

程序化购买系列-C3算法基础

@上海李立峰 2015.10.15

目录

- 信息检索
 - 倒排索引
 - 向量空间模型
- 机器学习
 - 贝叶斯
 - 最大熵与指数族分布
 - 混合模型与EM算法
- 最优化方法
 - 梯度下降法
 - 拉格朗日法
 - 拟牛顿法
 - Trust-Region法

信息检索-倒排索引

- 搜索引擎核心算法之一
- 核心目的是将从大量文档中查找包含某些词的文档集合这一任务用 $O(1)$ 或 $O(\log n)$ 的时间复杂度完成，其中 n 为索引中的文档数目。
- 也常称为反向索引、置入档案或反向档案，用来存储在全文搜索下某个单词在一个文档或者一组文档中的存储位置的映射。

信息检索-倒排索引-正排表

- pig dog cat dog

DocID	Word	HitCount	HitList
1	cat	1	3
1	dog	2	2,2
1	pig	1	1

- DocID: 文档ID
- WordID: 单词ID, 字典序
- HitCount: 命中次数
- HitList: 第一个数字表示单词在文档中的绝对位置, 后面的数字表示新位置与前一位置的距离

信息检索-倒排索引-倒排表

- Doc1: pig dog cat dog

Word	LocInfo
cat	0x0001
dog	0x0002
pig	0x0003

DocID	HitCount	HitList
1	1	3

DocID	HitCount	HitList
1	2	2,2

DocID	HitCount	HitList
1	1	1

信息检索-倒排索引-倒排表

- Doc1: pig dog cat dog
- Doc2: dog cat dog fish

Word	LocInfo
cat	0x0001
dog	0x0002
Fish	0x0004
pig	0x0003

DocID	HitCount	HitList
1	1	3
2	1	2

DocID	HitCount	HitList
1	2	2,2
2	2	1,2

DocID	HitCount	HitList
1	1	1

DocID	HitCount	HitList
2	1	4

信息检索-倒排索引-规模估算

- 齐普夫法则：第K个最经常出现的词，其词频与 $1/K$ 成正比
- 假设汉语词汇有50万个，根据齐普夫法则
有 $\sum_{i=1}^{500000} \frac{c}{i} = 1$ ，可以得到 $c=1/13$
- 假设每个文档平均词汇1000个，第i个最经常使用的词汇在文档中期望出现的次数为
 $1000 * 1/13i = 76i$
- 对于50万个词汇，则有 $500000/76=6500$ 词汇到文档的映射块，第i块出现在 n/i 个文档中

信息检索-倒排索引-规模估算

- 假设有100亿个平均词汇为1000的文档，则采取差分法存储这100亿个文档id至少需要10GB的空间

- 存储文档所需的空间估计为

$$\sum_{i=1}^{6500} 76 * 10GB * 1/i \approx 760GB * \ln(6500) = 6.7TB$$

- 加上HitList信息，大约需要10TB的空间
- 问题：这10TB的空间无法同时加载在内存，如何进行内存分配和优化检索？

信息检索-向量空间模型VSM

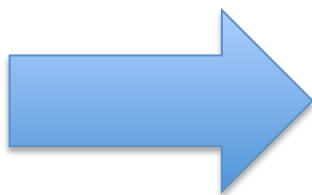
- **VSM**概念简单，把对文本内容的处理简化为向量空间中的向量运算，并且它以空间上的相似度表达语义的相似度，直观易懂。
- 当文档被表示为文档空间的向量，就可以通过计算向量之间的相似性来度量文档间的相似性。文本处理中最常用的相似性度量方式是余弦距离。

信息检索-向量空间模型-构建向量

- 每一篇文章用一个向量来表达，查询用一个向量来表达，通过向量运算来计算相似度
- 对于文档中每一个不同的词项，我们在向量中只记录一个分量
- 当词项出现几次，对应向量的分量就记几；未出现则记为0

信息检索-向量空间模型-构建向量

- D1: A, I
- D2: A, A
- D3: I
- Q: A, I



- D1: $\langle 1, 1 \rangle$
- D2: $\langle 2, 0 \rangle$
- D3: $\langle 0, 1 \rangle$
- Q: $\langle 1, 1 \rangle$

信息检索-向量空间模型-TF-IDF

- **TF（词频）**：某一个给定的词语在该文件中出现的频率。分子是该词在文件中的出现次数，而分母则是在文件中所有字词的出现次数之和
- **IDF（逆文档频率）**：一个词语普遍重要性的度量。某一特定词语的IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数

$$IDF_t = \log \frac{D}{DF_t}$$

$$IDF_t = \frac{D}{DF_t}$$

- **TF-IDF运算：**

$$TF - IDF_{t,d} = \frac{n_{t,d}}{n_d} \times \log \frac{D}{DF_t}$$

信息检索-向量空间模型-TF-IDF

- 湖畔的夏夜常常很凉爽， ...
- 湖畔有家“湖畔”啤酒花园，花园中常常是鼓鼓的蛙鸣一片， ...
- “蛙鸣”禅社举办“蛙鸣”诗会的消息...

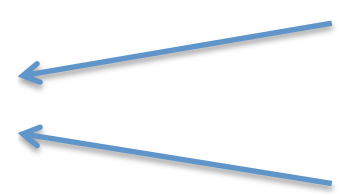
Term	...	湖畔	夏夜	的	常常	蛙鸣	禅社	诗会	...
<i>df</i>	...	2	1	3	2	2	1	1	...
<i>idf</i>	...	0.176	0.477	0	0.176	0.176	0.477	0.477	...

机器学习-贝叶斯-分类问题


- 分类问题
 - 医生治病
 - 已知集合： $C = \{y_1, y_2, \dots, y_n\}$ 和 $I = \{x_1, x_2, \dots, x_m, \dots\}$ 确定映射规则 $y = f(x)$ ，使得任意 $x_i \in I$ 有且仅有一个 $y_j = f(x_i)$ 使得 $y_j \in C$ 成立。
 - 其中 C 叫做类别集合，其中每一个元素是一个类别，而 I 叫做项集合，其中每一个元素是一个待分类项， f 叫做分类器。分类算法的任务就是构造分类器 f 。

机器学习-贝叶斯-贝叶斯定理

- 问题
 - 已知某条件概率，如何得到两个事件交换后的概率，也就是在已知 $P(A|B)$ 的情况下如何求得 $P(B|A)$
- 贝叶斯定理

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

$$P(A|B) = \frac{P(AB)}{P(B)}$$
$$P(B|A) = \frac{P(AB)}{P(A)}$$

机器学习-贝叶斯-朴素贝叶斯 (NB)

- 条件独立
- 朴素贝叶斯定义：
 1. 设 $x = \{a_1, a_2, \dots, a_m\}$ 为一个待分类项，而每个 a 为 x 的一个特征属性。
 2. 有类别集合 $C = \{y_1, y_2, \dots, y_n\}$ 。
 3. 计算 $P(y_1|x), P(y_2|x), \dots, P(y_n|x)$  $P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$
 4. 如果 $P(y_k|x) = \max\{P(y_1|x), P(y_2|x), \dots, P(y_n|x)\}$ ，则 $x \in y_k$ 。

机器学习-贝叶斯-朴素贝叶斯 (NB)

- 真假账号分类问题

- $C=0$ 表示真实账号， $C=1$ 表示不真实账号

- 特征属性

- a_1 : 日志数量/注册天数, $a_1: \{a \leq 0.05, 0.05 < a < 0.2, a \geq 0.2\}$

- a_2 : 好友数量/注册天数, $a_1: \{a \leq 0.1, 0.1 < a < 0.8, a \geq 0.8\}$

- a_3 : 是否使用真实头像, $a_3: \{a=0 \text{ (不是)}, a=1 \text{ (是)}\}$

- 训练样本

- 人工检测过的1万个账号

- 计算类别概率 $P(C=0) = 8900/100000 = 0.89$

$$P(C=1) = 110/100000 = 0.11$$

- 计算特征概率

$$P(0.05 < a_1 < 0.2 | C=0) = 0.5 \quad P(0.05 < a_1 < 0.2 | C=1) = 0.1$$

$$P(0.1 < a_2 < 0.8 | C=0) = 0.7 \quad P(0.1 < a_2 < 0.8 | C=1) = 0.2$$

$$P(a_3 = 0 | C=0) = 0.2 \quad P(a_3 = 0 | C=1) = 0.9$$

- 分类器鉴别

$$P(C=0)P(x|C=0) = P(C=0)P(0.05 < a_1 < 0.2 | C=0)P(0.1 < a_2 < 0.8 | C=0)P(a_3 = 0 | C=0) = 0.89 * 0.5 * 0.7 * 0.2 = 0.0623$$

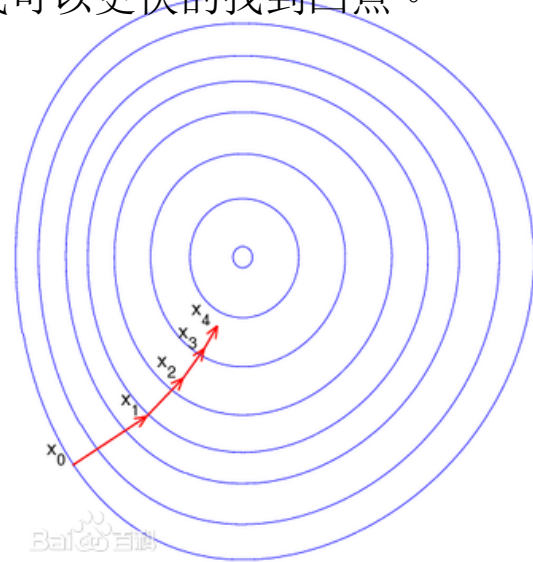
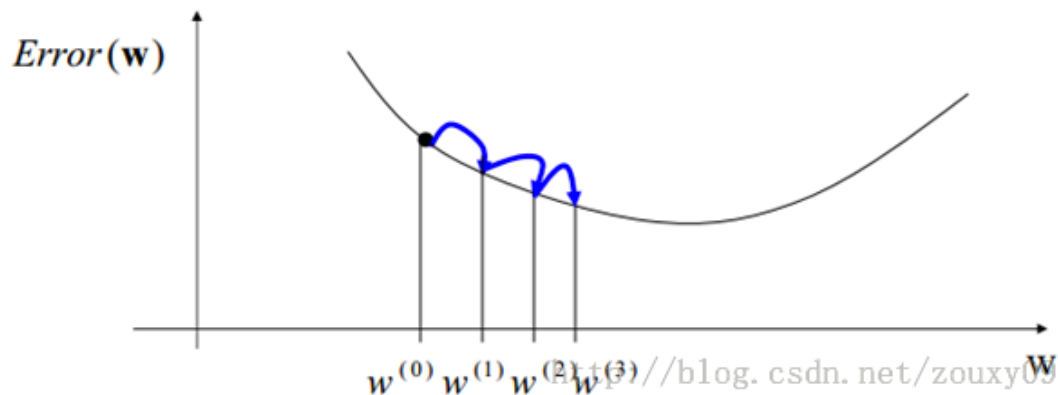
$$P(C=1)P(x|C=1) = P(C=1)P(0.05 < a_1 < 0.2 | C=1)P(0.1 < a_2 < 0.8 | C=1)P(a_3 = 0 | C=1) = 0.11 * 0.1 * 0.2 * 0.9 = 0.00198$$

最优化方法-最优化

- 最优化（**Optimization**），是应用数学的一个分支，主要研究以下形式的问题：
 - 给定一个函数，寻找一个元素满足A中的，取得最小化；或者最大化。
 - 有时被称为线性规划或运筹学
- 讨论内容：给定某个确定的目标函数以及该函数自变量的一些约束条件，求解该函数的最大或最小值的问题。
 - 无约束最优化问题：求解： $\min f(x)$
- 问题：如何使用min来求解max？

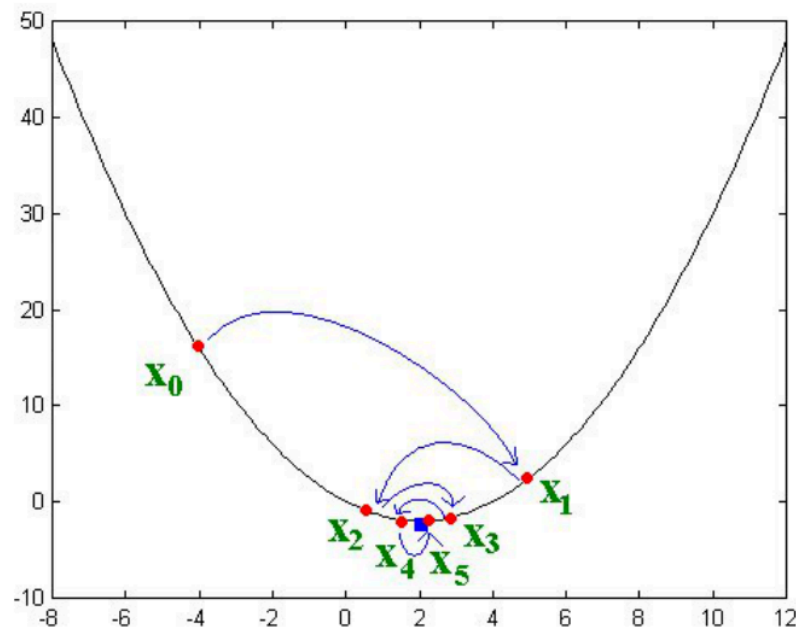
最优化方法-梯度下降法

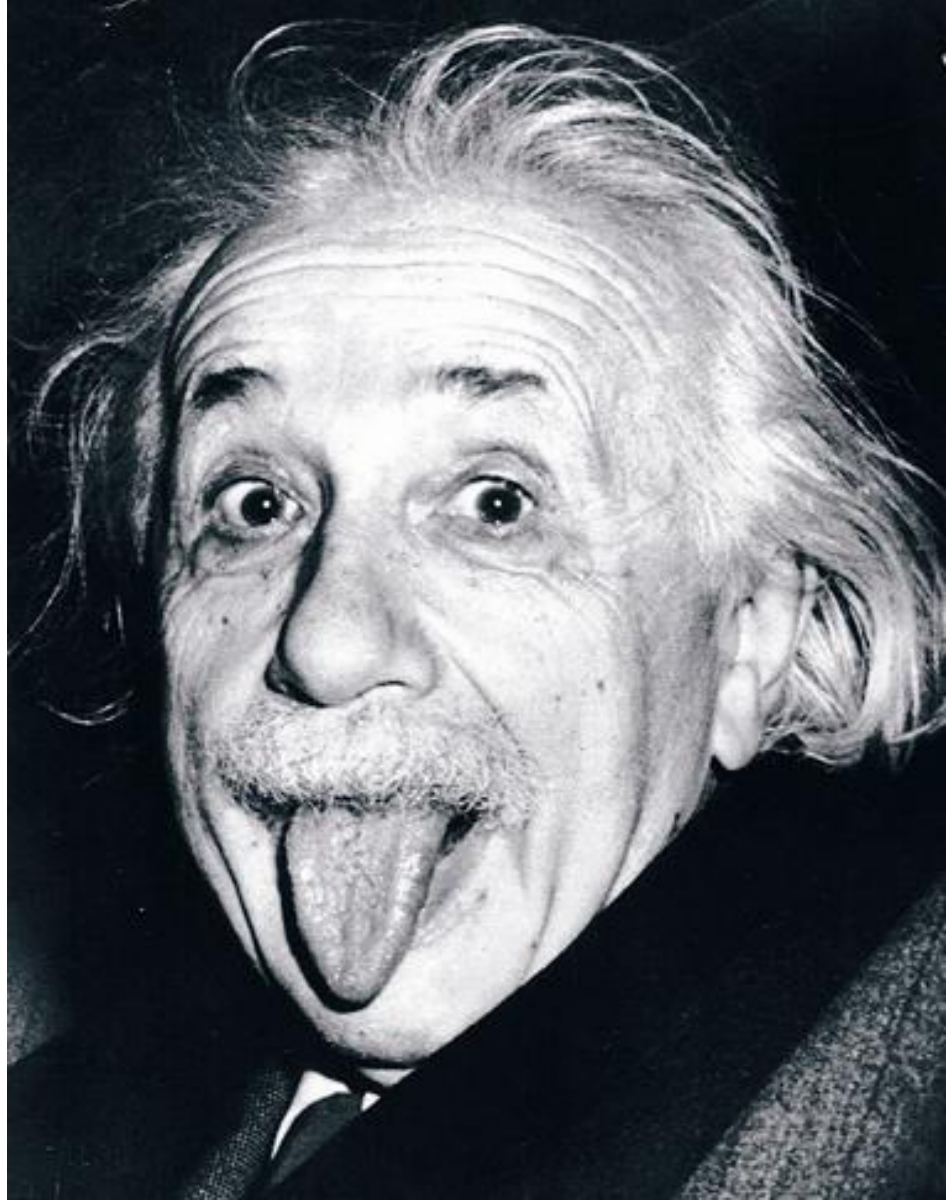
- 梯度下降(**gradient descent**)，是利用一阶的梯度信息找到函数局部最优解的一种方法，也是机器学习里面最简单最常用的一种优化方法。
 - 它的思想很简单，要找最小值，只需要每一步都找下坡最快的地方，也就是每一步走某个方向，都比走其他方向离最小值更近，那么我可以更快的找到凹点。



最优化方法-梯度下降法-举例

- 举例： $y = \frac{x^2}{2} - 2x$
- 计算过程
 - 任意给定一个初始点，设为 $x_0 = -4$
 - 首先给定两个参数： $\alpha = 1.5, \varepsilon = 0.01$
 - 计算导数（剃度）： $\frac{dy}{dx} = x - 2$
 - 计算当前导数： $y'_0 = x_0 - 2 = -6$
 - 计算新点： $x_1 = x_0 - \alpha \times y'_0 = -4 - 1.5 \times (-6) = 5.0$
 - 计算当前导数： $y'_1 = x_1 - 2 = 3.0$
 - 计算新点： $x_2 = x_1 - \alpha \times y'_1 = 5.0 - 1.5 \times (3.0) = 0.5$
 - 计算当前导数： $y'_2 = x_2 - 2 = -1.5$
 - 计算新点： $x_3 = x_2 - \alpha \times y'_2 = 0.5 - 1.5 \times (-1.5) = 2.75$
 - 计算当前导数： $y'_3 = x_3 - 2 = 0.75$
 - 计算新点： $x_4 = x_3 - \alpha \times y'_3 = 2.75 - 1.5 \times (0.75) = 1.625$
 - ...





福利-最优化约会策略

- 假设你征女朋友（男朋友），选定了某些约会对象，比如20个。假设：
 1. 可以将所有已约会对象按优劣排序，但无法得知他在所有人中的排名。在约会过程中，你知道某人是你目前已见到的最好的，但当时还不能确定是不是所有人里面最好的
 2. 如果你在约会当时决定放弃某人，后面再没有机会和好-好马不吃回头草
 3. 选定某人后约会结束-不再骑驴找马
 4. 排除一见钟情

福利-最优化约会策略

- 目标
 - 使得最大可能找到最喜欢的人
- 关系定得太早，会因为第3条假设-更精彩的还在后面
- 关系定得太晚，会因为第2条假设-而后悔莫及
- 什么策略能让你以最大概率找到你最满意的那个TA？ ？ ？

福利-最优化约会策略

- 一个简单的方法如下：
 - 待定 K ，与前 K 个人约会，不做任何选择。继续约会直到比这 K 个人还好的那个人为止
- 通过概率得出，这个方法比我们想象中要好得多，通过选取合适的 $K=7$ ，有接近40%的机会选中最好的那位，有70%的机会选中最好或次好的那位。

谢谢