ISTA 421 + INFO 521 Introduction to Machine Learning

Lecture 3: Linear Models & Moving to Higher dimensions

Clay Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

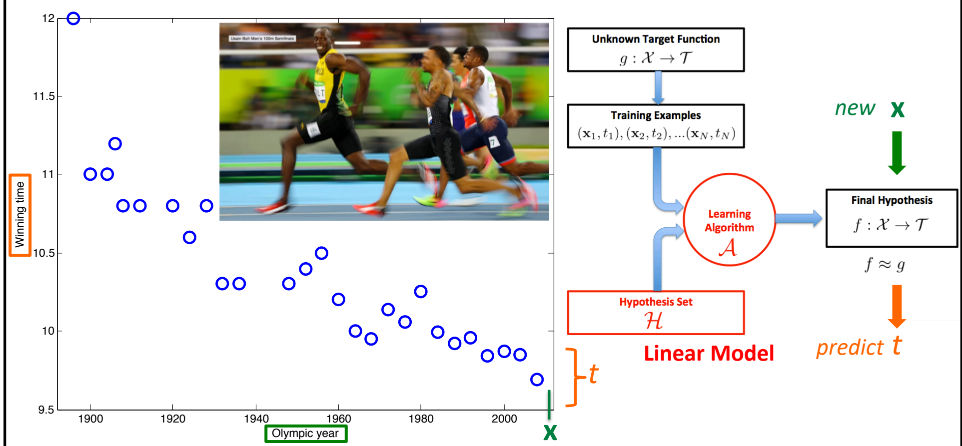
27 August 2017 1

Next Topics

- Linear Model (what is the *model*)
- Loss Function (what is a *good* model)
- Least Squares (finding the *best* model)
- Prediction
- Review LMS fit (the “normal equations”) for the single variable, 2 parameter **linear model**
- Moving to higher dimensions
  - Linear Algebra: matrix operators
  - Some Geometry of Linear Algebra
  - Least Mean Squares in Matrix formulation
  - The Geometry of LMS solution
- Nonlinear Response
- Generalization and Overfitting
- Regularized Least Squares

2

A simple learning problem



Want to describe **Winning time (t)** as a function of **Olympic year ($\mathbf{x}$)**

3

Defining a Model

- Define function that maps inputs (Olympics year, x_i) to output or target values (Winning times, t_i)

$$t = f(x)$$

- The model itself likely has parameters, which we'll generically refer to as 'θ' here. It is common to make them explicit within a function:

$$t = f(x\,;\,\theta)$$

4

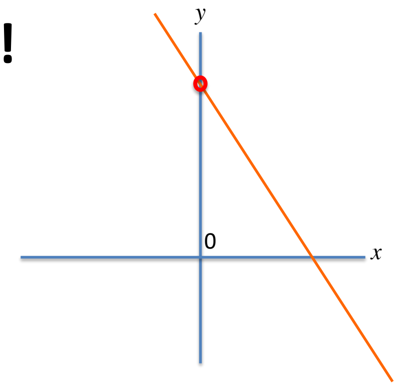
Lines!

- **Slope-intercept form**

$$y = \underline{m}x + \underline{b^*}$$

- General (standard) form

$$ax + by + c = 0$$



slope $m = -\dfrac{a}{b}$

y-intercept $b^* = -\dfrac{c}{b}$

x-intercept $= -\dfrac{c}{a}$

5

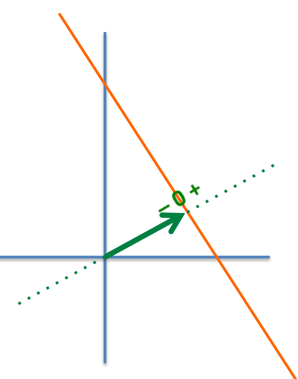
Lines!

- **General (standard) form**

$$\underline{a}x + \underline{b}y + \underline{c} = 0$$

- Slope-intercept form

$$y = \underline{m}x + \underline{b^*}$$



slope $m = -\dfrac{a}{b}$

y-intercept $b^* = -\dfrac{c}{b}$

x-intercept $= -\dfrac{c}{a}$

6

Linear Relationship

- $y = mx + b$ (or $t = w_1 x + w_0$)
  - the classic line (in 2D space)
  - For a given line, m and b are the ***parameters*** and x is a ***variable*** in the relationship: $y = f(x\ ;\ m, b)$
  - When considering alternate lines, we are adjusting m and b
- Generally, as long as the values that vary (assuming the others are constant) are not themselves involved in anything more than
  (1) addition and
  (2) scalar multiplication,
  … then the relationship is ***linear***.

Let's consider the relationship between y and x

$y = mx^2 + c$ $\quad y = \sin(x)$ $\quad \sqrt{y} = mx + c$ Not linear rel. btwn x,y

$y = mx + c^2$ $\quad y = x\sin(m) + c$ Is linear rel. btwn x,y

Linear Relationship

- $y = mx + b$ (or $t = w_1 x + w_0$)
  - the classic line (in 2D space)
  - For a given line, m and b are the ***parameters*** and x is a ***variable*** in the relationship: $y = f(x\ ;\ m, b)$
  - When considering alternate lines, we are adjusting m and b
- Generally, as long as the values that vary (assuming the others are constant) are not themselves involved in anything more than
  (1) addition and
  (2) scalar multiplication,
  … then the relationship is ***linear***.

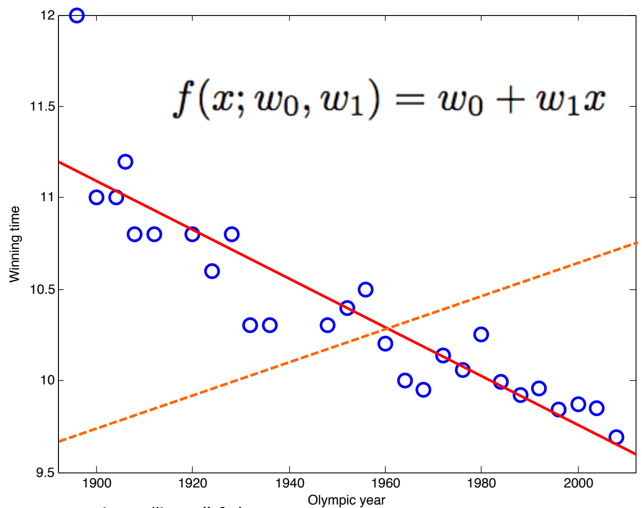
Let's consider the relationship between y and x

$y = mx^2 + c$ $\quad y = \sin(x)$ $\quad \sqrt{y} = mx + c$ Not linear rel. btwn y,m

$y = mx + c^2$ $\quad y = x\sin(m) + c$ Is linear rel. btwn y,m

What about the relationship between y and m (m is a parameter!)

Data with line (particular w_0 & w_1)



(The red line happens to be a "best" fit)
(The dashed orange line does not describe the trend in the data very well; not a good fit)

9

What is a good model?
Loss Function

$$\mathcal{L}_n(\ldots)$$

Squared Error:

$$(t_n - f(x_n; w_0, w_1))^2$$

$$\mathcal{L}_n(t_n, f(x_n; w_0, w_1)) = (t_n - f(x_n; w_0, w_1))^2$$

Mean Squared Error:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_n(t_n, f(x_n; w_0, w_1))$$

10

Goal: Find the "best" values for the parameters of model according to the loss fn

- If we're lucky (i.e., the model and the loss fn are "well-behaved") we can derive an ***analytic*** solution. Otherwise, we'll pick some (iterative) optimization method that is appropriate.
- Our first example, using a **mean squared error loss function** with a **linear model** permits a nice analytic solution!
  - Here (and in the book) we'll first look at the direct, analytic method.
  - Another method: gradient descent
    - Same loss function, but iterative algorithm and can be used in cases where we don't have an analytic solution for the parameters

11

Least Mean Squares Solution

(for single variable, 2 parameter linear model)

$$f(x; w_0, w_1) = w_0 + w_1 x$$

$$\begin{aligned}
\mathcal{L} &= \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_n(t_n, f(x_n; w_0, w_1)) \\
&= \frac{1}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1))^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}(t_n - (w_0 + w_1 x_n))^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}(w_1^2 x_n^2 + 2w_1 x_n w_0 - 2w_1 x_n t_n + w_0^2 - 2w_0 t_n + t_n^2) \\
&= \frac{1}{N}\sum_{n=1}^{N}(w_1^2 x_n^2 + 2w_1 x_n (w_0 - t_n) + w_0^2 - 2w_0 t_n + t_n^2)
\end{aligned}$$

The specific loss fn we're working with here

The specific model we're working with here

Multiply out and re-arrange to put into an easier-to deal with form.

12

Least Mean Squares Solution

(for single variable, 2 parameter linear model)

$$f(x; w_0, w_1) = w_0 + w_1 x$$ Our model family

Our loss fn

$$\mathcal{L} \;=\; \frac{1}{N}\sum_{n=1}^{N}(w_1^2 x_n^2 + 2w_1 x_n(w_0 - t_n) + w_0^2 - 2w_0 t_n + t_n^2)$$

Our goal: We want values for w_0 and w_1 that will ***minimize*** this loss function

I.e., we seek values for w_0 and w_1 that will make the loss function be the smallest when we actually sum over all the values of x and t in the dataset.

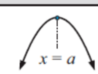
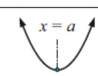
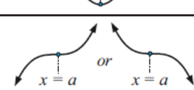
Because the loss function happens to be quadratic (in the two parameters) we can use a standard method from calculus for finding minima (maxima) directly: taking the derivative of the function and setting it to zero.

Our loss function has **two** parameters that we're trying set to minimize the loss fn, so we need to take the partial derivative (w.r.t. w_0 and w_1)

What we end up with are two functions, one for w_0 and one for w_1, and both will work with **any** data and give the best **least mean square** (LMS) fit!

Side note: Calculus Tools

First and Second Derivative
around particular x value (*x*=*a*)

Stationary point	*Sign diagram of $f'(x)$ near $x = a$*	*Shape of curve near $x = a$*
local maximum	+ | − (a)	
local minimum	− | + (a)	
horizontal inflection	+ | + (a) *or* − | − (a)	

Second derivative f''(x)

negative
(const if *f* is quadratic)

positive
(const if *f* is quadratic)

0 at inflection point
No longer a constant fn
(if *f* higher order than quad)

Least Mean Squares Solution

(for single variable, 2 parameter linear model)

$$\mathcal{L} \quad = \quad \frac{1}{N}\sum_{n=1}^{N}(w_1^2x_n^2 + 2w_1x_n(w_0 - t_n) + w_0^2 - 2w_0t_n + t_n^2)$$

- Partial derivative for w_0

First, since we're taking the partial w.r.t. w_0, can drop any terms without w_0.

$$\frac{1}{N}\sum_{n=1}^{N}\left[w_0^2 + 2w_1x_nw_0 - 2w_0t_n\right]$$

Next, move sums inward to put in easier form

$$w_0^2 + 2w_0w_1\frac{1}{N}\left(\sum_{n=1}^{N}x_n\right) - 2w_0\frac{1}{N}\left(\sum_{n=1}^{N}t_n\right)$$

Finally, take deriv. w.r.t. w_0

$$\frac{\partial\mathcal{L}}{\partial w_0} = 2w_0 + 2w_1\frac{1}{N}\left(\sum_{n=1}^{N}x_n\right) - \frac{2}{N}\left(\sum_{n=1}^{N}t_n\right)$$

15

Continued…

- Solve for $\frac{\partial\mathcal{L}}{\partial w_0} = 0$

$$2w_0 + 2w_1\frac{1}{N}\left(\sum_{n=1}^{N}x_n\right) - \frac{2}{N}\left(\sum_{n=1}^{N}t_n\right) = 0$$

$$2w_0 = \frac{2}{N}\left(\sum_{n=1}^{N}t_n\right) - w_1\frac{2}{N}\left(\sum_{n=1}^{N}x_n\right)$$

$$w_0 = \frac{1}{N}\left(\sum_{n=1}^{N}t_n\right) - w_1\frac{1}{N}\left(\sum_{n=1}^{N}x_n\right)$$

16

Least Mean Squares Solution

(for single variable, 2 parameter linear model)

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(w_1^2x_n^2 + 2w_1x_n(w_0 - t_n) + w_0^2 - 2w_0t_n + t_n^2)$$

- Partial derivative for w_1

Do the same for w_1 …

$$\frac{1}{N}\sum_{n=1}^{N}\left[w_1^2x_n^2 + 2w_1x_nw_0 - 2w_1x_nt_n\right]$$ only keep terms with w_1

$$w_1^2\frac{1}{N}\left(\sum_{n=1}^{N}x_n^2\right) + 2w_1\frac{1}{N}\left(\sum_{n=1}^{N}x_n(w_0 - t_n)\right)$$ move sums inside

$$\frac{\partial\mathcal{L}}{\partial w_1} = 2w_1\frac{1}{N}\left(\sum_{n=1}^{N}x_n^2\right) + \frac{2}{N}\left(\sum_{n=1}^{N}x_n(w_0 - t_n)\right)$$ now take partial derivative

$$= w_1\frac{2}{N}\left(\sum_{n=1}^{N}x_n^2\right) + \frac{2}{N}\left(\sum_{n=1}^{N}x_n\left(\overbrace{\bar{t} - w_1\bar{x}}^{\text{solution for } w_0} - t_n\right)\right)$$ plug in solution for w_0 … and rearrange terms

$$= w_1\frac{2}{N}\left(\sum_{n=1}^{N}x_n^2\right) + \bar{t}\frac{2}{N}\left(\sum_{n=1}^{N}x_n\right) - w_1\bar{x}\frac{2}{N}\left(\sum_{n=1}^{N}x_n\right) - \frac{2}{N}\left(\sum_{n=1}^{N}x_nt_n\right)$$

17

- Partial derivative for w_1 continued…

$$\frac{\partial\mathcal{L}}{\partial w_1} = w_1\frac{2}{N}\left(\sum_{n=1}^{N}x_n^2\right) + \bar{t}\frac{2}{N}\left(\sum_{n=1}^{N}x_n\right) - w_1\bar{x}\frac{2}{N}\left(\sum_{n=1}^{N}x_n\right) - \frac{2}{N}\left(\sum_{n=1}^{N}x_nt_n\right)$$

$$= 2w_1\left[\left(\frac{1}{N}\sum_{n=1}^{N}x_n^2\right) - \bar{x}\ \bar{x}\right] + 2\bar{t}\bar{x} - 2\frac{1}{N}\left(\sum_{n=1}^{N}x_nt_n\right)$$ replace remaining mean x with $\overline{x}$ and group w_1 terms

$$2w_1\left[\left(\frac{1}{N}\sum_{n=1}^{N}x_n^2\right) - \bar{x}\ \bar{x}\right] + 2\bar{t}\bar{x} - 2\frac{1}{N}\left(\sum_{n=1}^{N}x_nt_n\right) = 0$$ Solve for w_1 with $\frac{\partial\mathcal{L}}{\partial w_1} = 0$ …

$$2w_1\left[\left(\frac{1}{N}\sum_{n=1}^{N}x_n^2\right) - \bar{x}\ \bar{x}\right] = 2\frac{1}{N}\left(\sum_{n=1}^{N}x_nt_n\right) - 2\bar{t}\bar{x}$$

$$w_1 = \frac{\frac{1}{N}\left(\sum_{n=1}^{N}x_nt_n\right) - \bar{t}\bar{x}}{\left(\frac{1}{N}\sum_{n=1}^{N}x_n^2\right) - \bar{x}\bar{x}} = \frac{\left(\frac{1}{N}\sum_{n=1}^{N}x_nt_n\right) - \left(\frac{1}{N}\sum_{m=1}^{N}t_n\right)\left(\frac{1}{N}\sum_{m=1}^{N}x_n\right)}{\left(\frac{1}{N}\sum_{n=1}^{N}x_n^2\right) - \left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)^2} = \frac{\overline{xt} - \bar{x}\bar{t}}{\overline{x^2} - (\bar{x})^2}$$

18

Solving LMS: Method 1 (analytic)

(for single variable, 2 parameter linear model)

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(w_1^2x_n^2 + 2w_1x_n(w_0 - t_n) + w_0^2 - 2w_0t_n + t_n^2)$$

- Partial derivative for w_0:

$$\frac{\partial \mathcal{L}}{\partial w_0} = 2w_0 + 2w_1\frac{1}{N}\left(\sum_{n=1}^{N}x_n\right) - \frac{2}{N}\left(\sum_{n=1}^{N}t_n\right)$$

- Set $\frac{\partial \mathcal{L}}{\partial w_0} = 0$ and solve for w_0:

$$w_0 = \frac{1}{N}\left(\sum_{n=1}^{N}t_n\right) - w_1\frac{1}{N}\left(\sum_{n=1}^{N}x_n\right) = \boxed{\bar{t} - w_1\bar{x}}$$

- Partial derivative for w_1:

$$\frac{\partial \mathcal{L}}{\partial w_1} = 2w_1\frac{1}{N}\left(\sum_{n=1}^{N}x_n^2\right) + \frac{2}{N}\left(\sum_{n=1}^{N}x_n(w_0 - t_n)\right)$$

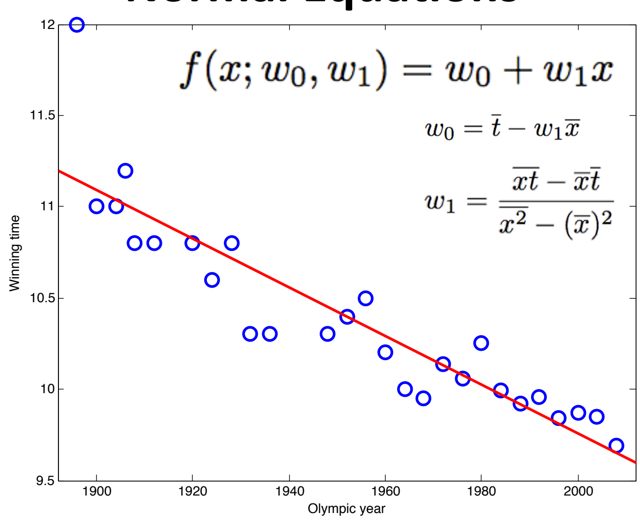
- Plug in w_0, set $\frac{\partial \mathcal{L}}{\partial w_1} = 0$ and solve for w_1:

$$w_1 = \frac{\left(\frac{1}{N}\sum_{n=1}^{N}x_nt_n\right) - \left(\frac{1}{N}\sum_{m=1}^{N}t_n\right)\left(\frac{1}{N}\sum_{m=1}^{N}x_n\right)}{\left(\frac{1}{N}\sum_{n=1}^{N}x_n^2\right) - \left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)^2} = \boxed{\frac{\overline{xt} - \bar{x}\bar{t}}{\overline{x^2} - (\bar{x})^2}}$$

The so-called **normal equations**!

19

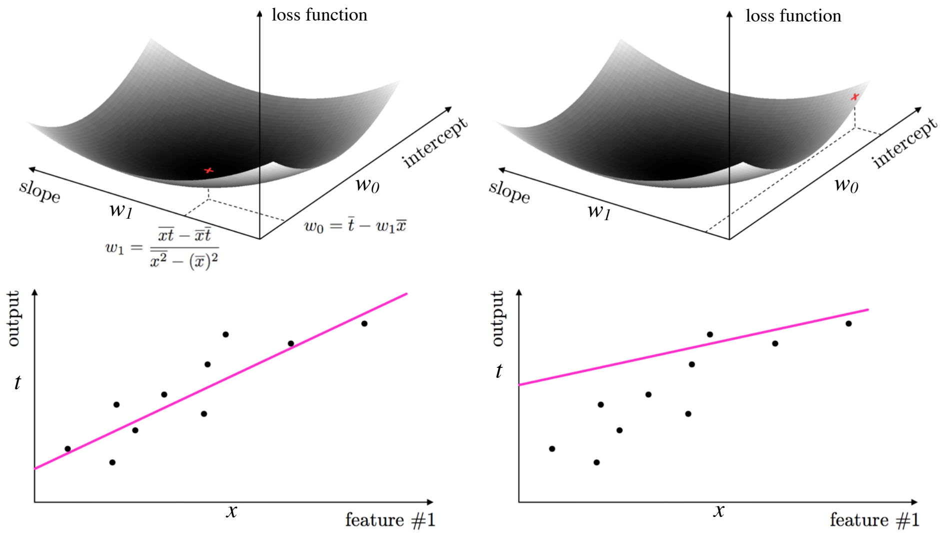
Linear Least Mean Square Normal Equations



20

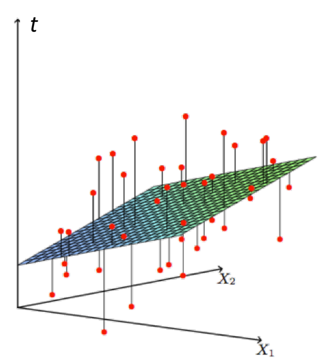
Visualizing the 2-dim loss fn surface

$f(x; w_0, w_1) = w_0 + w_1 x \qquad \mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}(w_1^2 x_n^2 + 2w_1 x_n(w_0 - t_n) + w_0^2 - 2w_0 t_n + t_n^2)$



How about more than 1 input?

- Most problems will involve more than just the relationship between 1 input attribute and a target.
- Extending our linear models to higher dimensions is desirable. For 2 inputs it is easy to visualize the geometry: now the "line" is a plane in 3D
- In general, a (regression) linear model with ***n*** input variables and ***n***+1 parameters (the ***w***'s, with their values determined) is an ***n***-dimensional "**hyperplane**" embedded in ***n***+1 dimensions.



22

Things quickly get messy as we increase the dimensions…

- Suppose we want a richer predictive model for the Olympic data: not only include the best overall time for the gold, but also the personal-best times of each sprinter that raced ($s_1, \ldots, s_8$)
- This is a 9 dimensional hyperplane with 10 parameters:

$$\begin{aligned} t = f(x, s_1, ..., s_8; w_0, ..., w_9) &= w_0 + w_1 x + w_2 s_1 + w_3 s_2 \\ &+ w_4 s_3 + w_5 s_4 + w_6 s_5 \\ &+ w_7 s_6 + w_8 s_7 + w_9 s_8 \end{aligned}$$

- The math is fundamentally the same, but to derive the normal equations, we need to take 10 partial derivatives, then have 10 equations to re-arrange and substitute back in…

23