机器学习 I

雷至祺

文本数据

- 无处不在：超过80%的大数据以文本形式存在
- 来源多样：社交媒体、新闻门户、电商平台、科研文献等
- 如何自动挖掘文本数据，得到有用的知识和信息？
  - 例如：情感倾向

文本数据

Great product!

Reviewed in the United States us on October 11, 2022

Verified Purchase

I've been using it for the last 4 months and it's very very good!

Helpful | Report

Not nearly as good as my previous Kindle.

Reviewed in the United States us on October 2, 2022

Verified Purchase

It does not immediately download a book when you buy it. This means that when you holiday somewhere without WiFi you don't have any books to read. If you try to read a book in the sun, after a very short time, the screen freezes. The battery is not nearly as long lasting as my previous Kindle. The organization of the books in its library is atrocious. It tells you haven't read a book because you only read 99%. The other 1% being an advertisement for the next book by that Author.

2 people found this helpful

Helpful | Report

热 一家三口能不能住标间？酒店规定引争议

界面新闻记者 |李如嘉 近日，有网友发帖称，自己预订北京东城区某酒店的标间，办理入住时，酒店前台却称一家三口不能住在同一个标间，并表示这是行业普遍规定。[详细]

2023年 3月29日 18:39 评论(1087)

热 钱都去哪里了？房地产泡沫中的货币供给

摘 要 中国房地产泡沫来自土地供应的不平衡，控制货币供应并无法精准地抑制地产泡沫，反而会影响到很多和地产毫无关系的产业。解决问题的关键是用财政和行政政策直接调节其...[详细]

2023年 3月29日 18:37 评论(592)

热 孔乙己为何总被误读？

北京晚报·五色土 | 作者 蔡辉 “十日晴。录文稿一篇讫，约四千余字，寄高一涵并函，由二弟持去。夜风。”这是鲁迅于1919年3月10日，写下的日记。[详细]

2023年 3月29日 15:17 鲁迅 孔乙己 周作人 评论(255)

热 百岁“扫地僧”，去了天堂图书馆

来源：环球人物杂志 天堂应该是图书馆的模样， 沈窥元先生只是去了另一个图书馆。 作者：许晓迪 今晨8点22分，版本目录学家沈窥元在南京去世，享年100岁。[详细]

2023年 3月29日 15:03 评论(666)

美团外卖没有大龄骑手？回应：近期未对年龄限制做调整

美团外卖没有大龄骑手？回应：近期未对年龄限制做调整 澎湃新闻记者 范佳来 近日，有用户在社交媒体晒出收到的美团《配送服务年龄到限通知》...[详细]

2023年 3月29日 14:34 评论(52)

文本情感分析

- 给定一篇文本:
  - 预测文本属于哪种情感
  - 预测文本在每种情感上的分值

- 如何构建机器学习模型，完成上述任务?

实验5

- 第1次课：朴素贝叶斯
- 第2次课：文本处理与k-NN

- 实验任务：文本情感分析
  - 在朴素贝叶斯分类、k-NN分类与k-NN回归中，三者至少完成两项；
  - 鼓励尝试多种算法及算法中的不同策略/参数，并进行结果对比分析；
  - 完成一份实验报告，注意实验报告要求。

朴素贝叶斯

雷至祺

2023.04.18

有监督学习

- 有监督学习的基本步骤：上课—考试
  - 给出带标签的训练数据
  - 用训练数据训练模型至一定程度
  - 用训练好的模型预测不带标签的数据的类别

- 常见的有监督学习问题：
  - 分类问题：预测离散值的问题（如预测明天是否会下雨）
  - 回归问题：预测连续值的问题（如预测明天气温是多少度）

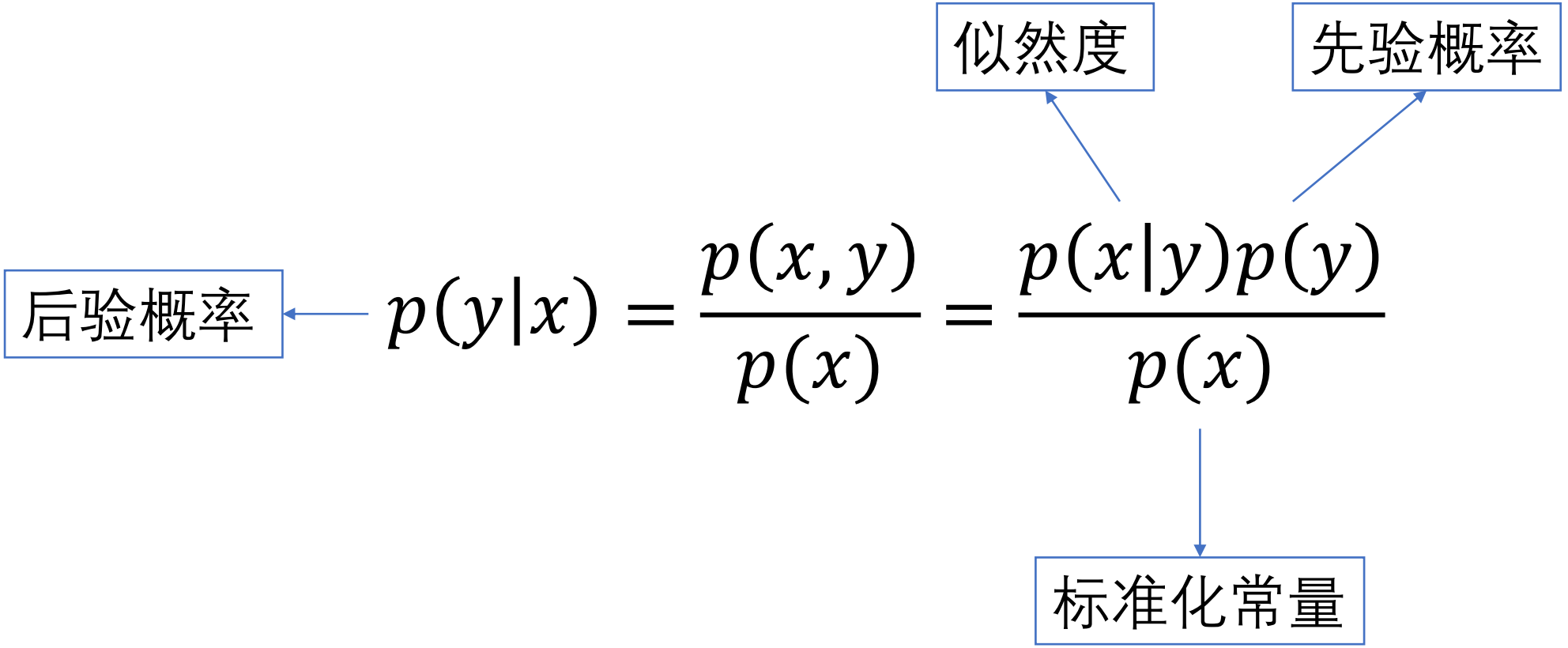
分类

- 若记文档为x，文档所属的类别为y（$y = y_1, y_2 \ldots$）
- 如何用条件概率的形式给出分类的目标?

$$p(y|x)$$

- 判别模型：如k-NN
  - 直接估计条件概率 $p(y|x)$ 。
- 生成模型：如朴素贝叶斯
  - 估计先验概率 $p(y)$ 和条件概率 $p(x|y)$;
  - 根据贝叶斯定理计算后验概率 $p(y|x)$ 。

贝叶斯定理



贝叶斯分类器

- 贝叶斯分类器根据贝叶斯定理来估计每个类别的后验概率。

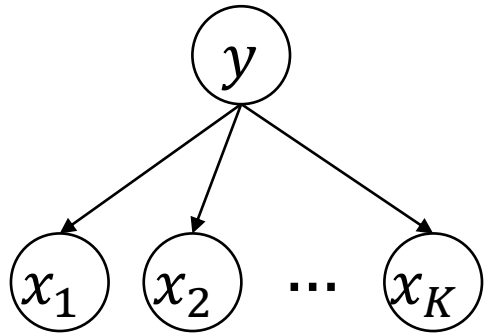
$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_i p(x|y_i)p(y_i)} \propto p(x|y)p(y)$$

- 贝叶斯分类器的目标是找到：

$$y = \arg\max_y p(y|x) = \arg\max_y p(x|y)p(y)$$

朴素贝叶斯分类：假设

$$y = \arg\max_y p(y|x) = \arg\max_y p(x|y)p(y)$$
$$= \arg\max_y p(x_1, x_2, \dots, x_K|y)p(y)$$



- 朴素贝叶斯（Naïve Bayes）的条件独立性假设
  - 假设在给定类别的条件下，特征变量间是独立的

$$p(x_1, x_2, \dots, x_K|y) = \prod_{k=1}^{K} p(x_k|y)$$

- 例如，对于文档x，其可由单词$x_1, x_2, \dots, x_K$构成特征

朴素贝叶斯分类

- 给定一个包含N篇文档的数据集，其中每个样本由一篇文档 x 和一个情感标签 e_i构成
- 给定包含K个单词的测试文档 $x = (x_1, \dots, x_K)$
- 为了预测测试文档的类别，需要估计：

$$\begin{aligned}\arg\max_{e_i} p(e_i|x) &= \arg\max_{e_i} \frac{p(x|e_i)p(e_i)}{p(x)} \\ &= \arg\max_{e_i} p(x|e_i)p(e_i) \\ &= \arg\max_{e_i} \prod_{k=1}^{K} p(x_k|e_i)p(e_i)\end{aligned}$$

朴素贝叶斯分类

- 朴素贝叶斯分类：
  - 训练：为各个情感类别估计先验概率 $p(e_i)$ 和似然度$\prod_{k=1}^{K} p(x_k|e_i)$；
  - 预测：对每篇测试文档，依照下式得到最符合的情感类别：

$$\arg\max_{e_i} p(e_i|x) = \arg\max_{e_i} \prod_{k=1}^{K} p(x_k|e_i)p(e_i)$$

- 如何建模 $p(x_k|e_i)$ 和 $p(e_i)$ ？

文本分类：例子

$$\arg\max_{e_i} p(e_i|x) = \arg\max_{e_i} \prod_{k=1}^{K} p(x_k|e_i)p(e_i)$$

- 如何建模 $p(e_i)$ ？
  - 在训练集中，类别为e_i的文档占比

	ID	text	class label
训练集	1	good,thanks	joy
	2	No impressive, thanks	sad
	3	Impressive good	joy
测试集	4	No, thanks	?

$$p(joy) = \frac{2}{3}$$

- 如何建模 $p(x_k|e_i)$ ？
- 模型1：类别为e_i的训练文档中，包含单词x_k的文档占比

$$p(thanks|joy) = \frac{1}{2}$$

- 称为：伯努利模型

文本分类：伯努利模型

- 如何建模 $p(x_k|e_i)$ 和 $p(e_i)$ ？
- 伯努利模型（Bernoulli Model）：

$$p(x_k|e_i) = \frac{n_{e_i}(x_k)}{N_{e_i}} \qquad p(e_i) = \frac{N_{e_i}}{N}$$

- 其中：
  - $n_{e_i}(x_k)$表示类别为e_i的文本中出现x_k的文本数量；
  - N_{e_i}表示类别为e_i的文本数量；
  - N表示全部文本的数量。

文本分类：例子

$$\arg\max_{e_i} p(e_i|x) = \arg\max_{e_i} \prod_{k=1}^{K} p(x_k|e_i)p(e_i)$$

- 如何建模 $p(e_i)$ ？
  - 在训练集中，类别为e_i的文档占比

	ID	text	class label
训练集	1	good,thanks	joy
	2	No impressive, thanks	sad
	3	Impressive good	joy
测试集	4	No, thanks	?

$$p(joy) = \frac{2}{3}$$

- 如何建模 $p(x_k|e_i)$ ？
- 模型2：类别为e_i文档的所有单词中，单词x_k出现次数的占比

$$p(thanks|joy) = \frac{1}{4}$$

- 称为：多项式模型

文本分类：多项式模型

- 如何建模 $p(x_k|e_i)$ 和 $p(e_i)$ ？
- 多项式模型（Multinomial Model）：

$$p(x_k|e_i) = \frac{nW_{e_i}(x_k)}{nW_{e_i}} \qquad p(e_i) = \frac{N_{e_i}}{N}$$

- 其中：
  - $nW_{e_i}(x_k)$表示类别为e_i的文本中出现x_k的总次数；
  - nW_{e_i}表示类别为e_i的文本中单词的总数；
  - N_{e_i}表示类别为e_i的文本数量；
  - N表示全部文本的数量。

思考题：伯努利模型和多项式模型分别有什么优缺点？

文本分类：例子

ID	text	class label
1	good,thanks	joy
2	No impressive, thanks	sad
3	Impressive good	joy
4	No, thanks	?

$$\arg\max_{e_i} p(e_i|x) = \arg\max_{e_i} \prod_{k=1}^{K} p(x_k|e_i)p(e_i)$$

以多项式模型为例

- $p(sad|d_4) \propto p(d_4|sad) \cdot p(sad)$

  $= p(thanks, no|sad) \cdot p(sad)$

  $= p(thanks|sad) \cdot p(no|sad) \cdot p(sad)$

  $= \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

- $p(joy|d_4) \propto p(d_4|joy) \cdot p(joy)$

  $= p(thanks, no|joy) \cdot p(joy)$

  $= p(thanks|joy) \cdot p(no|joy) \cdot p(joy)$

  $= \frac{1}{4} \times \frac{0}{4} \times \frac{2}{3} = 0$

平滑处理

- 在前面的文本分类算法中，如果测试文本中的单词没有在训练文本中出现会造成什么结果?
- 会影响到后验概率的计算结果，使分类产生偏差。
- 解决这一问题的方法是采用平滑处理 (smoothing):

$$Bernoulli: p(x_k|e_i) = \frac{n_{e_i}(x_k) + 1}{N_{e_i} + 2}$$

$$Multinomial: p(x_k|e_i) = \frac{nw_{e_i}(x_k) + 1}{nw_{e_i} + V_{e_i}}$$

- 这里V_{e_i}指得是类别为e_i的文本中不重复的单词数量。

平滑处理：例子

ID	text	class label
1	good,thanks	joy
2	No impressive, thanks	sad
3	Impressive good	joy
4	No, thanks	?

$$Bernoulli: p(x_k|e_i) = \frac{n_{e_i}(x_k) + 1}{N_{e_i} + 2}$$

$$Multinomial: p(x_k|e_i) = \frac{nw_{e_i}(x_k) + 1}{nw_{e_i} + V_{e_i}}$$

- 伯努利模型（Bernoulli Model）

$$p(no|joy) = \frac{0+1}{2+2} = \frac{1}{4}$$

- 多项式模型（Multinomial Model）

$$p(no|joy) = \frac{0+1}{4+3} = \frac{1}{7}$$

拉普拉斯平滑

- 更通用的一种平滑处理如下：

$$p(x_k|e_i) = \frac{x_k + \lambda}{\sum_{k=1}^{K} x_k + K\lambda}$$

- 其中，K为特征总数（词表大小），$\lambda \geq 0$。
  - 相当于：对随机变量各个取值的频数，加上一个正数 $\lambda \geq 0$。
  - 当 $\lambda = 0$ 时，即为极大似然估计。
  - 常取 $\lambda = 1$，此时称为拉普拉斯平滑 (Laplacian smoothing)。

实验5

- 第9周：朴素贝叶斯
- 第10周：文本处理与k-NN

- 实验任务：文本情感分析
  - 在朴素贝叶斯分类、k-NN分类与k-NN回归中，三者至少完成两项；
  - 鼓励尝试多种算法及算法中的不同策略/参数，并进行结果对比分析；
  - 完成一份实验报告，注意实验报告要求。

有监督学习：数据集划分

有监督学习：数据集划分

训练集 验证集 测试集的区别

数据类型	有无标签	作用
训练集(training set)	有	用来训练模型或确定模型参数的，如k-NN中权值的确定等。 相当于平时练习。
验证集(validation set)	有	用来确定网络结构或者控制模型复杂程度的参数，修正模型。 相当于模拟考试。
测试集(test set)	无	用于检验最终选择最优的模型的性能如何。 相当于期末考试。

任务5-1：朴素贝叶斯分类

- 使用朴素贝叶斯进行分类任务
- 在平滑处理中，要求至少实现拉普拉斯平滑
- 通过在验证集上尝试不同的概率模型和平滑处理等来筛选准确率最好的一组参数，并将该过程记录在实验报告中
- 数据目录为classification_dataset，其中train_set用于训练，validation_set是验证集
- 在测试集test_set上应用调参后参数做预测，输出结果保存为“**学号_姓名拼音_NB_classification.csv**”
  - 文件内部格式参考“21881234_pinyin_model_classification.csv”

分类实验数据介绍

```
Words (split by space),label
europe retain trophy with big win,joy
senate votes to revoke pensions,sad
```

- 数据一共有两列，其中每一列用英文逗号隔开。
- 第一列为文档，词之间用空格隔开；
- 第二列是标签。

实验提交

- **作业名称：实验5**
- **截止时间：5月10日 23:59**
- 本次实验提交样例：压缩包21******_wangxiaoming.zip，内含：
  - 21******_wangxiaoming.pdf
  - /code：文件夹，内含所有实验代码并附上readme
  - /result：文件夹，内含实验结果（根据完成情况，至少包含两个）
    - 21******_wangxiaoming_NB_classification.csv
    - 21******_wangxiaoming_KNN_classification.csv
    - 21******_wangxiaoming_KNN_regression.csv

实验验收

- 验收日期：4月23日/4月25日/5月9日实验课
- 验收形式：助教上传一个小数据集到Q群中，下载好然后课上验收时当场跑程序，TA会根据结果判断算法是否正确。

Q & A

附录

文件读写

C++:

http://blog.csdn.net/kingstar158/article/details/6859379/

Java:

http://blog.csdn.net/jiangxinyu/article/details/7885518/

Python:

http://www.cnblogs.com/allenblogs/archive/2010/09/13/1824842.html

字符串分割

C++:

http://blog.csdn.net/glt3953/article/details/11115485

Java:

http://blog.sina.com.cn/s/blog_b7c09bc00101d3my.html

Python:

http://blog.sina.com.cn/s/blog_81e6c30b01019wro.html