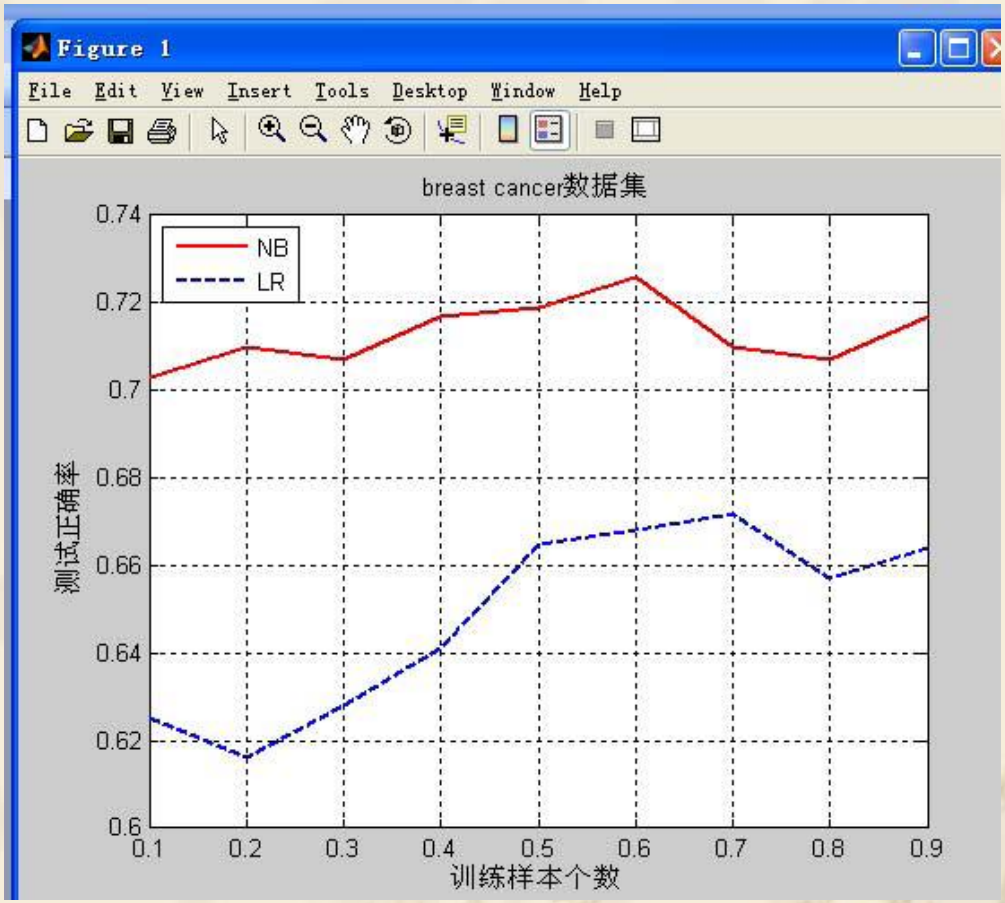

图1 breast cancer 数据集上NB与LR分类结果比较

实验分析

- 实验结果

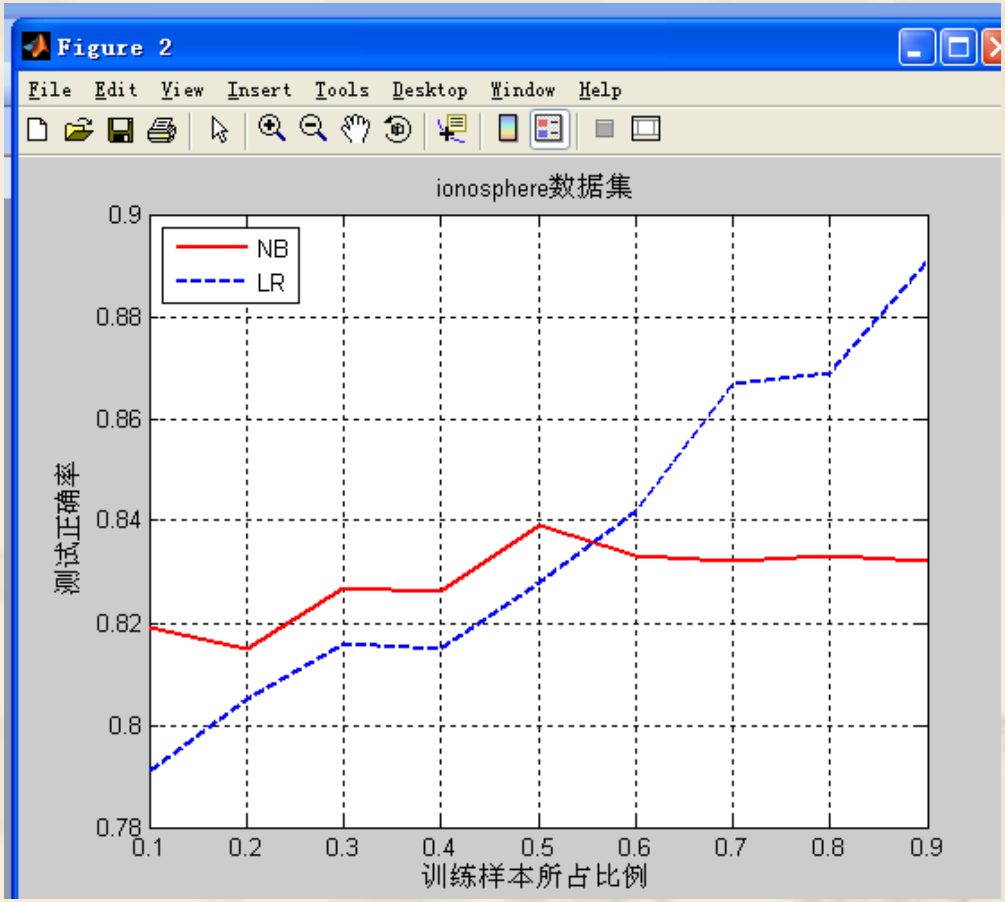

图2 ionosphere 数据集上NB与LR分类结果比较

实验分析

- 实验结果

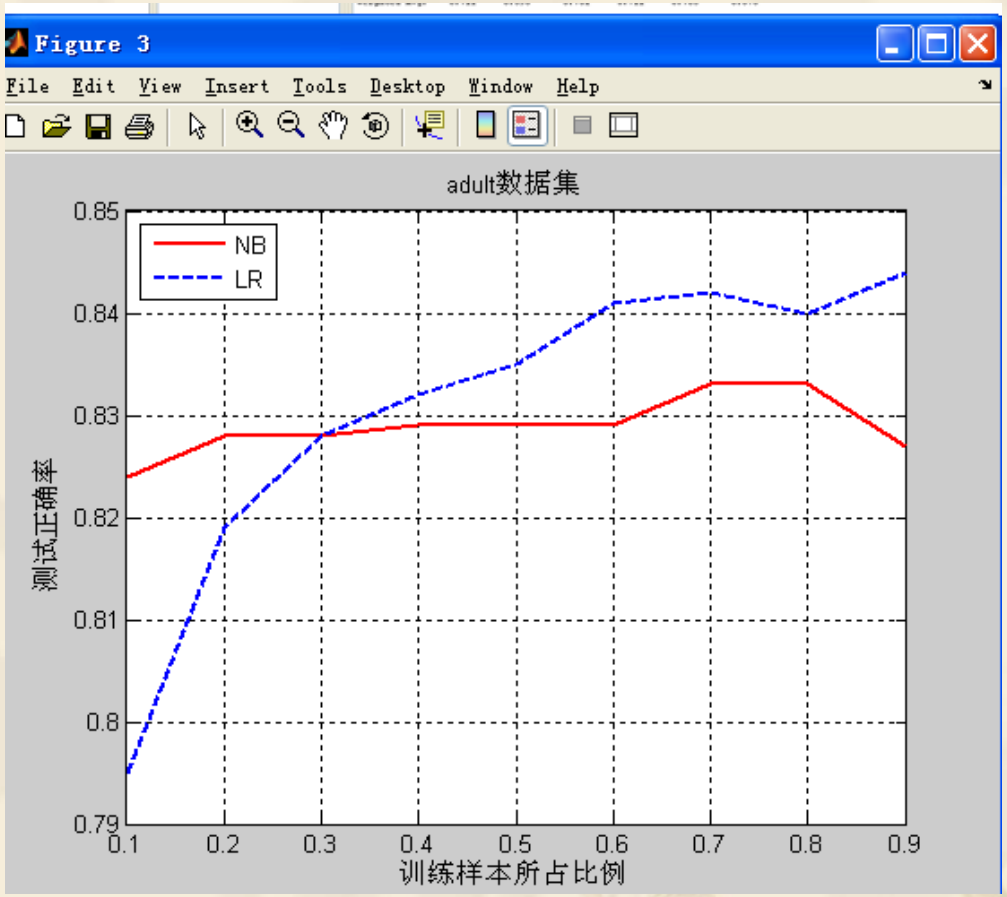

图3 adult 数据集上NB与LR分类结果比较

实验分析

- 实验结果

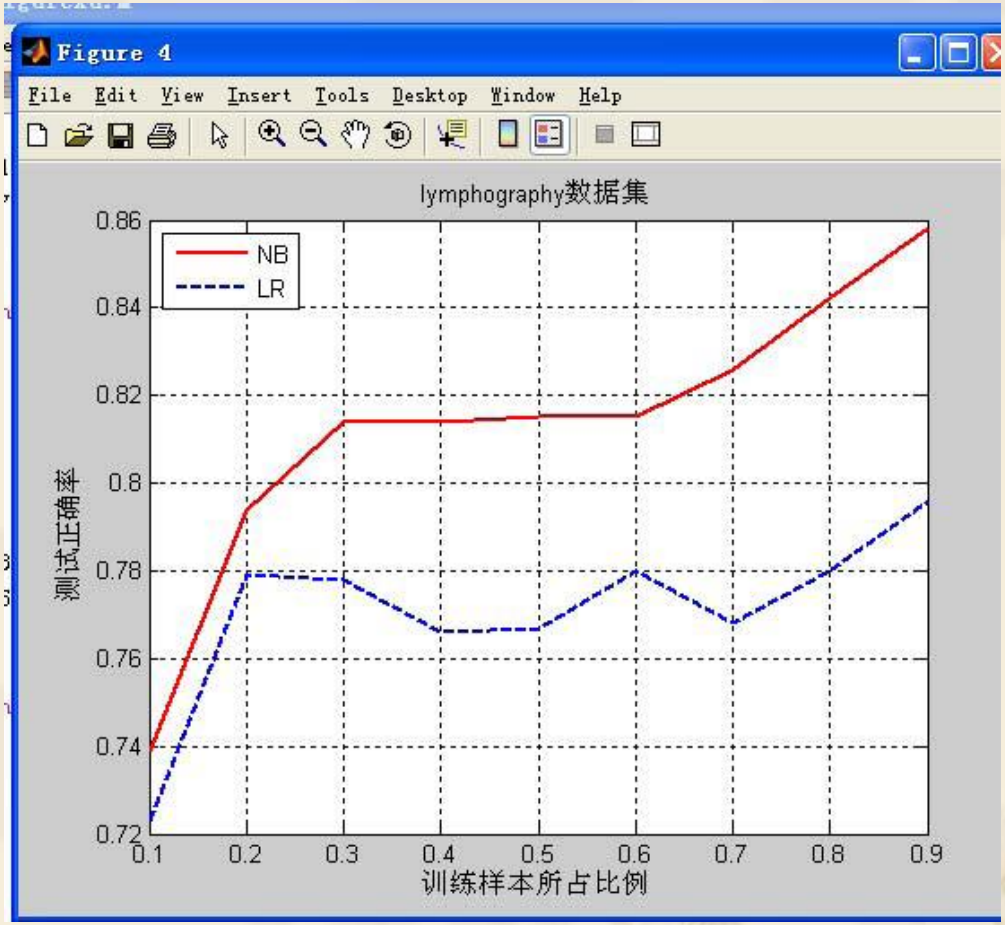

图4 lymphography 数据集上NB与LR分类结果比较

实验分析

- 结果分析
- 从实验结果可见，在**breast_cancer**和**lymphography**数据集上**NaiveBayes**分类器的分类效果明显优于**Logistic Regression**分类器
- 在四个数据集中，**adult**数据集的数据量最大，因此在这个数据集上可以看出**Logistic Regression**分类器的训练时间明显大于**NaiveBayes**分类器。

实验分析

- 结果分析
- 对于adult和ionosphere数据集，图中显示NaiveBayes分类器的分类正确率曲线与Logistic Regression分类器的分类正确率曲线有交叉的现象。当训练数据较少的时候Logistic Regression分类器的效果比较差，随着训练数据的增加其对测试数据的分类正确率快速增加。而NaiveBayes分类器对于训练数据的多少并不敏感，分类效果比较稳定。可见，在训练数据较少时应选择NaiveBayes分类器。