中国好创意
垃圾短信基于文本内容识别
技术方案说明

lifematrix
stevenliucx@gmail.com

2015.12

Hash Trick

- **传统做法**
  - 需维护一个字典，将feature 映射为索引(或feature id)
  - 保留字典需耗费极大内存，最大feature数会受到内存限制
  - 多机需维护同一个字典，不利于并行
- Feature Hashing
  - 将Feature映射到一个有限的空间(2^n)
  - 不需字典
  - 只要hash算法统一，多机极易并行

Definition 1 *Denote by h a hash function $h : \mathbb{N} \to \{1, \ldots, m\}$. Moreover, denote by ξ a hash function $\xi : \mathbb{N} \to \{\pm 1\}$. Then for vectors $x, x' \in \ell_2$ we define the hashed feature map ϕ and the corresponding inner product as*

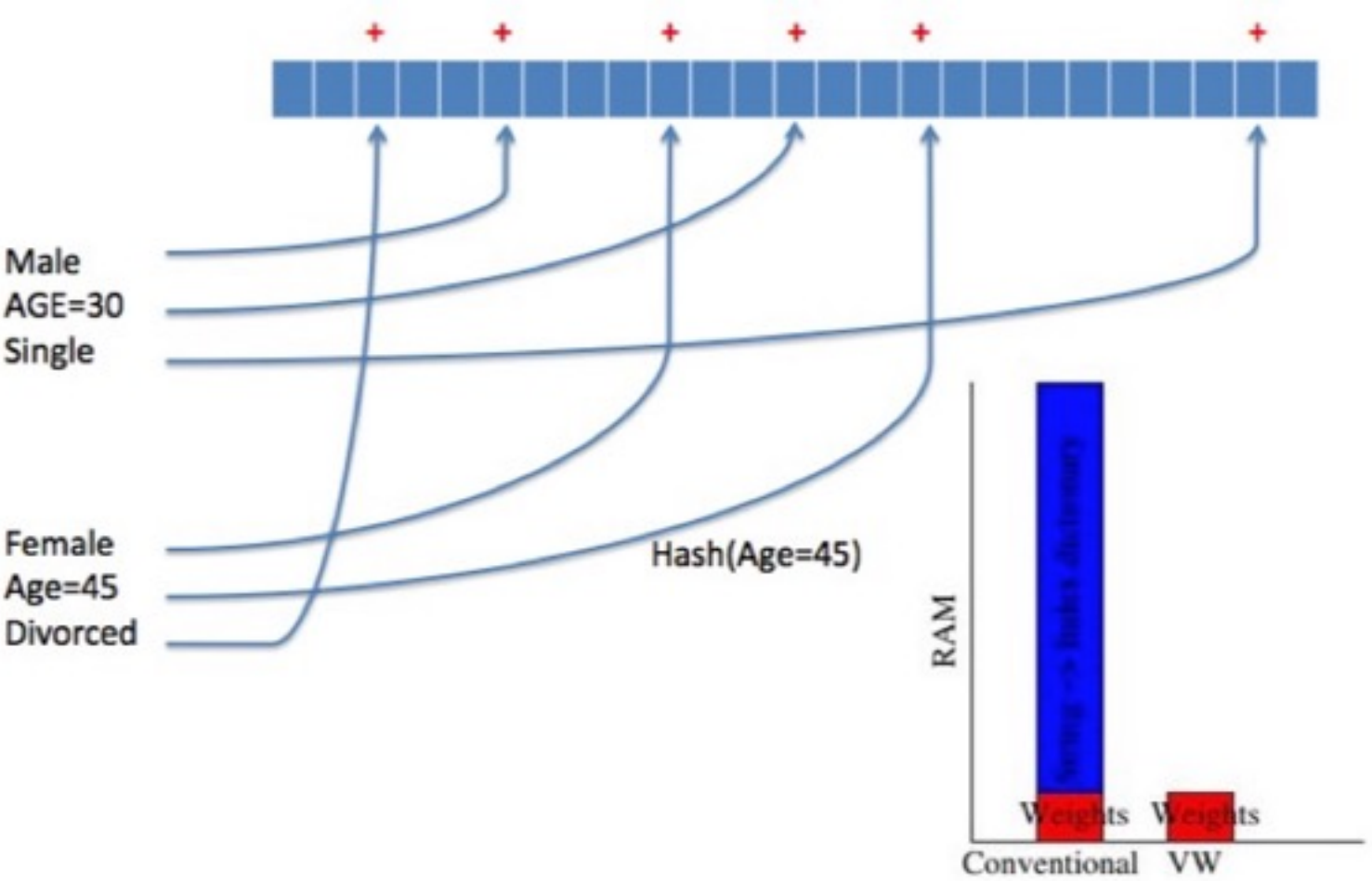
$$\phi_i^{(h,\xi)}(x) = \sum_{j:h(j)=i} \xi(i) x_i \tag{2}$$

$$\text{and } \langle x, x' \rangle_\phi := \left\langle \phi^{(h,\xi)}(x), \phi^{(h,\xi)}(x') \right\rangle. \tag{3}$$

Lemma 2 *The hash kernel is unbiased, that is $\mathbf{E}_\phi[\langle x, x' \rangle_\phi] = \langle x, x' \rangle$. Moreover, the variance is $\sigma^2_{x,x'} = \frac{1}{m}\left(\sum_{i \neq j} x_i^2 {x'_j}^2 + x_i x'_i x_j x'_j\right)$, and thus, for $\|x\|_2 = \|x'\|_2 = 1, \sigma^2_{x,x'} = O\left(\frac{1}{m}\right)$.*

Hash Trick

哈希表，2^n. n 为bits，最大32，支持4billon个特征

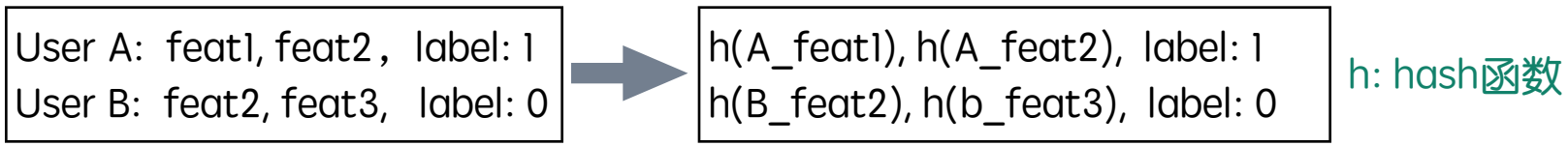


多过滤器的对应

- 在Spam filter中，如果不同的用户或群体有各自的过滤器。分别训练和维护不同的过滤器模型，开销和时间将会极大

- 通过Hash Trick将所有的feature映射同一空间，即可训练单一模型

训练实例



- 单一模型对应多任务，在email-spam中很常用
  - 虽然短信过滤中，多过滤器不甚突出，但也会存在。会是给用户的特色服务
  - Hash trick为短信过滤提供了极大的灵活性

分词后features字典规模

分词及ngram 2增加了feature数，弥补了feature的不足。
而ngram 3则会造成feature过多，增加了噪声

	字符数	分词后单词数	*ngram 2*
测试集	*8,757*	*379,302*	*3,747,171*
训练集	*7,173*	*17,326*	*1,277,956*
合集	*9,013*	*429,306*	*4,420,476*

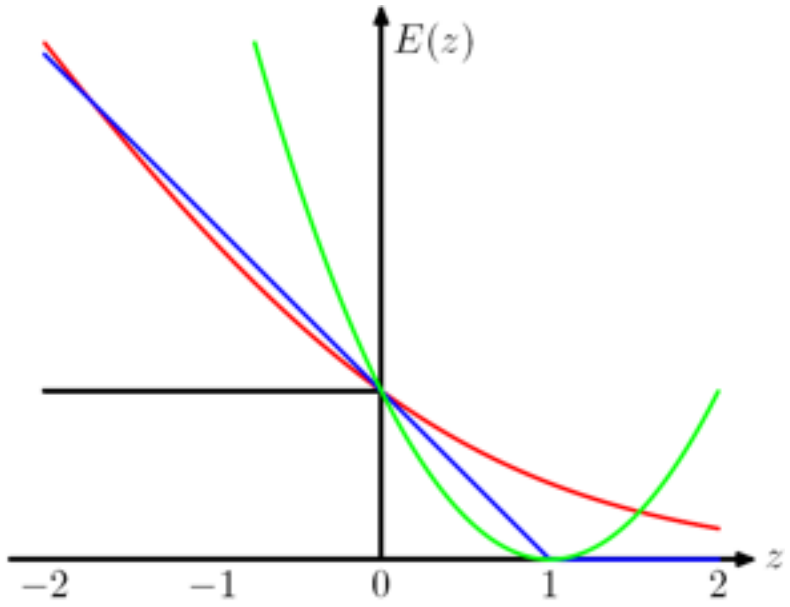
在线学习

- **传统的学习方法：批量学习(batch learning)**
  - 需要将整个训练集数据读入，进行训练。如果有新的实例，必须对新的训练集再次训练
- **在线学习**
  - 也是增量学习。当有新实例产生时，可以继续学习，优化模型
  - 不需将整个训练集读入内存，节约内存
  - 适合**“流式计算”范式**，可从socket/消息队列等不断读入新实例去训练
- **短信过滤的应用，很适合在线学习的方式**

损失函数

- Logistic
  - log(1 + exp(−yp)),
- Hinge
  - max(0, 1 − yp)
  - 采用hinge比logistic得分高0.1%
  - 文本分类，相对hinge更适合

y: 真值，p: 为预测，属于{0,1}



logistic 红色，Hinge 蓝色。Square loss 绿色。
0/1 error 黑色。z = yp

为什么选择Vowpal Wabbit

- **目前世界前沿的学习系统**
  - 曾用于Yahoo垃圾邮件过滤
  - 速度极快，很轻松处理几百万记录
  - Hash Trick技术，能够化解文本处理中的维度膨胀
    - 对于短信文本，如果不分词，feature 等于汉字的字符数(几万)。如果分词，feature 有几十万。但如果采用二元模型，feature理论上就会过亿。

- **工业级的学习系统**
  - 在线学习系统(online learning)
  - 支持1000+台机器的集群，Allreduce
    https://github.com/JohnLangford/vowpal_wabbit/wiki/Cluster_parallel.pdf

- **极具灵活性**
  - 多种损失函数、学习方法、参数…

性能

#	步骤	说明	耗时
1	转换测试集	80万条	3m48s
2	转换训练集	20万条	57s
4	训练模型	hash bits = 31	1m38s
5	预测、生成提交文件		8s

训练和测试不到2分钟

测试环境：Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz
Amazon EC2 Linux

- features转换花了较多时间，因为采用了通常的python写法，没有特别优化
- vowpal wabbit被称为机器学习的“瑞士军刀”，模型训练百万条记录，非常轻松
- 如果降低hash bits到24, 训练和预测模型，几乎瞬间完成，而且准确性只稍降低

谢 谢!