Feature Hashing Summary
特征哈希小结

VISION@OUC

目 录

- Hierarchical Feature Hashing (层级特征哈希)
- Feature Hashing (特征哈希)

  Brief Introduction（简 介)

  Theoretic Proof (理论证明)

  Application (应 用)
- Support Vector Machine 支持向量机

目 录

特征哈希——核心

在一个较短的时间内，把高维特征向量压缩成较低维特征向量，且尽量不损失原始特征的表达能力

——时间与准确率的优化问题

特征哈希——优点

- 实现简单，所需额外计算量小（时间）
- 准确率持平或者较高?
- 不同任务（或类别、图片）之间的相互干扰小
- 可以添加新任务（如新用户、新图片），或者新的原始特征而保持哈希转换后的特征长度不变，适合任务数频繁变化的问题
- 保持原始特征的稀疏性，哈希转换时，只有非0特征才起作用
- 可以只哈希一部分原始特征，而保留另一部分原始特征（如那些出现collision就会影响分类精度的）

1. K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature Hashing for Large Scale Multitask Learning. In *ICML*, *2009*. *2,3*

特征哈希——核心

在一个较短的时间内，把高维特征向量压缩成较低维特征向量，且尽量不损失原始特征的表达能力

——时间与准确率的优化问题

特征哈希——优点

- 实现简单，所需额外计算量小（时间）？
- 准确率持平或者较高？
- 不同任务（或类别、图片）之间的相互干扰小
- 可以添加新任务（如新用户、新图片），或者新的原始特征而保持哈希转换后的特征长度不变，适合任务数频繁变化的问题
- 保持原始特征的稀疏性，哈希转换时，只有非0特征才起作用
- 可以只哈希一部分原始特征，而保留另一部分原始特征（如那些出现collision就会影响分类精度的）

1. K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature Hashing for Large Scale Multitask Learning. In *ICML*, *2009. 2,3*

特征哈希——理论证明

1. Unbiased estimate (无偏估计)

引入有符号的哈希后的特征和代替原本[2]无符号的特征和，这一修正可以产生无偏估计。

定义哈希函数h：N→{1,⋯..,m}（m是哈希后的特征维度），定义哈希函数ξ：N→{±1}。对于特征向量x，x'，特征哈希后的向量$\phi \in R^m$的第i个元素值和内积如下：

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

$$< x, x' >_{\phi} := < \phi^{(h,\xi)}(x), \phi^{(h,\xi)}(x') > \qquad (3)$$

2. Shi, Q., Petterson, J., Dror, G., Langford, J., Smola, A., Strehl, A., &Vishmanathan, V. (2009). Hash kernels. *Proc. Intl. Workshop on Artificial Intelligence and Statistics 12*.

特征哈希——理论证明

h(n)是能把{1,….,N}均匀哈希到{1,….,m}的函数

1. Unbiased estimate (无偏估计)

引入有符号的哈希后的特征和代替原本[2]无符号的特征和，这一修正可以产生无偏估计。

定义哈希函数h：N→{1,⋯..,m} （m是哈希后的特征维度），定义哈希函数ξ：N→{±1}。对于特征向量x，x'，特征哈希后的向量$\phi \in R^m$的第i个元素值和内积如下：

$$\phi_i^{(h,\xi)}(x) = \sum_{j:h(j)=i} \boxed{\xi(j)} x_j \quad (2)$$

$$< x, x' >_{\phi} := < \phi^{(h,\xi)}(x), \phi^{(h,\xi)}(x') > \quad (3)$$

这个就是引入的二值哈希函数，用来去除[2]中哈希核的固有偏差，这样才能保证内积的无偏性。通过这个哈希函数可以将N维映射为两维{-1，1}

满足h(j)=i的j的取值，j是哈希前的序号

2. Shi, Q., Petterson, J., Dror, G., Langford, J., Smola, A., Strehl, A., &Vishmanathan, V. (2009). Hash kernels. *Proc. Intl. Workshop on Artificial Intelligence and Statistics 12*.

特征哈希——理论证明

1. Unbiased estimate (无偏估计)

引入有符号的哈希后的特征和代替原本[1]无符号的特征和，这一修正可以产生无偏估计。

定义哈希函数h：N→{1,⋯..,m} （m是哈希后的特征维度），定义哈希函数ξ：N→{±1}。对于特征向量x，x'，特征哈希后的向量ϕ和相应的内积如下：

$$\phi_i^{(h,\xi)}(x) = \sum_{j:h(j)=i} \boxed{\xi(j)} x_j \quad (2)$$

这个就是引入的二值哈希函数，用来去除[2]中哈希核的固有偏差，这样才能保证内积的无偏性

$$< x, x' >_{\phi} := < \phi^{(h,\xi)}(x), \phi^{(h,\xi)}(x') > \quad (3)$$

内积相等是为了下面这个最终的预测

$$< \phi_0(x) + \phi_u(x), \omega_h > = < x, \omega_0 + \omega_u >$$

$$\omega_h = \phi_0(\omega) + \sum_{u \in U} \phi_{u(\omega_u)}$$

ω_u是某个具体类别（或者某张图片？）的学习参数，ω_0是全局参数，通过计算$< x, \omega_0 + \omega_u >$来获得最终的预测值。而

2. Shi, Q., Petterson, J., Dror, G., Langford, J., Smola, A., Strehl, A., &Vishmanathan, V. (2009). Hash kernels. *Proc. Intl. Workshop on Artificial Intelligence and Statistics* 12.

有了内积的等式，哈希前的预测结果也就和哈希后的预测结果相同了。

实际上，

$$<\phi_0(x)+\phi_u(x),\omega_h>=<x,\omega_0+\omega_u>+\epsilon_d+\epsilon_i$$

有了内积的等式，哈希前的预测结果也就和哈希后的预测结果相同了。

实际上，
$< \phi_0(x) + \phi_u(x), \omega_h > = < x, \omega_0 + \omega_u > + \epsilon_d + \epsilon_i$

$$\phi_i^{(h,\xi)}(x) = \sum_{j:h(j)=i} \boxed{\xi(j)} x_j$$

ξ到底做了什么？

ξ到底做了什么?

2. Analysis&Exponential tail bounds (哈希函数分析以及指数尾部边界)

辅助定理 2

哈希核是无偏的，即 $\mathbf{E}_{\emptyset}[<x, x'>_{\emptyset}]=<x,x'>$ 方差$\sigma^2_{x,x'}=\frac{1}{m}(\sum_{i\neq j}x_i^2\, x'^2_j + x_i x'_i x_j x'_j)$

故由$||x||_2=||x'||_2=1$, $\sigma^2_{x,x'}=O(\frac{1}{m})$ ➡ 哈希核的值就会被限制在目标值的$O(\frac{1}{\sqrt{m}})$

Chebyshev's 不等式证明了一半的观测值都在$\sqrt{2}\sigma$，然后，再根据凸距离不等式，就可以构建尾部指数边界

2.1 Concentration of Measure Bounds (量程聚焦)

这一部分主要展示了在哈希后的特征映射下，每个向量的长度都被尽可能的保留。（长度保留了，信息也就可以保留了）

令$\epsilon < 1$为不变常数，$\eta = \frac{||x||_\infty}{||x||_2}$，根据上面已给定的假设，哈希核满足下列不等式：

定理3

$$P\{\frac{||x||_\emptyset^2 - ||x||_2^2}{||x||_2^2} \geq \sqrt{2}\sigma_{x,x} + \varepsilon\} \leq \exp(-\frac{\sqrt{\epsilon}}{4\eta})$$

推论4　对两个向量x和x′，定义：

$$\sigma := \max\left(\sigma_{x,x}, \sigma_{x',x'}, \sigma_{x-x',x-x'}\right)$$

$$\eta := \min\left(\frac{||x||_\infty}{||x||_2}, \frac{||x'||_\infty}{||x'||_2}, \frac{||x-x'||_\infty}{||x-x'||_2}\right)$$

令 $\Delta = ||x||^2 + ||x'||^2 + ||x - x'||^2$，基于上述假设，有：

$$P[|\langle x, x'\rangle_\emptyset - \langle x, x'\rangle| > (\sqrt{2}\sigma + \epsilon)\Delta/2] < 3e^{-\frac{\sqrt{\epsilon}}{4\eta}}$$

令$\epsilon < 1$为不变常数，$\eta = \frac{||x||_\infty}{||x||_2}$，根据上面已给定的假设，哈希核满足下列不等式:

定理3

$$P\{\frac{||x||_\emptyset^2 - ||x||_2^2}{||x||_2^2} \geq \sqrt{2}\sigma_{x,x} + \varepsilon\} \leq \exp(-\frac{\sqrt{\epsilon}}{4\eta})$$

$$2\langle x, x'\rangle_\emptyset = \|x\|_\emptyset^2 + \|x'\|_\emptyset^2 - \|x - x'\|_\emptyset^2$$

$\longrightarrow$

标准内积不等式：$|2\langle \emptyset_u(x), \emptyset_u(x)\rangle - 2\langle x, x'\rangle|$
$\leq \left|\|\emptyset_u(x)\|^2 - \|x\|^2\right| + \left|\|\emptyset_u(x')\|^2 - \|x'\|^2\right|$
$+ \left|\|\emptyset_u(x - x')\|^2 - \|x - x'\|^2\right|$

$\Big\downarrow$ 定理3+内积标准不等式

$$P[|\langle x, x'\rangle_\emptyset - \langle x, x'\rangle| > (\sqrt{2}\sigma+\epsilon)\Delta/2] < 3e^{-\frac{\sqrt{\epsilon}}{4\eta}}$$

令$\epsilon < 1$为不变常数，$\eta = \frac{||x||_\infty}{||x||_2}$，根据上面已给定的假设，哈希核满足下列不等式：

定理3

$$P\{\frac{||x||_\emptyset^2 - ||x||_2^2|}{||x||_2^2} \geq \sqrt{2}\sigma_{x,x} + \varepsilon\} \leq \exp(-\frac{\sqrt{\epsilon}}{4\eta})$$

尽可能地保留了信息

定理3+内积标准不等式

$$P[|\langle x, x'\rangle_\emptyset - \langle x, x'\rangle| > (\sqrt{2}\sigma+\epsilon)\Delta/2] < 3e^{-\frac{\sqrt{\epsilon}}{4\eta}}$$

对理论3的边界进行扩展，大规模数据集向量（$x_1,\ldots,x_n$）之间的距离最大典型失真如下：

$$P\left(\frac{\left|\|x_i-x_j\|_\phi^2-\|x_i-x_j\|_2^2\right|}{\|x_i-x_j\|_2^2}\leq\sqrt{\frac{2}{m}}+64\eta^2 log^2\frac{n}{2\delta}\right)=1-\delta \qquad \text{推论5}$$

其中X=($x_1,\ldots,x_n$)是一系列满足$\|x_i-x_j\|_\infty\leq\eta\|x_i-x_j\|_2$ 的向量，这意味着观测值n的数量（或对应的非哈希矩阵的大小）在分析中只以对数的方式增加。

2.2 Multiple Hashing 多重哈希

从推论5看出，对于数值较大的η，只要x中的某些项数值很大，即使是单个的冲突，也会导致嵌入时明显的扭曲。为防止冲突带来的负面影响，对特征值比较大的多哈希几次。哈希C次后 $x'=\frac{1}{\sqrt{c}}(x,\ldots,x)$

辅助定理6　　若 $x'=\frac{1}{\sqrt{c}}(x,\ldots,x)$

1. 保范性：$\|x\|_2=\|x'\|_2$
2. 通过$\frac{1}{\sqrt{c}}=\frac{\|x'\|_\infty}{\|x\|_\infty}$ 减小成员大小
3. 方差增加为$\sigma_{x',x'}^2=\frac{1}{c}\sigma_{x,x}^2+\frac{c-1}{c}2\|x\|_2^4$

将辅助定理6应用到定理3中，大的值就可以以增加方差为代价来降低

$$\frac{\|x'\|_\infty}{\|x\|_2}=\frac{1}{\sqrt{c}}\eta=\eta'$$

2.3 Approximate 近似正交 ——为了证明互不影响

每一个任务都要学习一个不同的参数向量，当映射到同一个哈希特征空间时，我们希望不同参数向量间不存在相互影响。

令U为一系列不同的任务集合，某一任务u∈ U，ω是$U\backslash\{u\}$参数向量的集合. 对u中任意一个观测值x，在特征哈希的空间中 ω对x的干扰很小。记$\emptyset_u(x) = \emptyset^{(\xi,h)}((x,u))$

定理7 令$\omega \in R^m$是$U\backslash\{u\}$参数向量的集合,$\langle\omega, \emptyset_u(x)\rangle$受到如下限制

$$P\{|\langle\omega, \emptyset_u(x)\rangle| > \epsilon\} \leq 2e^{-\frac{\epsilon^2/2}{m^{-1}\|\omega\|_2^2\|x\|_2^2+\epsilon\|\omega\|_\infty\|x\|_\infty/3}}$$

特征哈希——应用 存储压缩，降低计算量，保留稀疏性

Personalization 垃圾邮件分类器

个人(local)权重 + 公共(global)权重

假定有数千个使用者U，使用者负责提供邮件的标签数据。预测器ω_u只是针对用户u，为了防止用户懒得标注，需要一个全局预测器ω_0。用不同的哈希函数$\emptyset_0,\ldots,\emptyset_{|U|}$哈希所有的权重向量$\omega_0,\ldots,\omega_{|U|}$：

$$\omega_h = \emptyset_0(\omega_0) + \sum_{u \in U} \emptyset_u(\omega_u)$$

预测垃圾邮件就是计算$\langle \emptyset_0(x_0) + \emptyset_u(x), \omega_h \rangle$，更精确些：

$$\langle \emptyset_0(x) + \emptyset_u(x), \omega_h \rangle = \langle x, \omega_0 + \omega_u \rangle + \epsilon_d + \epsilon_i$$

特征哈希——应用 存储压缩，降低计算量，保留稀疏性

Personalization 垃圾邮件分类器

个人(local)权重 + 公共(global)权重

假定有数千个使用者U，使用者负责提供邮件的标签数据。预测器ω_u只是针对用户u，为了防止用户懒得标注，需要一个全局预测器ω_0。用不同的哈希函数$\emptyset_0,\ldots,\emptyset_{|U|}$哈希所有的权重向量$\omega_0,\ldots,\omega_{|U|}$：

$$\omega_h = \emptyset_0(\omega_0) + \sum_{u \in U} \emptyset_u(\omega_u)$$

预测垃圾邮件就是计算$\langle \emptyset_0(x_0) + \emptyset_u(x), \omega_h \rangle$，更精确些：

$$\langle \emptyset_0(x) + \emptyset_u(x), \omega_h \rangle = \langle x, \omega_0 + \omega_u \rangle + \underbrace{\epsilon_d}_{\text{内积偏差，来源于自冲突}} + \epsilon_i$$

来自其他权重向量的影响

内积偏差，来源于自冲突

$$\epsilon_i = \sum_{v \in U, v \neq 0} \langle \emptyset_0(x), \emptyset_v(\omega_v) \rangle + \sum_{v \in U, v \neq u} \langle \emptyset_u(x), \emptyset_v(\omega_v) \rangle$$

$$\epsilon_i = \sum_{v \in U, v \neq 0} \langle \emptyset_0(x), \emptyset_v(\omega_v) \rangle + \sum_{v \in U, v \neq u} \langle \emptyset_u(x), \emptyset_v(\omega_v) \rangle$$

根据定理7，$\sum_{v \in U, v \neq 0} \omega_v$ 并不依赖于$\emptyset_0(x)$，第二项也可以由定理7的边界进行限制，所以ϵ_i很小

$$P\{|\langle \omega, \emptyset_u(x) \rangle| > \epsilon\} \leq 2e^{-\frac{\epsilon^2/2}{m^{-1}\|\omega\|_2^2\|x\|_2^2 + \epsilon\|\omega\|_\infty\|x\|_\infty/3}}$$

$$\epsilon_i = \sum_{v\in U, v\neq 0}\langle \emptyset_0(x), \emptyset_v(\omega_v)\rangle + \sum_{v\in U, v\neq u}\langle \emptyset_u(x), \emptyset_v(\omega_v)\rangle$$

根据定理7，$\sum_{v\in U, v\neq 0}\omega_v$ 并不依赖于$\emptyset_0(x)$，第二项也可以由定理7的边界进行限制，所以ϵ_i很小

$$\epsilon_d = \sum_{v\in\{u,0\}}|\langle \emptyset_v(x), \emptyset_v(\omega_v)\rangle - \langle x, \omega_v\rangle|$$

$$\epsilon_i = \sum_{v \in U, v \neq 0} \langle \emptyset_0(x), \emptyset_v(\omega_v) \rangle + \sum_{v \in U, v \neq u} \langle \emptyset_u(x), \emptyset_v(\omega_v) \rangle$$

根据定理7，$\sum_{v \in U, v \neq 0} \omega_v$ 并不依赖于$\emptyset_0(x)$，第二项也可以由定理7的边界进行限制，所以ϵ_i很小

$$\epsilon_d = \sum_{v \in \{u,0\}} |\langle \emptyset_v(x), \emptyset_v(\omega_v) \rangle - \langle x, \omega_v \rangle|$$

根据推论4可知ϵ_d通常也很小

$$\mathrm{P}[|\langle x, x' \rangle_{\emptyset} - \langle x, x' \rangle| > (\sqrt{2}\sigma + \epsilon)\Delta/2] < 3e^{-\frac{\sqrt{\epsilon}}{4\eta}}$$

$$\epsilon_i = \sum_{v\in U, v\neq 0} \langle \emptyset_0(x), \emptyset_v(\omega_v)\rangle + \sum_{v\in U, v\neq u} \langle \emptyset_u(x), \emptyset_v(\omega_v)\rangle$$

根据定理7，$\sum_{v\in U, v\neq 0} \omega_v$并不依赖于$\emptyset_0(x)$，第二项也可以由定理7的边界进行限制，所以$\epsilon_i$很小

$$\epsilon_d = \sum_{v\in\{u,0\}} |\langle \emptyset_v(x), \emptyset_v(\omega_v)\rangle - \langle x, \omega_v\rangle|$$

根据推论4可知ϵ_d通常也很小

数据： 3.2 million 邮件（已标记好），433167个用户

特征： 40 million 不同的词汇

对单个用户数据进行哈希的方法：

对单个用户数据进行哈希的方法：

It's weird that the mails you sent to me before were always sorted into the garbage box, and I tried to send a mail to you just now, but blocked.

邮件文本内容

构建词典 →

weird
mail
sort
…

每个词汇复制称两个 →
其中一个标记上用户标签

weird
常琳_weird
mail
常琳_mail
sort
常琳_sort
…

→ 哈希后的稀疏向量

$\boxed{\emptyset_0(x)} + \emptyset_u(x)$

谢谢