大数据分析

Scalable Machine Learning
least square regression

刘盛华

1

What is machine learning

- **Study of algorithms that**
  - **improve their performance**
  - **at some task**
  - **with experience**

Data ⇨ **Learning algorithm** ⇨ Understanding

(experience) **(task)** **(performance)**

Barnabás Póczos, CMU

2

Warnings about the Class

"There is nothing more practical than a good theory"

Lewin (1952)

3

Linear Regression

Sketching

4

Massive data sets

- **Examples**
  - 网络流量日志
  - 金融数据
  - 社交网络
  - …
- **Algorithms**
  - 需要线性时间复杂度或者更优
  - 经常需要以随机近似为代价

5

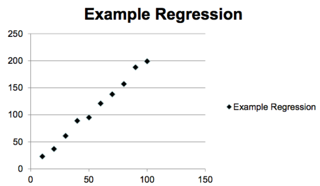
Regression analysis

- **回归分析**
  - 在噪声存在的情况下，运用统计的方法发现变量之间的关系。

6

Regression analysis

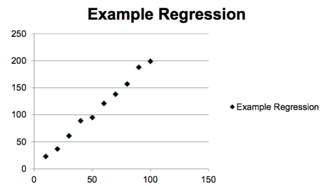
- **线性回归**
  - 在噪声存在的情况下，运用统计的方法发现变量之间的线性关系。
- **Example**
  - **欧姆定律 V = R · I**



7

Regression analysis

- **线性回归**
  - 在噪声存在的情况下，运用统计的方法发现变量之间的线性关系。
- **Example**
  - 欧姆定律 V = R · I
  - 寻找最能拟合数据的线性函数



8

Regression analysis

- **Linear Regression**
  - 在噪声存在的情况下，运用统计的方法发现变量之间的线性关系。
- **Standard Setting**
  - One measured variable b
  - A set of predictor variables $a_1, \ldots, a_d$
  - Assumption:

    $$b = x_0 + a_1 x_1 + \ldots + a_d x_d + \varepsilon$$
  - ε is assumed to be noise and the x_i are model parameters we want to learn
  - Can assume $x_0 = 0$
  - Now consider n observations of b

9

Regression analysis

- **矩阵形式**

Input: $n \times d$-matrix A and a vector $b=(b_1,\ldots, b_n)$
n is the number of observations; d is the number of predictor variables

Output: x^* so that Ax^* and b are close

- Consider the over-constrained case, when $n \gg d$
- Can assume that A has full column rank

10

Regression analysis

- **最小二乘方法**
  - Find x^* that minimizes $|Ax-b|_2^2 = \Sigma (b_i - \langle A_{i*}, x\rangle)^2$
  - A_{i*} is i-th row of A
  - Certain desirable statistical properties

11

Regression analysis

- **回归的几何形式**
  - We want to find an x that minimizes $|Ax-b|_2$
  - The product Ax can be written as

$$A_{*1}x_1 + A_{*2}x_2 + \ldots + A_{*d}x_d$$

where A_{*i} is the i-th column of A

  - This is a linear d-dimensional subspace
  - The problem is equivalent to computing the point of the column space of A nearest to b in l_2-norm

12

Time Complexity

- **通过正规方程求解最小二乘回归**
  - 需要计算 x = A⁻b
    - Moore-Penrose Pseudoinverse（伪逆） $A^- = V\Sigma^{-1}U^T$
  - 一般方法需要 nd^2 时间复杂度
  - 通过快速矩阵乘法能达到$nd^{1.376}$
  - 但我们想要更快！

13

Sketching to solve least squares regression

- How to find an approximate solution x to $\min_x |Ax-b|_2$?
- Goal: output x‘ for which $|Ax'-b|_2 \leq (1+\varepsilon) \min_x |Ax-b|_2$ with high probability
- Draw S from a k x n random family of matrices, for a value k << n
- Compute S*A and S*b
- Output the solution x‘ to $\min_{x'} |(SA)x-(Sb)|_2$
  - $x' = (SA)^- Sb$

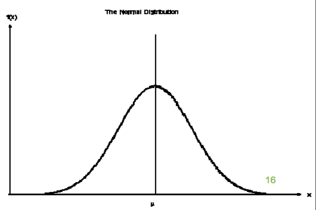
14

How to choose the right sketching matrix S?

- Recall: output the solution x‘ to $\min_{x'} |(SA)x-(Sb)|_2$
- Lots of matrices work
- S is d/ε^2 x n matrix of i.i.d. Normal random variables
- **S is a subspace embedding for colomn space of *A***

For all x, $|SAx|_2 = (1\pm\varepsilon)|Ax|_2$

* poof skipped



ref: David P. Woodruff, Sketching as a Tool for Numerical Linear Algebra, Foundations and Trends in Theoretical Computer Science, vol 10, issue 1-2, pp. 1-157 (ref to 10-40)

15

Subspace Embeddings for Regression

- Want x so that $|Ax-b|_2 \leq (1+\varepsilon) \min_y |Ay-b|_2$
- Consider subspace L spanned by columns of A together with b
- Then for all y in L, $|Sy|_2 = (1\pm \varepsilon) |y|_2$
- Hence, $|S(Ax-b)|_2 = (1\pm \varepsilon) |Ax-b|_2$ for all x
- Solve $\text{argmin}_y |(SA)y - (Sb)|_2$
- Given SA, Sb, can solve in poly(d/ε) time

Only problem is computing SA takes $O(nd^2)$ *time*

16

Faster Subspace Embeddings S

- CountSketch 矩阵S
- 定义 k x n 矩阵 S, for $k = O(d^2/\varepsilon^2)$
- S 非常稀疏: 每列随机选择非零元素的位置

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Can compute $S \cdot A$ in nnz(A) time! $<< nd < nd^2$

nnz(A) 是A矩阵中非零元素的数量

O(nnZ(A)) :每个A中的非零元素最多与S中的一个非零元素相乘

17

High Probability and Complexity

- **Theorem 2.5.** ([27]) For $\mathbf{S}$ a sparse embedding matrix with $r = O(d^2/\varepsilon^2 \mathrm{poly}(\log(d/\varepsilon)))$ rows, for any fixed $n \times d$ matrix $\mathbf{A}$, with probability .99, $\mathbf{S}$ is a $(1 \pm \varepsilon)$ ℓ_2-subspace embedding for $\mathbf{A}$. Further, $\mathbf{S} \cdot \mathbf{A}$ can be computed in $O(\mathrm{nnz}(\mathbf{A}))$ time.
- **Theorem 2.14.** The ℓ_2-Regression Problem can be solved with probability .99 in $O(\mathrm{nnz}(A)) + \mathrm{poly}(d/\varepsilon)$ time.

18