CS 15-851: Algorithms for Big Data Spring 2025

Lecture 1 Part 1 — 01/16/2025

Prof. David Woodruff *Scribe: Lauren Sands*

1 Introduction

There are many large data sets that need to be processed (e.g., internet traffic logs, financial data). Because these data sets are so large, we need algorithms that are linear or sublinear to analyze them. Other algorithms will simply be too slow. Usually, this means that we will introduce some randomness (randomness over the algorithm's choices, not over the input).

2 Regression

2.1 Definition and Representation

Regression: Statistical method to study dependencies between variables in the presence of noise.

Linear Regression: Statistical method to study **linear** dependencies between variables in the presence of noise.

For example, Ohm's law says $V = I \cdot R$. The graph below plots I vs. V. We want to find the linear function that best fits the data, which would represent the resistance.

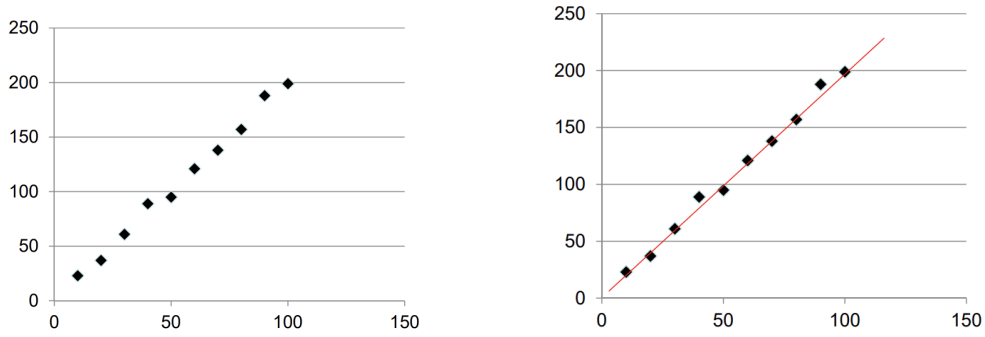

Figure 1: Example graphs of V vs. I, and the line of best fit.

Standard Setting:

Standard representation of a linear relationship between the measured and predictor variables: $b = x_0 + a_1x_1 + ... + a_dx_d + \epsilon$

- One measured variable b: In the example, voltage V.
- A set of predictor variables $a_1, ..., a_d$: In the example, a_1 would be the current I.

- Model parameters x_i which are the unknowns: In the example, resistance R.
- The noise, ϵ
- Replacing d with d + 1 allows us to ignore the free variable, so we can assume $x_0 = 0$

Consider n observations of b. These are like the data points in the graph. It can be useful to represent our regression in matrix form.

Matrix Form for Linear Regression:

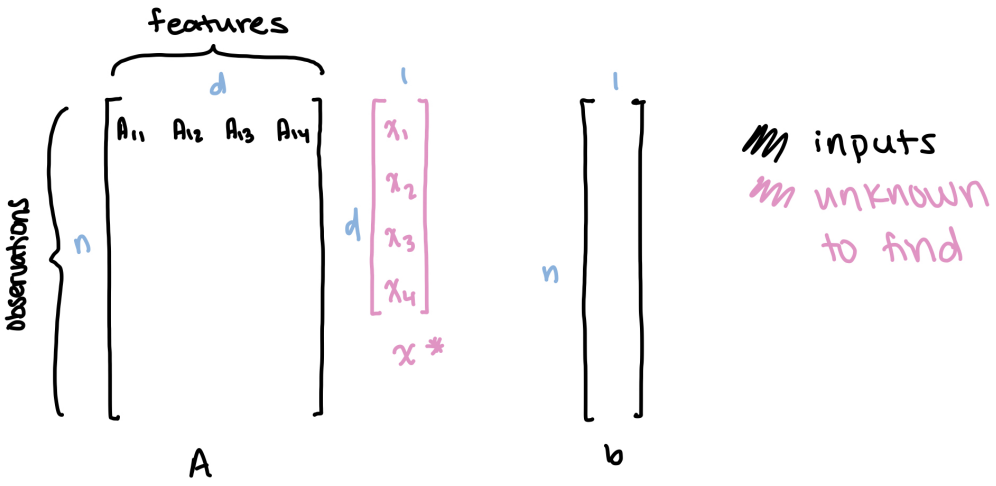

Figure 2: Visual representation of the matrix form for linear regression.

The *input* to the regression consists of:

- An $n \times d$ matrix A, with a row per observation, and a column per predictor variable.
- An $n \times 1$ vector b.

The *output* is a $d \times 1$ vector x^* that minimizes the difference between Ax^* and b. Note here that we want to minimize the error because we cannot always find a vector x^* such that $Ax^* = b$. In an extreme case, consider if A contains all 0's, and b is non-zero. Then there is certainly no x^* such that $Ax^* = b$. We are working in the over-constrained case. That is, $n \gg d$. IN this case we likely cannot find a solution that satisfies $Ax^* = b$, so we aim to reduce error.

2.2 Least Squares Method

A common way to measure closeness is the least squares method. We want to find an x^* that minimizes the following error:

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

In the equation above, b_i is an observation (the dependent variable), A_{i*} is the ith row of A, and $\langle A_{i*}, x \rangle$ is the prediction for the ith datapoint.

The least squares method also has certain desirable statistical properties.

2.3 Geometry of Regression

Wan to find an x such that Ax is close to b. That is, we want to minimize $\|Ax - b\|$.

Now we can rewrite Ax as follows, where A_{*i} is the ith column of A:

$$Ax = A_{*1}x_1 + A_{*2}x_2 + ... + A_{*d}x_d$$

We can see that this is a d-dimensional subspace of $\mathbb{R}^n$. In particular, this is the column space of A, or $\text{Col}(A)$. Thus, this problem is equivalent to computing the point in the column space of A that is closest to b.

Therefore, the best-fit solution Ax^* is the **projection** of b onto $\text{Col}(A)$. This is represented in the image below.

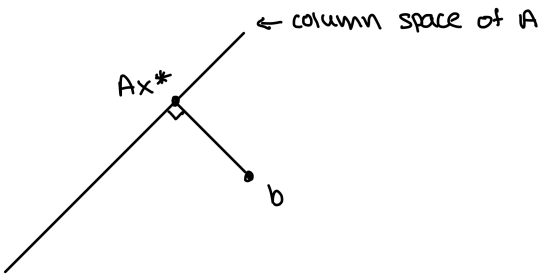

Figure 3: To minimize $\|Ax - b\|$, we want to find the projection of b onto the column space of A.

2.4 Solving Least Squares Regression via the Normal Equations

Our goal is to find the x that minimizes $\|Ax - b\|$. This is equivalent to minimizing $\|Ax - b\|^2$, since minimizing the distance and the square of the distance are the same.

We will decompose b into a part that is in $\text{Col}(A)$, and a part that is orthogonal to $\text{Col}(A)$. Geometrically, this is represented in the image below.

Ax' is the component of b that is in $\text{Col}(A)$, and b' is the component that is orthogonal to $\text{Col}(A)$. Thus, $b = Ax' + b'$.

Claim 1. The x that minimizes $\|Ax - b\|$ occurs when $\|A(x - x')\|^2 = 0$.

Proof. By the Pythagorean theorem, we know $\|Ax - b\|^2 = \|A(x - x')\|^2 + \|b'\|^2$. The value we want to minimize is $\|Ax - b\|^2$, so this is equivalent to minimizing $\|A(x - x')\|^2 + \|b'\|^2$. We will pay the cost of $\|b'\|^2$ no matter what. Therefore, our optimal solution occurs when $\|A(x - x')\|^2 = 0$. ■

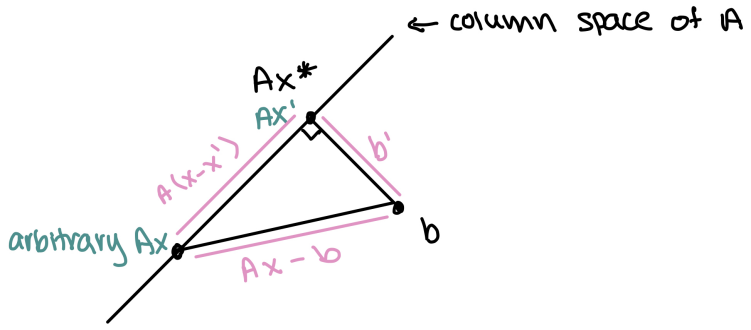

Figure 4: Geometric representation of decomposing b.

Claim 2. x is an optimal solution iff $A^T(Ax - b) = 0$

Proof. By claim 1, we know our optimal solution occurs when $\|A(x - x')\|^2 = 0$. Thus,

$$\begin{aligned}
\langle A(x-x'), A(x-x')\rangle &= 0 && \text{because } \|v\| = \sqrt{\langle v, v\rangle} \\
(A(x-x'))^T A(x-x') &= 0 \\
(x-x')^T A^T A(x-x') &= 0 \\
A^T A(x-x') &= 0 \\
A^T(Ax - Ax') &= 0
\end{aligned}$$

$$\begin{aligned}
A^T(Ax-b) &= A^T(Ax - Ax' - b') && \text{because } b = Ax' + b' \\
&= A^T(Ax - Ax') && \text{because } b' \perp \text{Col}(A), \text{ so multiplying by } A^T \text{ cancels} \\
&= A^T(A(x-x'))
\end{aligned}$$

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

Normal Equation: $A^T Ax = A^T b$ for any optimal solution x.

This is derived from claim 2, because $A^T(Ax - b) = 0 \implies A^T Ax - A^T b = 0 \implies A^T Ax = A^T b$.

If the columns of A are linearly independent, we can simply solve for $x = (A^T A)^{-1} A^T b$.

If the columns of a are *not* linearly independent, there are multiple possible solutions. If x^* is an optimal solution, then $x^* + y$ will also be optimal, where $y \in \text{Kernel}(A)$. In this case, we want to pick some canonical solution from the family $x^* + y$. We will choose the one with the smallest norm (which will be unique). We can use the Moore-Penrose pseudoinverse to get the minimum norm solution x.

2.5 Moore-Penrose Pseudoinverse

Singluar Value Decomposition (SVD): Any matrix A can be written in the form $U \cdot \Sigma \cdot V^T$, where

- U is $n \times d$ and has orthonormal columns (unit length and orthogonal to each other).
- Σ is a diagonal $d \times d$ matrix, where $\Sigma_{1,1} \geq \Sigma_{2,2} \geq \Sigma_{3,3} \geq ....$
- V^T is $d \times d$ and has orthonormal rows. Note that since V^T is square, it also has orthonormal columns.

In this class, we should always be thinking about writing matrices in SVD form!

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

- Σ^{-1} is a diagonal $d \times d$ matrix with $\Sigma^{-1}_{i,i} = 1/\Sigma_{i,i}$ if $\Sigma_{i,i}$ is positive, and 0 otherwise. Visually, this would look something like:

$$\begin{bmatrix} \frac{1}{\Sigma_{11}} & & & & \\ & \ddots & & & \\ & & \frac{1}{\Sigma_{rr}} & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}$$

What does A^-A look like? We can analyze it in SVD form:

$$\begin{aligned} A^-A &= (V\Sigma^{-1}U^T)(U\Sigma V^T) \\ &= V\Sigma^{-1}\Sigma V^T \qquad\qquad \text{because } U^TU = I \text{ for orthogonal matrices} \end{aligned}$$

Now let's see what $\Sigma\Sigma^{-1}$ looks like:

$$\Sigma\Sigma^{-1} = \begin{bmatrix} \Sigma_{11} & & & \\ & \ddots & & \\ & & & \Sigma_{dd} \end{bmatrix} \begin{bmatrix} \frac{1}{\Sigma_{11}} & & & & \\ & \ddots & & & \\ & & \frac{1}{\Sigma_{rr}} & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}$$

Putting it all together, we see that A^-A is equal to:

$$A^-A = V \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} V^T$$

Note that if A has full column rank, $\Sigma\Sigma^{-1}$ will be the identity, so the pseudoinverse is the same as the inverse in this case (because V is orthogonal so $VV^T = I$).

Claim 3. $x = A^-b$ is an optimal solution to the least squares regression.

Proof. The solution is optimal iff the normal equation holds. So we need to check if $A^TAx = A^Tb$.

First, let's write out the SVD of some matrices that will be helpful:

$$\begin{aligned} A &= U\Sigma V^T \\ A^T &= V\Sigma^T U^T = V\Sigma U^T && \text{because } \Sigma \text{ is diagonal so } \Sigma^T = \Sigma \\ A^- &= V\Sigma^{-1}U^T \end{aligned}$$

Now we will show that the normal equation holds:

$$\begin{aligned} A^TAx &= A^TAA^-b && \text{substitute } x = A^-b \\ &= V\Sigma U^TU\Sigma V^TV\Sigma^{-1}U^Tb && \text{substitute SVD forms written above} \\ &= V\Sigma\Sigma\Sigma^{-1}U^Tb && U \text{ and } V \text{ have orthonormal columns} \\ &= V\Sigma U^Tb \\ &= A^Tb \end{aligned}$$

■

Claim 4. Any optimal solution has the form $A^-b + (I - V'V'^T)z$, where V'^T corresponds to the rows of V^T for which $\Sigma_{i,i} > 0$.

Proof. Recall from the section defining normal equations that if x^* is an optimal solution, then $x^* + y$ will also be optimal, where $y \in \text{Kernel}(A)$.

We will now show that Kernel(A) = the set of vectors $(I - V'V'^T)z$ where z is arbitrary. Recall that Kernel(A) is the set of vectors x such that $Ax = 0$. Therefore, we will show that $A(I - V'V'^T)z = 0$ for all z.

$$A(I - V'V'^T)z = U\Sigma V^T(I - V'V'^T)z$$

ΣV^T is in the rowspan of V'^T. This is because V'^T is made up of precisely the non-zero rows in V^T, so all the rows in V^T can be made of linear combinations of rows in V'^T.

$(I - V'V'^T)$ is a projection matrix. This is because V^T has orthonormal columns, so V'^T must also have orthonormal columns. Therefore, $V'V'^T$ is a projection matrix. $I - P$ is a projection matrix if P is a projection matrix. Therefore, $(I - V'V'^T)$ is a projection matrix. Furthermore, since $V'V'^T$ is a projection matrix that projects onto the subspace spanned by the rows of V'^T, we know that $(I - V'V'^T)$ projects onto the orthogonal complement of that subspace. That means it projects onto a subspace that is outside the rowspan of V'^T.

When multiplying a matrix in the rowspan of V by a projection matrix P that projects onto a subspace outside the rowspan of V, the result will be the zero matrix. Thus, we know that $\Sigma V^T(I - V'V'^T)$ is the zero matrix. Therefore, we have shown that $(I - V'V'^T)z$ is the kernel of A, so an optimal solution has the form $A^-b + (I - V'V'^T)z$. ■

Claim 5. Of all the optimal solutions, A^-b is the one with minimum norm.

Proof. The equations below use V_i to represent the ith column of matrix V.

$$
\begin{aligned}
A^- b = V\Sigma^{-1}U^T b &= \begin{bmatrix} & & & & \\ V_1 & \dots & V_r & \dots & V_n \\ & & & & \end{bmatrix} \begin{bmatrix} \frac{1}{\Sigma 11} & & & & \\ & \ddots & & & \\ & & \frac{1}{\Sigma rr} & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} U^T b \\
&= \begin{bmatrix} & & & & \\ V_1 & \dots & V_r & 0 & 0 \\ & & & & \end{bmatrix} U^T b \\
&= V' U^T b
\end{aligned}
$$

Therefore, $A^- b$ is in the column span of V'. Recall that $(I - V'V'^T)$ projects onto a subspace orthogonal to V', so $A^- b \perp (I - V'V'^T)z$. Then by the Pythagorean theorem,

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

. Therefore, all solutions have norm $\geq \|A^- b\|^2$, so it must be the solution with minimum norm. ■