DRP HW 3 - Univariate and Multivariate Linear Regression

Owen Huang

July 3, 2024

Notes

This homework rigorously examines the foundations of linear regression, which is, in some sense, the backbone of machine learning. We begin with the univariate case (ie., linear regression in the plane), then culminate in the derivation of the *normal equations* for linear regression in multiple variables.

1 Introduction and Warmup

Here is the setting for the rest of the problem set, and probably for the rest of the course. Suppose you are a machine, whose goal is to learn the relationship between d observations $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{d}$.[1] For simplicity (and generality), we might assume that each $x^{(i)}$ is a vector in $\mathbb{R}^n$ (which we think of as a list of variables called *regressors*) and $y^{(i)} \in \mathbb{R}$, which should hopefully depend on the x_i somehow. Now without further evidence, there is no reason to guess that $y^{(i)}$ and $x^{(i)}$ should have any connection, but we can do what mathematicians do best, and *assume* that they do, then see what interesting statements follow. In the simplest case (modulo the case where they are not related at all), they are related linearly. That is, for each i,

$$y^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_n x_n^{(i)}$$

The point of the homework will be for you to figure out which choice of $\beta = (\beta_0 \dots \beta_n)$ is the best. Of course, we need to define what we even mean by best, amongst other issues, such as uniqueness and computational efficiency. We will settle all of that in a second, but firstly, you will solve a warmup problem which examines when a variant of linear regression gives a "perfect" solution.

1.1 Polynomial Fitting

Consider the scenario where one has two points $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\} \subset \mathbb{R}^2$. Is it possible to define a (unique) line $\ell \subset \mathbb{R}^2$ which passes through the elements of $\mathcal{D}$? By some high school math, this is certainly possible. Let us step up one dimension to gain a little more intuition.

Exercise 1.1. Let $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$ be three distinct points in the plane such that $x_1 \neq x_2 \neq x_3$. Show that there exists a unique quadratic equation f so that $f(x_i) = y_i$ for $1 \leq i \leq 3$.

So not too hard, right? At this stage, the astute might conjecture that such a result is true for polynomials of arbitrarily high degree, and it turns out that this is indeed true. Your task is to prove the following theorem.

Theorem 1.2 (Polynomial Interpolation Theorem)**.** *Let $\{(x_0, y_0), \dots, (x_n, y_n)\}$ be $n+1$ distinct points in the plane such that $x_1 \neq \dots \neq x_n$. There exists a unique polynomial f of degree n so that $f(x_i) = y_i$ for $1 \leq i \leq n$.*

[1] By convention, superscripts with parentheses indicate the observation number. Subscripts indicate the index of the vector.

Proof. Firstly, recall what it means for this polynomial to *interpolate* the points (x_i, y_i). This means that if $f(x) = a_0 + a_1^x + \dots a_n x^n$, then

$$f(x_i) = a_0 + a_1 x_i + \cdots + a_n x_i^n = y_i \tag{1}$$

for each i.

Exercise 1.3. Equation 1 is linear in the a_i, and there are $n+1$ of them. Find a $(n+1) \times (n+1)$ matrix V so that $Va = y$, where $a = (a_0, \dots a_n)^T$ and $y = (y_0, \dots y_n)^T$.

Now, the matrix in the above exercise is called a *Vandermonde* matrix, and it has some special properties, including a formula for its determinant. Note that if V is invertible, then $a = V^{-1}y$, and since a uniquely determines the polynomial f, we will be done! Hence, we must show that V is invertible, or equivalently that it has nonvanishing determinant.

Exercise 1.4. Find a row reduction operation which makes the top rightmost entry of V equal to zero. Perform a similar operation to kill the top, second-rightmost element, and continue until all entries of the top row, except the top left entry is zero.

Hopefully you see where we are going with this.

Exercise 1.5. Using the Laplace expansion formula for the determinant, and induction, conclude that

$$\det V = \prod_{1 \le i < j \le n} (x_j - x_i) \tag{2}$$

Exercise 1.6. Why does the above formula imply the main theorem? Why must the polynomial f be unique?

□

And finally, we note that in *most* cases, no line goes through three points in the plane, no quadratic goes through 4 points, and no degree n polynomial interpolates $n+2$ points. Conversely, given n points in the plane, there are infinitely many degree n polynomials which interpolate them. Such a theorem may be difficult to prove using analytical techniques, but linear algebra can simplify the proof drastically.

Exercise 1.7. Prove the above phenomenon using dimensionality arguments on V.

2 Univariate Linear Regression

We begin by studying the simplest case with $\mathcal{D} = \{(x^{(0)}, y^{(0)}), \dots, (x^{(d)}, y^{(d)})\} \subset \mathbb{R}^2$, where we pick β_0, β_1 to fit the function $\hat{f}(x; \beta_0, \beta_1) = \beta_0 + \beta_1 x$. If we assume that the relationship between $y^{(i)}$ and $x^{(i)}$ is linear, it means that we believe the underlying "true" function should be of the form $y^{(i)} = mx^{(i)} + b$, and hence that β_0 and β_1 should approximate b and m respectively. We introduce the *loss function*

$$\mathcal{L}(\beta_0, \beta_1) = \sum_{i=1}^{d} (\hat{f}(x^{(i)}; \beta_0, \beta_1) - y^{(i)})^2 \tag{3}$$

This loss function represents how far away our prediction of m and b were, and hence a natural objective will be to minimize it. Using facts we know from standard calculus, we can derive the first order conditions

Exercise 2.1. Compute the first order equations

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = 0 \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \beta_1} = 0$$

and then solve for the optimal values of β_0, β_1. [Hint: you should have to solve a linear system of two equations in two unknowns].

Great! So you've found what is known as the *line of best fit*. It is the line ℓ in $\mathbb{R}^2$ which minimizes the sum of the vertical distances to each of the points in $\mathcal{D}$. Unfortunately, the expressions for β_0 and β_1 are quite messy. One can imagine that extending this to multiple dimensions will be even messier. We can leverage some linear algebra to make our computations a lot cleaner.

3 Multivariate Regression

To generalize your previous results, we need to extend to multiple dimensions. That is, we now consider data set $\mathcal{D} = \{(x^{(0)}, y^{(0)}), \ldots, (x^{(d)}, y^{(d)})\} \subset \mathbb{R}^{n+1}$, with $x^{(i)} \in \mathbb{R}^n$, and we are now to pick $\beta = (\beta_0, \ldots \beta_n)^T$ to fit the function $\hat{f}(x; \beta) = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n$, for $x \in \mathbb{R}^n$. We can define a loss function similar to Equation 3 by

$$\mathcal{L}(\beta) = \sum_{i=1}^{d} (\hat{f}(x^{(i)}; \beta) - y^{(i)})^2 \tag{4}$$

but note that if we adopt the same technique of computing first order conditions explicitly and solving a system of $n+1$ equations and $n+1$ unknowns will not be a tractable problem. Hence, we need to adapt to a more elegant solution.

Exercise 3.1. Let $v \in \mathbb{R}^n$. Show that $v^T v = \|v\|^2$.

Exercise 3.2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(v) = \|v\|^2$. Compute ∇f.

Now, we might not know how to take derivatives of complicated multivariate functions, but the above shows that when we restrict to just the norm function, its gradient has a simple form. Fortunately, the loss function $\mathcal{L}(\beta)$ is actually the norm of a vector!

Exercise 3.3. Show that there exists a vector $v \in \mathbb{R}^d$ so that $\mathcal{L}(\beta) = v^T v$. Using this, compute the first order condition $\nabla \mathcal{L}$ and set it to zero. Use the chain rule to simplify the gradient to derive the *normal equations* $X^T X \beta = X^T Y$.

Exercise 3.4. Let A be an $n \times m$ matrix. Show that $A^T A$ is invertible if and only if the rank of A is m.

Exercise 3.5. Assume further that $d \geq n$ (ie., the number of data points is larger than the number of regressors). Use the previous problem to argue that in most cases, there is a unique solution of β which minimizes $\mathcal{L}$. [2]

Hopefully, this problem set convinces you the power of linear algebra - both as an organizational tool and then as informative one. Next week, we will combine the multivariate regression solution with ideas from the polynomial interpolation theorem to discuss regression with respect to *basis functions*.

[2] Note that this will not be a formal proof unless one studies measure theory/ perturbations but should nonetheless be convincing.