CS 15-851: Algorithms for Big Data Spring 2024

Lecture 1 — January 18

Prof. David Woodruff *Scribe: Mahbod Majid*

1 Motivation

In many instances, we have a large amount of data, and we want to extract information from it. For example, when dealing with internet traffic logs, financial data, etc. In these instances, naive algorithms are not efficient enough, and we usually want to design algorithms that are *nearly linear* time or *sublinear* in the size of the data. As we will see in this course, in many instances, we can achieve this by using randomization.

2 Linear Regression

The first problem we consider is linear regression.

Linear Regression. A statistical model to study linear dependencies between variables in the presence of noise.

One example of a linear relationship between two variables is the following:

Example 1 (Ohm's Law)**.** One example of this is Ohm's law, which states that the current through a conductor between two points is directly proportional to the voltage across the two points. Mathematically, we can write this as $V = IR$, where V is the voltage, I is the current, and R is the resistance. We can also write this as $I = V/R$. This is a linear relationship between I and V.

We can formalize the linear regression problem as follows:

Standard Linear Regression. Let b denote the measured variable, and $a_i, \ldots, a_d$ be a set of predictor variables. We assume that the relationship between b and a_i is linear, i.e.

$$b = x_0 + a_1 x_1 + \cdots + a_d x_d + \varepsilon,$$

where ε is a random variable representing the noise, and the goal is to learn $x_0, \ldots, x_d$. Note that here we may assume that $x_0 = 0$ without loss of generality, as we may always add a new predictor variable a_0 that is always 1.

Now let's assume we have n observations of b and a_i's. In this setting it is convenient to consider the regression problem in matrix form.

Linear Regression Matrix Form. Suppose n observations of the standard linear regression model are given. Namely, we have access to variables $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times d}$, where $A_{i,j}$ is the j-th predictor variable of the i-th observation. Our goal is to find the d-dimensional vector x^* such that Ax^* is close to b.

Specifically, we are interested in the over constrained setting where $n \gg d$. In this setting, we have more observations than the number of variables, and we want to find the best fit. In this setting, we can't find a solution that satisfies all the equations, so we want to find a solution that minimizes the error.

2.1 Least Squares Method

Probably the most common way to measure closeness is the least squares model. Here what we are trying to do is find x^* that minimizes the sum of the squares of the errors. We can write this as follows:

$$\|Ax - b\|_2^2 = \sum_{i=1}^{n} (\langle A_{i,*} x \rangle - b_i)^2,$$

where $A_{i,*}$ is the i-th row of A. The least squares estimator has desirable statistical properties. For example, one can show that the least squares estimator is a maximum likelihood estimator.

2.2 Geometric Interpretation of Least Squares

We want to find an x^* such that Ax^* is close to b, i.e. x^* minimizes $\|Ax - b\|_2$. The product Ax is a linear combination of the columns of A with coefficients x_i. Therefore, we can write Ax as follows:

$$Ax = A_{*,1} x_1 + \cdots + A_{*,d} x_d,$$

where $A_{*,i}$ is the i-th column of A. This forms a d-dimensional subspace of $\mathbb{R}^n$. The problem is equivalent to computing the point of the column space of A that is closest to b, in ℓ_2 norm.

2.3 Solving Least Squares Regression via the Normal Equations

Our goal is to find the solution x to $\min_x \|Ax - b\|_2$. We can write b as the sum of a member of the column space of A, and a vector orthogonal to the column space of A.

$$b = Ax' + b', \text{where } b' \text{ is orthogonal to the columns of } A.$$

Now by the Pythagorean theorem we can write,

$$\|Ax - b\|_2^2 = \|Ax - Ax' - b'\|_2^2 = \|A(x - x') - b'\|_2^2 = \|A(x - x')\|_2^2 + \|b'\|_2^2 \tag{1}$$

Claim 1. x is an optimal solution if and only if,

$$A^T(Ax - b) = A^T(Ax - Ax') = 0.$$

Proof. First note that $A^T(Ax - b) = A^T(Ax - Ax')$. This is because b' is orthogonal to the columns of A, and $b = Ax' + b'$. Now note that by Equation 1, we have that

$$\arg\min_x \|Ax - b\|_2^2 = \arg\min_x \|A(x - x')\|_2^2$$

Therefore, x is an optimal solution iff $A(x - x') = 0$. It is easy to see if this is true then $A^T(Ax - Ax') = 0$. It remains to show that if $A^T(Ax - Ax') = 0$, then $A(x - x') = 0$. To see this, note that

$$A^T(Ax - Ax') = 0 \implies (x - x')^T A^T A(x - x') = 0 \implies \|A(x - x')\|_2^2 = 0 \implies A(x - x') = 0.$$

■

Therefore any optimal solution x must satisfy the normal equations.

Definition (Normal Equations)**.** For a linear regression problem and variables $A \in \mathbb{R}^{n \times d}$, and $b \in \mathbb{R}^n$, we say $x \in \mathbb{R}^d$ satisfies the normal equations iff $A^T Ax = A^T b$.

If the columns of A are independent then we can write the solution as:

$$x = (A^T A)^{-1} A^T b$$

What can we do if the columns of A are not independent? In that case the solution is not unique: addition with any member of the kernel would still be a solution. We can use the psuedo-inverse to find x in this case.

2.4 Moore-Penrose Pseudo-inverse

Theorem 1 (Singular Value Decomposition (SVD))**.** *Any matrix A can be written as*

$$A = U \cdot \Sigma \cdot V^T,$$

where,

- *U has orthonormal columns,*
- *Σ is diagonal with non-increasing non-negative entries down the diagonal,*
- *V has orthonormal columns or equivalently V^T has orthonormal rows.*

Definition (Pseudo-inverse)**.** The pseudo-inverse of A is defined as

$$A^- = V\Sigma^{-1}U^T,$$

where Σ^{-1} is the diagonal matrix with the reciprocals of the diagonal entries of Σ. The i-th diagonal entry is equal to $1/\Sigma_{ii}$ if $\Sigma_{ii} \neq 0$ and 0 otherwise.

Theorem 2 (Finding an Optimal Solution to Least Squares Regression)**.** *Consider a least squares regression problem with variables $A \in \mathbb{R}^{n \times d}$, and $b \in \mathbb{R}^n$. Let $x = A^- b$. Then x is an optimal solution to the problem.*

Proof. We want to show that x satisfies the normal equations. Using Theorem 1, we can write the following:

$$\begin{aligned} A &= U\Sigma V^T \\ A^T &= V\Sigma^T U^T \\ A^- &= V\Sigma^{-1} U^T \end{aligned}$$

Now let's substitute, and see if x satisfies the normal equations, i.e. $A^T Ax = A^T b$.

$$\begin{aligned} A^T Ax &= A^T AA^- b \\ &= V\Sigma^T U^T U\Sigma V^T V\Sigma^{-1} U^T b \\ &= V\Sigma^T \Sigma\Sigma^{-1} U^T b \\ &= V\Sigma U^T b \\ &= A^T b, \end{aligned}$$

where we used the fact that U is orthonormal, V is orthonormal. Moreover, and $\Sigma^T \Sigma\Sigma^{-1} = \Sigma$ since Σ is diagonal. Therefore, we have that $A^T Ax = A^T b$, as desired. ■

Theorem 3 (Characterizing the Set of Optimal Solutions)**.** *Any optimal solution x has the form $A^- b + (I - V'V'^T)z$ where V'^T corresponds to the rows i of V^T for which $\Sigma_{i,i>0}$. Moreover, among all solutions $A^- b$ is the one with the smallest norm.*

Proof. Suppose Σ has r non-zero entries. We can write A as follows:

$$A = U\Sigma \begin{bmatrix} V'^T \\ V''^T \end{bmatrix},$$

where V'^T is $r \times d$ and V''^T is the rest of the rows of V^T. Since V has orthonormal columns, we have that $I = P_{V'} + P_{V''} = V'V'^T + V''V''^T$, where $P_{V'}$ is the projection onto the column space of V'. Therefore, $I - V'V'^T = V''V''^T$. We need to show that x satisfies the normal equations, i.e. $A^T Ax = A^T b$. To see this, note that

$$\begin{aligned} A^T Ax - A^T b &= A^T A(A^- b + (I - V'V'^T)z) - A^T b \\ &= A^T AA^- b + A^T A(I - V'V'^T)z - A^T b \\ &= A^T AV''V''^T z \\ &= 0. \end{aligned}$$

The last equality is because $AV'' = 0$, since, $A = U\Sigma \begin{bmatrix} V'^T \\ V''^T \end{bmatrix}$, and V''^T corresponds to the rows of V^T for which $\Sigma_{i,i} = 0$. Therefere, any point of the form $A^- b + (I - V'V'^T)z$ satisfies the normal equations. We need to show that any optimal solution has this form. To see this, note that this set is a $d - \text{rank}(A)$-dimensional affine space so it spans all optimal solutions.

It remains to show that $A^- b$ is the solution with the smallest norm. To see this, note that $A^- b$ is in the column span of V'. This is because,

$$A^- b = \begin{bmatrix} V'V'' \end{bmatrix} \begin{bmatrix} \Sigma'^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T b$$

So all of the terms affected by V'' are zero. Therefore A^-b is in the column span of V'. By the Pythagorean theorem, we have that

$$\|A^-b + (I - V'V'^T)z\|_2^2 = \|A^-b\|_2^2 + \|(I - V'V'^T)z\|_2^2 \geq \|A^-b\|_2^2,$$

as desired. ■

2.5 Time Complexity

We need to compute $x = A^-b$. Naively, this takes $O(nd^2)$ time. If we use fast matrix multiplication results we can reduce this to get $O(nd^{1.376})$ time. This would still be slow if we have a large number of observations n. Here's where sketching comes in.

2.6 Sketching to Solve Least Squares Regression

Our goal is to find an approximate solution x to $\min_x\|Ax - b\|_2$, i.e. we want to find an x such that

$$\|Ax' - b\|_2 \leq (1 + \varepsilon)\|Ax^* - b\|_2,$$

with high probability.

Sketching. Consider a $k \times n$ matrix S that is fat and wide ($k \ll n$). The idea is to instead solve the problem under a random S, and hope that the solution would still be an approximate minimizer and it would save us time by reducing the time complexity to $O(kd^2)$. We can write the problem as follows:

$$\min_x\|SAx - Sb\|_2$$

Now the question is what choice of S works for us. First, let's consider S to be a random matrix with independent Gaussian entries. We can show that this works for us. Let $k = O(\frac{d}{\varepsilon^2})$, and assume

$$S_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/k),$$

where $\mathcal{N}(0, 1/k)$ is the normal distribution with mean 0 and variance $1/k$.

Theorem 4 (Subspcae Embedding)**.** *For any fixed d-dimensional subspace, i.e. the column space of an $n \times d$ matrix A, we have that with high probability over the choice of S,*

$$\forall x : \|SAx\|_2 \leq (1 + \varepsilon)\|Ax\|_2$$

Note that the order of the quantifiers here is important. We want to show that with high probability *for any* x the inequality holds. We don't want to show that for each x the inequality holds with high probability.

Proof. We can simplify the problem.

Claim 2. Without loss of generality, we may assume that the columns of A are orthonormal, and that x is norm 1.

This is because we can instead pick another A' where the columns of A' form an orthonormal basis for the column space of A. The second part is because we can scale both sides by $\|x\|_2$.

Claim 3. If A's columns are orthonormal, then SA itself is a $k \times d$ matrix of i.i.d. $\mathcal{N}(0, 1/k)$ Gaussian entries.

We use Facts 1 and 2. First we know that each row of SA is independent, as they use different random bits. Second, we know that each row of SA has the following form

$$\begin{bmatrix} \langle g, A_1 \rangle & \cdots & \langle g, A_d \rangle \end{bmatrix}$$

From the facts below we know that each entry is independent and is distributed as $\mathcal{N}(0, 1/k)$. Therefore, we have that SA is a $k \times d$ matrix of i.i.d. $\mathcal{N}(0, 1/k)$ Gaussian entries.

Please refer to the next lecture note for the rest of the proof. ■

Fact 1. For any two independent random variables X and Y drawn from $N(0, a^2)$, and $N(0, b^2)$ we have that $X + Y$ is distributed as $N(0, a^2 + b^2)$.

Fact 2. If u and v are vectors with $\langle u, v \rangle = 0$, then $\langle g, u \rangle$ and $\langle g, v \rangle$ are independent, where g is a vector of i.i.d. $N(0, 1/k)$ Gaussian entries. Moreover, this property holds for set of orthogonal set of vectors.