CS 15-859: Algorithms for Big Data

Lecture 1

Prof. David Woodruff *Scribe: Ainesh Bakshi*

1 Course Information

- *Professor :* David Woodruff .
- *TA:* Dhivya Eswaran.

2 Overview

The motivation for studying algorithms for big data stems from the massive data sets that are available today and the requirement to process them efficiently. Data sets such as internet traffic logs, financial data etc. are large enough to warrant near linear time processing and often cannot be entirely stored in memory. Algorithms achieving such constraints usually do so at the cost of a randomized approximation. We consider the regression problem which is ubiquitous in machine learning, statistics and data mining. Regression is a statistical method to study dependencies between the variables in the presence of noise. More specifically, we look at linear regression, where the dependencies are linear.

3 Linear Regression

In the standard setting for linear regression, there is a measured variable b and a set of predictor variables $a_1, a_2 \dots a_d$. The working assumption is that $b = x_0 + a_1x_1 + a_2x_2 + \dots a_dx_d + \epsilon$, where ϵ is the noise and the x_is are the coefficients of a hyperplane (model parameters) that we wish to learn. We can assume $x_0 = 0$ by adding a_0 to our model and always setting it to 1. Thus, w.l.o.g. we work with $b = a_1x_1 + a_2x_2 + \dots a_dx_d + \epsilon$. The noise ϵ may be adversarial or may come from a distribution we have no information about.

Consider an experiment in which we receive n observations of the form $(a_{i,1}, a_{i,2} \dots a_{i,d}, b_i)$, for all $i \in [1, n]$. It is convenient to think about the observations in matrix form where we are given a $n \times d$ matrix $\mathbf{A}$, where each row i has d predictor variables corresponding to the i^{th} observation. Additionally, there is a column vector $\mathbf{b}$, where the i^{th} entry is b_i. The goal of the regression problem is to output a vector $\mathbf{x}$ such that $\mathbf{Ax}$ is close to $\mathbf{b}$ under an appropriate notion of closeness. We also assume that the number of observations, n, is much larger than the number of predictor variables, d.

3.1 Least Squares Method

One of the most common notions of closeness between $\mathbf{Ax}$ and $\mathbf{b}$ is the least squares method, which minimizes the Euclidean distance between them. Formally,

$$argmin_x||\mathbf{Ax}-\mathbf{b}||_2 = argmin_x \sum_i^n (b_i - \langle \mathbf{A}_{i,*}\mathbf{x}\rangle)^2$$

where $\mathbf{A}_{i,*}$ is the i^{th} row of matrix $\mathbf{A}$ and b_i is the i^{th} entry of vector $\mathbf{b}$. As an aside, if ϵ is independently sampled gaussian noise, then x is the Maximum Likelihood Estimator for the data.

The least squares method also has an aesthetic geometric interpretation. The matrix vector product $\mathbf{Ax}$ can be rewritten as $\mathbf{A}_{*,1}x_1 + \mathbf{A}_{*,2}x_2 + \dots \mathbf{A}_{*,d}x_d$, where $\mathbf{A}_{*,i}$ is the i^{th} column of $\mathbf{A}$. Note, this is a linear d-dimensional subspace. The least squares problem is then equivalent to finding a point in the column space of A that is nearest to b in Euclidean distance.

In order to find the solution to $argmin_x||\mathbf{Ax}-\mathbf{b}||_2$, we can instead consider the equivalent problem, $argmin_x||\mathbf{Ax}-\mathbf{b}||_2^2$. Let $b = \mathbf{Ax'} + \mathbf{b'}$, where $\mathbf{b'}$ is orthogonal to the column space of $\mathbf{A}$. By Pythagorean theorem, the cost is then $argmin_x||\mathbf{A}(\mathbf{x}-\mathbf{x'})||_2^2 - ||\mathbf{b'}||_2^2$. Observe, x is an optimal solution if and only if $\mathbf{A}^T(\mathbf{Ax}-\mathbf{b}) = \mathbf{A}^T\mathbf{Ax} - \mathbf{A}^T\mathbf{Ax'} - \mathbf{A}^T\mathbf{b'} = \mathbf{A}^T\mathbf{Ax} - \mathbf{A}^T\mathbf{Ax'} = 0$. The first equality follows from $b = \mathbf{Ax'} + \mathbf{b'}$, the second follows from $\mathbf{A}^T$ being orthogonal to $\mathbf{b}$ and the last follows from $||\mathbf{A}(\mathbf{x}-\mathbf{x'})||_2^2 \geq 0$ and thus the cost is minimized when it is set to 0. The equation $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ is known as the *normal equation* and any optimal x satisfies it. Note, if the columns of A are linearly independent, then $\mathbf{A}$ has full rank and $x = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$. If the columns of $\mathbf{A}$ are not linearly independent, the Moore-Penrose pseudoinverse gives an optimal minimum norm solution.

3.1.1 Moore-Penrose Pseudoinverse

Let $\mathbf{A} = \mathbf{U\Sigma V}^T$ be the Singular Value Decomposition (SVD) of $\mathbf{A}$, where $\mathbf{U}$ is a $n \times d$ matrix with orthonormal columns, $\mathbf{\Sigma}$ is a $d \times d$ diagonal matrix with non-zero non-increasing entries down the diagonal, and $\mathbf{V}^T$ is a $d \times d$ matrix with orthonormal rows. Then, the Moore-Penrose pseudoinverse, $\mathbf{A}^\dagger$, is the $d \times n$ matrix $\mathbf{V\Sigma}^\dagger\mathbf{U}^T$, where $\mathbf{\Sigma}^\dagger$ is a $d \times d$ diagonal matrix with $\mathbf{\Sigma}^\dagger_{i,i} = \frac{1}{\mathbf{\Sigma}_{i,i}}$ if $\mathbf{\Sigma}_{i,i} > 0$ and 0 otherwise.

Claim 1. *The solution $\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{b}$ is optimal and has minimum norm.*

Proof. Substituting $\mathbf{x} = \mathbf{A}^\dagger, \mathbf{b}$ in the normal equation, it is sufficient to show that $\mathbf{A}^T\mathbf{A}(\mathbf{A}^\dagger\mathbf{b}) = \mathbf{A}^T\mathbf{b}$. By definition, $\mathbf{A} = \mathbf{U\Sigma V}^T$ and $\mathbf{A}^T = \mathbf{V\Sigma U}^T$. Then, $\mathbf{A}^T\mathbf{A}(\mathbf{A}^\dagger\mathbf{b}) = \mathbf{V\Sigma U}^T\mathbf{U\Sigma V}^T\mathbf{V\Sigma}^\dagger\mathbf{U}^T\mathbf{b}$. Note $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ since they have orthonormal rows and columns respectively. Further, $\mathbf{\Sigma\Sigma}^\dagger = \mathbf{I}$. Therefore, we get $\mathbf{V\Sigma U}^T\mathbf{b} = \mathbf{A}^T\mathbf{b}$, proving $\mathbf{x}$ is an optimal solution.

Observe, we can generate an affine space of solutions by adding a vector orthogonal to the column space of A. Therefore, any optimal solution has the form $\mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{V'V'}^T)z$, where $\mathbf{V'}^T$ corresponds to the rows i of $\mathbf{V}^T$ for which $\mathbf{\Sigma}_{i,i} > 0$. Since $\mathbf{A}(\mathbf{I} - \mathbf{V'V'}^T)z = 0$, this is an optimal solution. Additionally, $\mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{V'V'}^T)z$ is a ($d-$ rank($\mathbf{A}$)-dimensional affine space that spans all optimal solutions. Since $\mathbf{A}^\dagger\mathbf{b}$ is in the column spam of $\mathbf{V'}$, using the Pythagorean theorem, $||\mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{V'V'}^T)z||_2^2 = ||\mathbf{A}^\dagger\mathbf{b}||_2^2 + ||(\mathbf{I} - \mathbf{V'V'}^T)z||_2^2 \geq ||\mathbf{A}^\dagger\mathbf{b}||_2^2$. Therefore, $\mathbf{x} = \mathbf{A}^\dagger\mathbf{b}$ is the solution with the minimum norm. □

The main problem with the above solution is that on sufficiently large data sets, such as those in practice, matrix multiplication and singular value decomposition are prohibitively expensive. Naively computing the SVD requires $min(nd^2, dn^2)$ time. Using fast matrix multiplication we can bring it down to $nd^{1.376}$. However, we are interested in algorithms that run much faster.

3.1.2 The Sketch and Solve Paradigm

In this section, we explore sketching techniques to improve upon the above time complexities, at the cost of settling for a randomized approximation algorithm. Let us consider the relaxed version of the problem:

$$argmin'_x||\mathbf{Ax'} - \mathbf{b}||_2 = (1+\epsilon)argmin_x||\mathbf{Ax} - \mathbf{b}||_2$$

where $\mathbf{x}$ is the optimal hyperplane. The sketch and solve paradigm uses the following high-level algorithm:

- Draw S from a $k \times n$ random family of matrices, where $k << n$.
- Computer **S*A** and **S*b**
- Output solution $\mathbf{x'}$ to $min_{x'}||(\mathbf{SA})\mathbf{x} - \mathbf{Sb}||_2^2$

Many families of random matrices work. Let $\mathbf{S}$ be a $d/\epsilon^2 \times n$ matrix of i.i.d. Normal random variables with mean 0 and variance $1/k$, where $k = O(d/\epsilon^2)$. For any fixed d-dimensional subspace, i.e. the column space of $\mathbf{A}$, w.h.p for all $x \in \mathbf{R}^d$, $||\mathbf{SAx}||_2 = (1 \pm \epsilon)||\mathbf{Ax}||_2$. To see why this is true, we segue into subspace embeddings.

3.1.3 Subspace Embeddings

In this section, we want to prove that w.h.p for all $x \in R^d$, $||\mathbf{SAx}||_2 = (1 \pm \epsilon)||\mathbf{Ax}||_2$. Since we want to prove this for all x, we can assume that the columns of A are orthonormal. The first property we need is the two stability of normal random variables.

Claim 2. *Let X and Y be two random variables such that X is drawn from $\mathcal{N}(0, a^2)$ and Y is drawn from $\mathcal{N}(0, b^2)$. Then, X+Y is drawn from $\mathcal{N}(0, a^2 + b^2)$.*

Proof. Observe that the probability density function of f_z of $Z = X + Y$ is a convolution of probability density functions f_x and f_y. By definition, $f_z(z) = \int f_x(z-y)f_y(y)dy$, where $f_x(x) = \frac{1}{\sqrt{2\pi}a}e^{-x^2/2a^2}$ and $f_y(y) = \frac{1}{\sqrt{2\pi}b}e^{-y^2/2b^2}$. Then,

$$f_z(z) = \int \frac{1}{\sqrt{2\pi}a}e^{-(z-y)^2/2a^2}\frac{1}{\sqrt{2\pi}b}e^{-y^2/2b^2}dy$$

$$= \frac{1}{\sqrt{2\pi(a^2+b^2)}}e^{-z^2/2(a^2+b^2)}\int \frac{\sqrt{(a^2+b^2)}}{\sqrt{2\pi}ab}e^{-\frac{(y-\frac{b^2 z}{a^2+b^2})^2}{2(\frac{a^2b^2}{a^2+b^2})}}dy$$

Observe that the integral evaluates to 1 since it is a Guassian distribution and Z $\sim \mathcal{N}(0, a^2+b^2)$. □

The second property that we require is the rotational invariance of Gaussian random variables. This is shown in the second part of the lecture.