法律声明

- 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

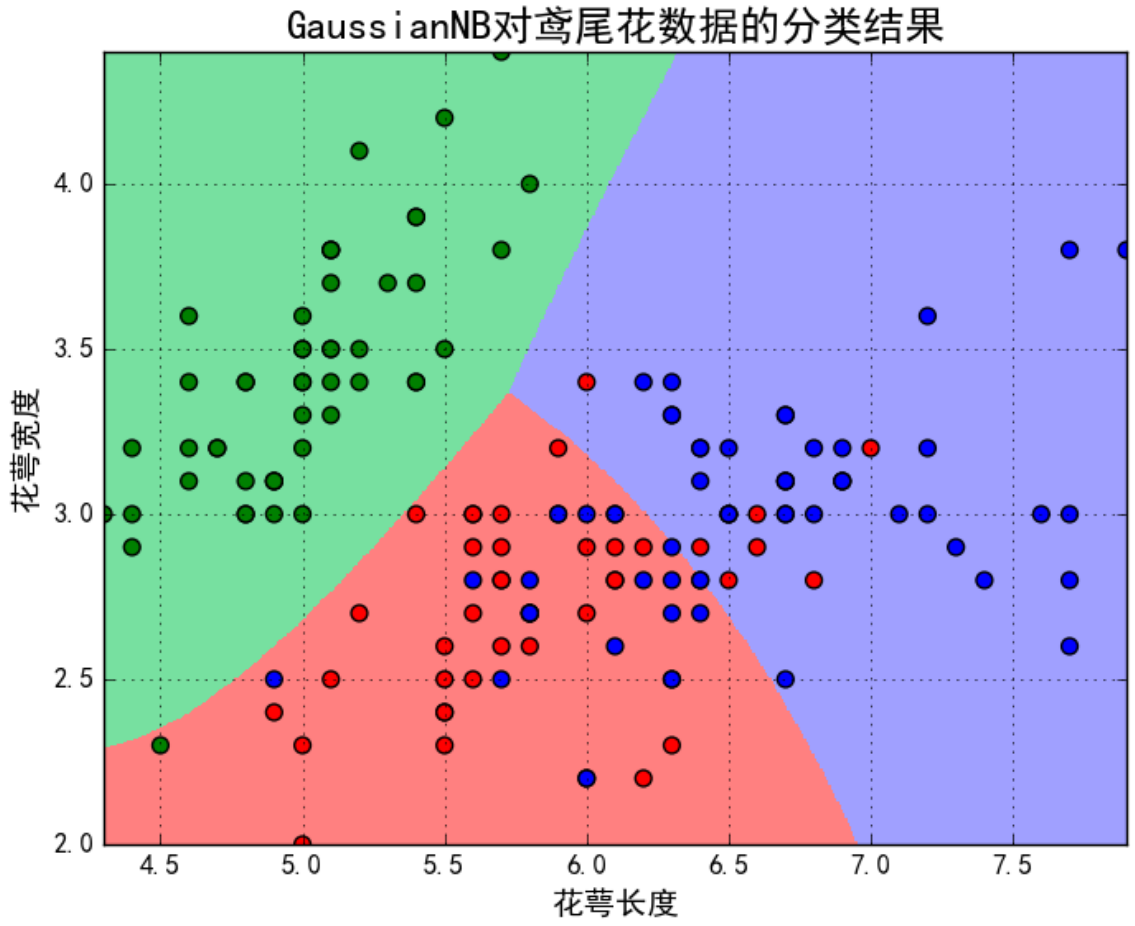
贝叶斯网络实践

邹博

主要内容

- 朴素贝叶斯的推导和应用
- 使用马尔科夫模型计算临近点概率
- 文本数据的处理流程
- 使用TF-IDF得到文本特征
- Word2vec的使用

GaussianNB


GaussianNB对鸢尾花数据的分类结果
花萼宽度
花萼长度
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5

GaussianNB / MultinomialNB

```
np.random.seed(0)
M = 20
N = 5
x = np.random.randint(2, size=(M, N))     # [low, high
x = np.array(list(set([tuple(t) for t in x])))
M = len(x)
y = np.arange(M)
print '样本个数：%d，特征数目：%d' % x.shape
print '样本：\n', x
mnb = MultinomialNB(alpha=1)    # 动手：换成GaussianNB(
mnb.fit(x, y)
y_hat = mnb.predict(x)
print '预测类别：', y_hat
print '准确率：%.2f%%' % (100*np.mean(y_hat == y))
print '系统得分：', mnb.score(x, y)
```

20.1.Iris_GaussianNB | 20.2.MultinomialNB_intro | 20.3.text_classification

```
[0 0 0 0 1]
[1 0 0 1 0]
[1 1 1 1 1]
[0 1 1 1 1]
[1 1 0 0 0]]
预测类别： [ 0  1  0  3  4  5  6  7  8  9 10 11 12 13  2 15 16]
准确率：88.24%
系统得分： 0.882352941176
2 :    [0 0 0 0 0] 被认为与 [1 1 0 1 0] 一个类别
14 :    [1 1 1 1 1] 被认为与 [0 0 0 0 0] 一个类别
```


小象学院 ChinaHadoop.cn

朴素贝叶斯的假设

- 一个特征出现的概率，与其他特征(条件)独立(特征独立性)
  - 其实是：对于给定分类的条件下，特征独立
- 每个特征同等重要(特征均衡性)

朴素贝叶斯的推导

- 朴素贝叶斯(Naive Bayes，NB)是基于“特征之间是独立的”这一朴素假设，应用贝叶斯定理的监督学习算法。
- 对于给定的特征向量 $x_1, x_2, \cdots, x_n$
- 类别y的概率可以根据贝叶斯公式得到：

$$P(y \mid x_1, x_2, \cdots, x_n) = \frac{P(y)P(x_1, x_2, \cdots, x_n \mid y)}{P(x_1, x_2, \cdots, x_n)}$$

朴素贝叶斯的推导

- 使用朴素的独立性假设：

$$P(x_i \mid y, x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) = P(x_i \mid y)$$

- 类别y的概率可简化为：

$$P(y \mid x_1, x_2, \cdots, x_n) = \frac{P(y)P(x_1, x_2, \cdots, x_n \mid y)}{P(x_1, x_2, \cdots, x_n)} = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, x_2, \cdots, x_n)}$$

- 在给定样本的前提下，$P(x_1, x_2, \cdots, x_n)$ 是常数：

$$P(y \mid x_1, x_2, \cdots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

- 从而：

$$\hat{y} = \arg\max_{y} P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

高斯朴素贝叶斯Gaussian Naive Bayes

- 根据样本使用MAP(Maximum A Posteriori)估计$P(y)$，建立合理的模型估计$P(x_i \mid y)$，从而得到样本的类别。

$$\hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

- 假设特征服从高斯分布，即：

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi}\sigma_y}\exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

  - 参数使用MLE估计即可。

多项分布朴素贝叶斯Multinomial Naive Bayes

- 假设特征服从多项分布，从而，对于每个类别y，参数为$\theta_y = (\theta_{y1}, \theta_{y2}, \cdots, \theta_{yn})$，其中n为特征的数目，$P(x_i \mid y)$的概率为$\theta_{yi}$。
- 参数θ_y使用MLE估计的结果为：$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha \cdot n}$，$\alpha \geq 0$
- 假定训练集为T，有：$\begin{cases} N_{yi} = \sum_{x \in T} x_i \\ N_y = \sum_{i=1}^{|T|} N_{yi} \end{cases}$
- 其中，
  - $\alpha = 1$称为Laplace平滑，
  - $\alpha < 1$称为Lidstone平滑。

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

以文本分类为例

- 样本：1000封邮件，每个邮件被标记为垃圾邮件或者非垃圾邮件
- 分类目标：给定第1001封邮件，确定它是垃圾邮件还是非垃圾邮件
- 方法：朴素贝叶斯

分析

- 类别c：垃圾邮件c_1，非垃圾邮件c_2
- 词汇表，两种建立方法：
  - 使用现成的单词词典；
  - 将所有邮件中出现的单词都统计出来，得到词典。
  - 记单词数目为N
- 将每个邮件m映射成维度为N的向量**x**
  - 若单词w_i在邮件m中出现过，则$x_i=1$，否则，$x_i=0$。即邮件的向量化：$m \rightarrow (x_1,x_2.....x_N)$
- 贝叶斯公式：$P(c|\mathbf{x})=P(\mathbf{x}|c)*P(c) / P(\mathbf{x})$
  - $P(c_1|\mathbf{x})=P(\mathbf{x}|c_1)*P(c_1) / P(\mathbf{x})$
  - $P(c_2|\mathbf{x})=P(\mathbf{x}|c_2)*P(c_2) / P(\mathbf{x})$
    - 注意这里**x**是向量

分解

- $P(c|\mathbf{x})=P(\mathbf{x}|c)*P(c) / P(\mathbf{x})$
- $P(\mathbf{x}|c)=P(x_1,x_2\ldots x_N|c)=P(x_1|c)*P(x_2|c)\ldots P(x_N|c)$
  - 特征条件独立假设
- $P(\mathbf{x})=P(x_1,x_2\ldots x_N)=P(x_1)*P(x_2)\ldots P(x_N)$
  - 特征独立假设
- 带入公式： $P(c|\mathbf{x})=P(\mathbf{x}|c)*P(c) / P(\mathbf{x})$

- 等式右侧各项的含义：
  - $P(x_i|c_j)$：在c_j(此题目，cj要么为垃圾邮件1，要么为非垃圾邮件2)的前提下，第i个单词x_i出现的概率
  - $P(x_i)$：在所有样本中，单词x_i出现的概率
  - $P(c_j)$：在所有样本中，邮件类别c_j出现的概率

拉普拉斯平滑

- $p(x_1|c_1)$是指的:在垃圾邮件c_1这个类别中，单词x1出现的概率。
  - x_1是待考察的邮件中的某个单词
- 定义符号
  - n_1：在所有垃圾邮件中单词x_1出现的次数。如果x_1没有出现过，则$n_1=0$。
  - n：属于c_1类的所有文档的出现过的单词总数目。
- 得到公式：$$p(x_1|c_1)=\frac{n_1}{n}$$
- 拉普拉斯平滑：$$p(x_1|c_1)=\frac{n_1+1}{n+N}$$
  - 其中，N是所有单词的数目。修正分母是为了保证概率和为1
- 同理，以同样的平滑方案处理$p(x_1)$

对朴素贝叶斯的思考

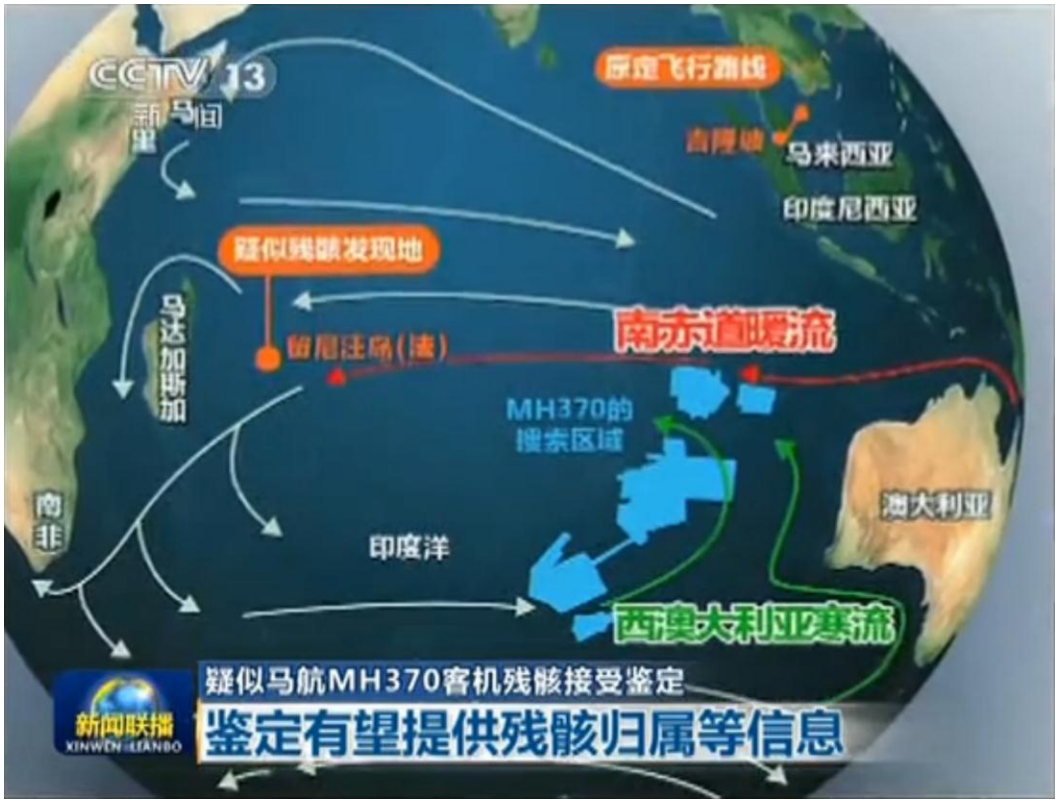
- 拉普拉斯平滑能够避免0/0带来的算法异常
- 要比较的是P(c1|x)和P(c2|x) 的相对大小，而根据公式P(c|x) =P(x|c)*P(c) / P(x)，二者的分母都是除以P(x)，实践时可以不计算该系数。
- 编程的限制：小数乘积下溢出怎么办？
- 问题：一个词在样本中出现多次，和一个词在样本中出现一次，形成的词向量相同
  - 由0/1向量改成频数向量或TF-IDF向量
- 如何判断两个文档的距离
  - 夹角余弦
- 如何给定合适的超参数 $\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha \cdot n}, \quad \alpha \geq 0$
  - 交叉验证

寻找马航MH370

- 2014年3月8日，马来西亚航空公司MH370航班(波音777-200ER)客机凌晨0:41分从吉隆坡飞往北京；凌晨1:19分，马航MH370与空管失去联系。凌晨2:14分飞机最后一次出现在军事雷达上之后人间消失。
- 2015年7月29日在法属留尼汪岛(l'île de la Reunion)发现襟副翼残骸；2015年8月6日，马来西亚宣布，该残骸确属马航MH370。随后法国谨慎宣布，“有很强的理由推测认为，...残骸属于马航MH370航班的波音777客机...但最终的比对结果还需要进一步的技术验证加以确认。”

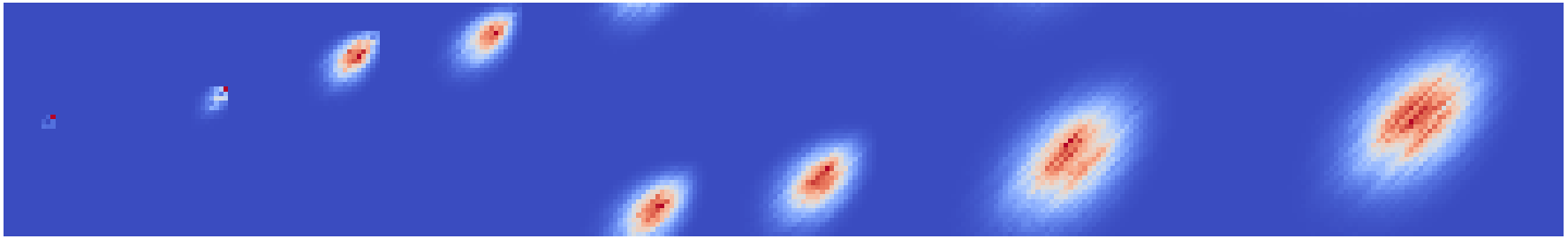
MH370最后消失区域

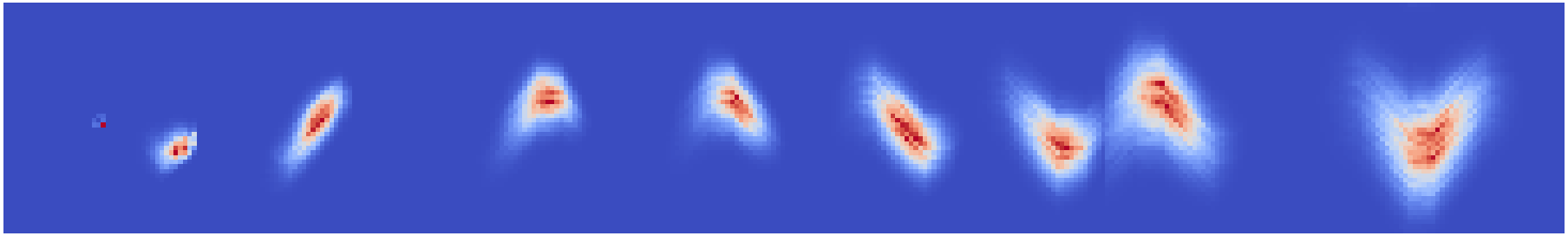
- 可否根据雷达最后消失区域和洋流、大气等因素；
- 判断留尼汪岛是否位于可能区域？
- 残骸漂流到该岛屿的概率有多大？

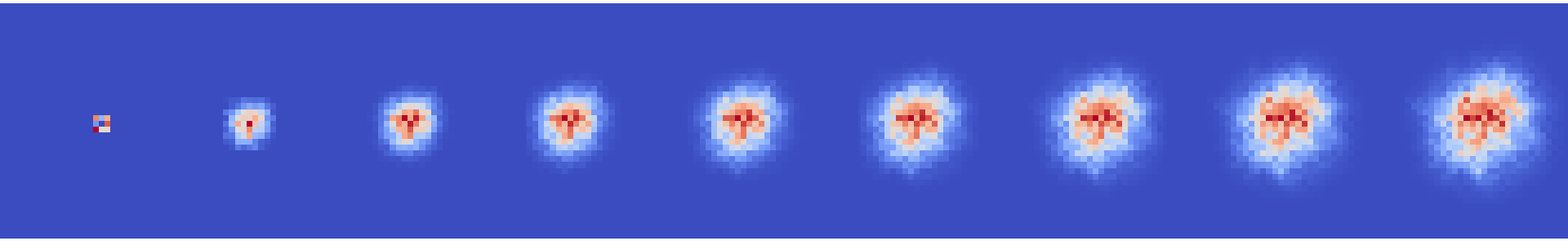


马尔科夫模型模拟实验

- 概率优势方向：Direct/Sin/Random







Code

```
save_image = True
style = 'Direct'    # Sin/Direct/Random
m, n = 50, 100
directions = np.random.rand(m, n, 8)

if style == 'Direct':
    directions[:,:,1] = 10
elif style == 'Sin':
    x = np.arange(n)
    y_d = np.cos(6*np.pi*x/n)
    theta = np.empty_like(x, dtype=np.int)
    theta[y_d > 0.5] = 1
    theta[~(y_d > 0.5) & (y_d > -0.5)] = 0
    theta[~(y_d > -0.5)] = 7
    directions[:, x.astype(np.int), theta] = 10
for i in np.arange(m):
    for j in np.arange(n):
            directions[i, j] /= np.sum(directions[i, j])
print directions

loc = np.zeros((m, n), dtype=np.float)
loc[m/2, n/2] = 1
loc_prime = np.empty_like(loc)
loc_prime = loc
fig = plt.figure(figsize=(8, 6), facecolor='w')
im = plt.imshow(loc/np.max(loc), cmap='coolwarm')
anim = animation.FuncAnimation(fig, update, frames=1000, interval=50, blit=True)
plt.tight_layout(1.5)
plt.show()
```

总结

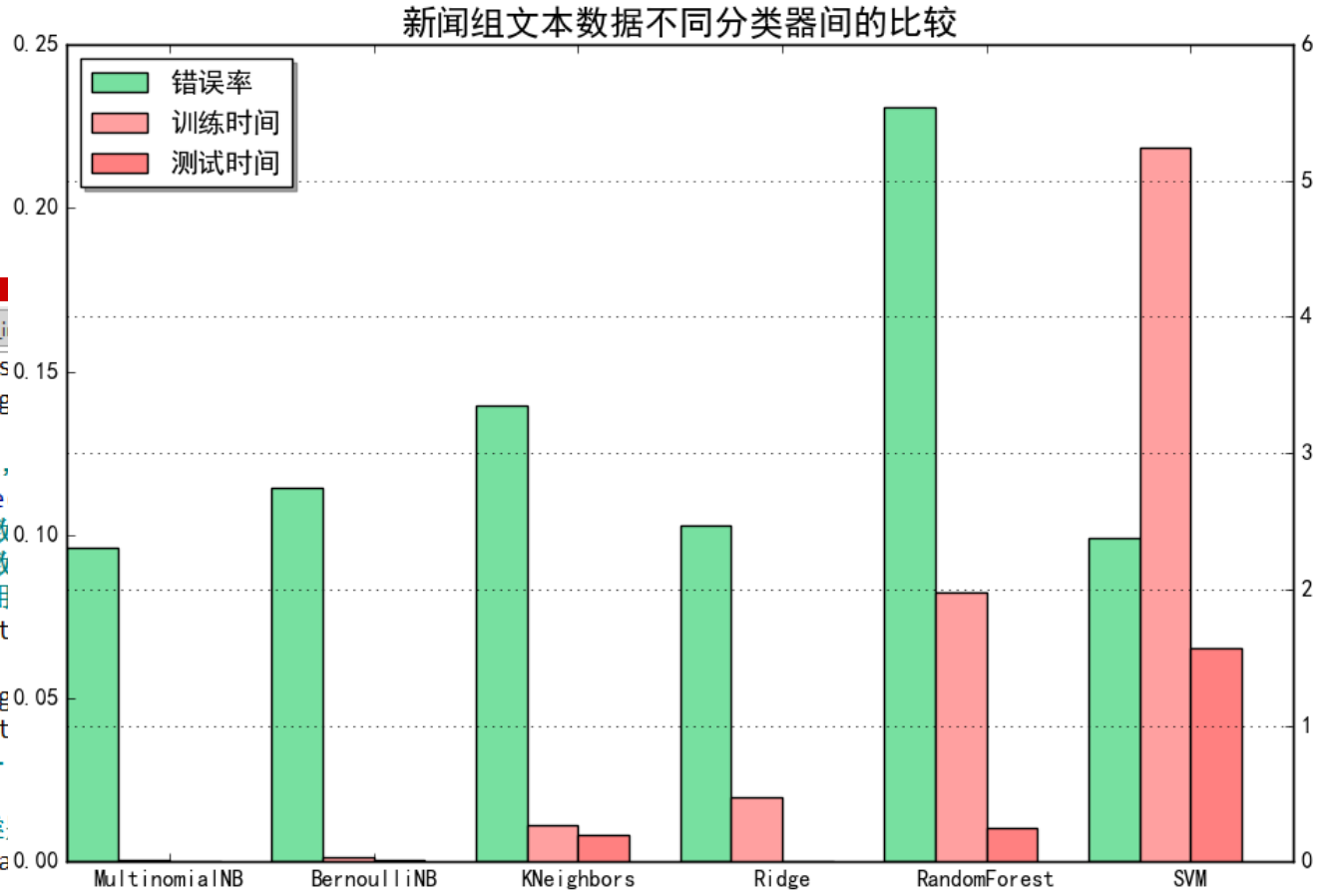
- 在每个时刻，物体的当前可能区域是上一时刻所有可能区域和相应转移概率的乘积和，这恰好是矩阵乘法(矩阵和向量乘法)的定义。
- 当前可能区域只和上一个时刻的区域有关，而与更上一个时刻无关，因此，是马尔科夫模型。
- 思考：可以使用“漂流位置”建立马尔科夫模型，该可能位置是不可观察的，而将“转移位置”认为是“漂流位置”的转换结果，“转移位置”是残骸的最终真实位置，使用增强的隐马尔科夫模型。
  - 不要过得累加模型的复杂度，适时使用奥卡姆剃刀(Occam‘s Razor)。
  - 该模型仅个人观点。

文本分类实验

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

- 实验数据：新闻组中的20个类别，原始文本数目约两万个，根据新闻组中文本的时间前后，划分成训练集(60%)和测试集(40%)。
  - 该数据最初应该是Ken Lang搜集整理。
- 数据获取：
  - 可使用sklearn.datasets.fetch_20newsgroups获取原始文本
  - 或者使用sklearn.datasets.fetch_20newsgroups_vectorized返回文本向量
- 该原始数据可以在该网页完整下载：
  - http://qwone.com/~jason/20Newsgroups/
    - 该课程的配套数据中已经包含该原始数据。

实验结果



20.1.Iris_GaussianNB.py × 20.2.MultinomialNB_i

```
    data_train = fetch_20news
    data_test = fetch_20newsg
    t_end = time()
    print u'下载/加载数据完成,
    print u'数据类型: ', type
    print u'训练集包含的文本数
    print u'测试集包含的文本数
    print u'训练集和测试集使用
    categories = data_train.t
    pprint(categories)
    y_train = data_train.targ
    y_test = data_test.target
    print u' -- 前10个文本 --
    for i in np.arange(10):
        print u'文本%d(属于类
        print data_train.data
        print '\n\n'
vectorizer = TfidfVectorizer(input='content', stop_words='english'
x_train = vectorizer.fit_transform(data_train.data)  # x_train是稀
x_test = vectorizer.transform(data_test.data)
print u'训练集样本个数: %d, 特征个数: %d' % x_train.shape
print u'停止词:\n',
pprint(vectorizer.get_stop_words())
feature_names = np.asarray(vectorizer.get_feature_names())

print u'\n\n====================\n分类器的比较: \n'
clfs = (MultinomialNB(),                    # 0.87(0.017), 0.002, 90.3
        BernoulliNB(),                      # 1.592(0.032), 0.010, 88.
        KNeighborsClassifier(),             # 19.737(0.282), 0.208, 86
        RidgeClassifier(),                  # 25.6(0.512), 0.003, 89.7
        RandomForestClassifier(n_estimators=200),   # 59.319(1.977
```

小象学院 ChinaHadoop.cn

word2vec

- Word2vec本质是建立了3层神经网络，将所有词都映射为一定长度(如200)的向量；取一定的窗口范围作为当前词的邻域，估计窗口内的词。
  - 词潜入
- 实验中使用2015年5月爬取的网页新闻作为输入文本，使用gensim的Word2vec包训练词向量。

词典

词典中词的个数： 19542

石块 办公会 基建 最后 着眼于 土 商户 郑重 组织纪律 沿街 岛内 基廷 周靖 工地 圈 律师 极乐 农村土地 期望 广东 央广 人身权利 圆 亲情 集群 眼角 给出 男子 合理 难看
事 城市 伊拉克 亏 于 云 升空 井 区长 为期 些 亚 千克 认定书 青年报 雅安 亡 考察 农业部 未曾 赞同 交 直至 亦 营房 打动 亭 京 亮 检 默契 人畜 亲 中央政治局常委
这时 灵璧县 外人 前往 修房 运动会 中纪委 烂尾 更改 延庆 收购 租房 政商 外事 申请 误工费 建议 租户 主战 宪政 抽屉 张嘉伟 开工 首席 入院 夺下 力争 出卖 场 爆 唐
安全事故 明确 姐姐 解决方案 雁过拔毛 前方 战斗机 心思 工作证 国民政府 陈水扁 铁轨 浅 有权 购物中心 覆盖面 地势 停诊 热潮 侵占 心疼 因素 比哈尔邦 响起 身患 开启
律师函 行贿 寄出 冷战 区位 宣判 拽 不至于 莱西 分析 形象 遮拦 是因为 剖析 药房 顿时 穿过 过半 录取 两难 秀屿 越发 客户端 巡视 拉票 鄱阳县 PAKDA 独家 玉皇 埃塞俄
丢弃 斐然 境界 反馈 净利润 政权 风格 文明 最起码 常会 情节 事端 经查 谈及 生命 涉外 恨不得 民事裁定 独立王国 违法 美发 战争 配 文工团 票据 A 素 党纪国法 版本
放火 运走 儿媳妇 劳埃德 实验室 相应 行政长官 求医 此番 有 领土问题 负伤 广度 候选人 牵动 最 校长 不够 疾病 望 不备 朝 期 总政治部 配置 累 人行道 朗 打斗 木 学院
业态 魏先生 这种 监禁 角度 王明 口罩 一致 黄磨 学文 贩毒 高原 企图 炮塔 捍卫 过世 鼓吹 咱们 巡逻机 产 货物 攻坚 APP 车位 郑绍鑫 读卖新闻 啤酒瓶 其后 秦玉海 松井
北极圈 十几年来 张北县 纽约市 北京电视台 美国共和党 砖 要出 动员 boyangcongpeople 砍 每天晚上 大军 亲朋 想见 平方米 荷兰 消防员 名头 笃信 先后 生计 填报 蛋白质
供认不讳 管窥 阵营 坤叔 房屋 储备 川 组织部 收集 屈某 校外 重新启动 策略 十月 屋外 差 外包 姚某 世界各地 左 房山 指导 郝纯毅 伟大 视为 已 风景区 并行 巴 铁道
纠葛 起点 汽油 否认 农业局 裁定 愿望 集团军 是从 重工 北京市公安局 工作作风 立法 非正式 点儿 海报 协和医院 咳嗽 打电话 乐队 此举 全局 姑息 日台 男童 拥堵 勘察
发到 法律文书 论述 老屋 战胜 马克思主义 逝者 编译 坦率地 遇上 体外 无期 西湖 带动 冲毁 产物 委派 上当受骗 去世 此后 大厂县 中国佛教协会 保护者 提取 互信 可怕
二维码 吴艾莉 施工方 选举 张通荣 哪位 分流 人事变动 想出 出发 冈比亚 面前 二甲基 性行为 厂房 军长 陶瓷 应有 储蓄 申冤 土耳其政府 签字 守候 马伊 成分 学业 边检
带队 放在 冷孟梅 爱心 袭击 韩某 定案 不明真相 预定 宋先生 办公厅 术语 会长 鱼鹰 背后 向永林 雨量 防微杜渐 创始人 王庄 以往 康定 news 衢州市 二妹 交通管理 跳车
九州 网讯 焦兰生 其中 油漆 师范大学 问起 买不起 前两天 高洪顺 成都市委 失事 造势 其一 高尔夫球 信息中心 开区 平时 拒收 其三 国门 隔壁 商务通 人大常委会 资产
亲戚朋友 安全监管 应该 我行我素 非要 其次 总监 藏品 晏涛 庄河市 咖啡 新西兰 毛发 分开 遗留问题 控股 治病 经济社会 官僚 深刻 纵容 汉口 进一步 真经 力气 视频 脱团
湖南省委 购 眼线 深切 木材 留给 处警 大字 初步统计 增长极 自强 大概 砍掉 确为 轰炸 商家 丁学君 原因 近些年 等级 被害人 打仗 心怀 遗物 协调一致 奶 反对者 木板
负责 适用 外来 数以亿计 清江 更何况 桥 落到实处 大方 军委 县市区 字词 案 年纪 机翼 阐明 桌 桂 龙岩 检方 受到限制 层级 女 法定程序 父母 洪金洲 陈 节点 该处 何等
克制 二环路 碎 生病 摄制组 电梯门 时会 碗 测量 知情人 打开 主管 成人 碘 威廉 时说 碧 联合国大会 沙子 慰问 远远 火爆 夺走 激情 性别 切实 击碎 火箭 樊某 碰 区县
站不住脚 毛利率 进而 数学 幸福感 环保部门 五月 十日 冰块 遗址 股份制 远东地区 数字 进军 小山头 薛凤 捡起 Facebook 两学 开善 产出 科长 护栏 昨夜 免税 今日 医健
夏伯渝 已经 图案 就医 可谓 原判 王纪平 整理 奈何 下跌 阅读 亲友 脏器 暴涨 三叉戟 考上 包庇 不懈努力 脱口秀 难料 鼓楼 金融时报 空气质量 救灾 积极主动 萨吉姆
礼仪 确信 斗鱼 停在 发愁 手中 弊端 不雅 承接 美国大学 底特律 商业秘密 战线 砖混 道德 艰苦奋斗 月入 取暖 放出 合法权利 癌变 医疗事故 此人 德州 产科 手上 手下
爱情 所知 相对来说 患 光环 副作用 恶劣 悦 之一 工程学院 安抚 您 之上 贡献力量 悬 建国 严厉 E 振源 宣泄 巨大损失 发帖人 天津 华夏 痛恨 放管 民事 小心 复决 三板
美国国务院 锅 锁 偷盗 补发 不出 回来 半岛 随身 刑警大队 残留量 钻营 用心 应酬 医疗保险 键 盎司 纪念日 蹲点 不减 站上 更多 饮片 王强 卡号 不准 潜入 刘建军 普通人
集中统一 张贴 假警察 质量 真诚 贷款 事前 山洪 研究生 甲乙双方 刑警 高手 韩军 敲诈勒索罪 仓库 普查 背上 标注 鹿邑县 臭味 随身携带 俄联邦 医用 朱明国 擒获 发觉
外交部 皇冠 形成 桂江 尘埃落定 有害信息 分拆 墨西哥 怀上 脱逃 益康 再三 脑子 领跑 堂而皇之 永久性 不对劲 读者 漂亮 主治医师 MA 酒席 细胞 姑娘 灵 成绩 伤残 灰 灯
热度 白 纯粹 车门 上面 车间 嘴 机长 市场机制 阿里 欧内斯特 元素 公职人员 巴西利亚 国务院办公厅 欠下 难忍 夜晚 缓和 前兆 这里 规划图 酣睡 发育 集约 巨额 海宁
表格 三死一伤 爱浜 操办 招募 卢某诚 刘会文 警戒线 友善 年迈 算账 药业 约见 名录 恒频 长春 刘峻成 相距 二氧化硫 满脸 通了 卡通 更是 的的 会面 填 上来 布满 外侧
干部队伍 准确 排放 那种 根源 辐射 党性原则 农副产品 有违 将军 幕府 紧锁 行列 桥梁 成功 勇气 大雨 推动 杂物 华裔 推演 通知书 物种 江苏队 电警棍 郊区 核准 通畅

训练结果

中国 和 祖国 的相似度为：0.532070
中国 和 毛泽东 的相似度为：0.312460
中国 和 人民 的相似度为：0.371986
祖国 和 毛泽东 的相似度为：0.882594
祖国 和 人民 的相似度为：0.759524
毛泽东 和 人民 的相似度为：0.583692

苹果 三星 美的 海尔 离群词： 海尔
中国 日本 韩国 美国 北京 离群词： 北京
医院 手术 护士 医生 感染 福利 离群词： 福利
爸爸 妈妈 舅舅 爷爷 叔叔 阿姨 老婆 离群词： 舅舅

中国 + 城市 - 学生 ：
经济 0.842732787132
国际 0.842595815659
战略 0.839224457741
发展 0.816590726376
我国 0.792158007622
全球 0.782959520817
推动 0.781142055988
军事 0.778445780277
台海 0.776162207127
税制 0.7729408741
男 + 工作 - 女 ：
两学 0.74424546957
做好 0.723003566265
督察 0.708856344223
党校 0.68696641922
电视电话会议 0.679121792316
各级 0.677579462528
明确提出 0.668009757996
中央 0.664826393127
充分认识 0.661263227463
谈话 0.659905314445
俄罗斯 + 美国 + 英国 - 日本 ：
法国 0.944284677505
安倍 0.94345831871
菲律宾 0.923454284668
俄 0.923405408859
朝鲜 0.921864390373
普京 0.919387102127
敌对 0.918260753155
大选 0.915652871132
国民党 0.911968111992
主流 0.911087334156

与 学习 最相近的词：
结构性 0.948330461979
侧 0.94113522768
创新 0.939934253693
树立 0.939579963684
注重 0.933748006821
发挥 0.93070846796
精益求精 0.925910234451
人才 0.924418389797
供给 0.923088252544
党校 0.92102253437
与 公安局 最相近的词：
分局 0.98934006691
大队 0.975338459015
支队 0.963695168495
鹿邑县 0.961336493492
西城 0.95579791069
刑警 0.955546677113
公安分局 0.953128516674
隆化县 0.952146053314
控申 0.951352596283
专案组 0.950993955135
与 大学 最相近的词：
秦川 0.889399170876
中央党校 0.88586807251
中科院 0.856984019279
学历 0.856915056705
吉安 0.84980738163
研究生 0.847551643848
优秀 0.846781611443

我们在这里

- http://wenda.ChinaHadoop.cn
  - 视频/课程/社区
- 微博
  - @ChinaHadoop
  - @邹博_机器学习
- 微信公众号
  - 小象
  - 大数据分析挖掘

感谢大家！

恳请大家批评指正！