1.3 Vector / Matrix Notation

Add more parameters:

$t = f(x, s_1, \ldots, s_8 : w_0, \ldots, w_9) =$

$w_0 + w_1 x + w_2 s_1 + w_3 s_2 \ldots$

As you add more paras the number of equations and partial derivatives becomes unfeasibly large

the solution is to use vectors & matricies

X^T = vector or matrix transpose

it is easier to read as a row so the transpose sign is used to tell us that a row is actually a col vect

$X_n = [X_{n1}, \ldots, X_{nD}]^T$

Recall matrix is called rows x cols

Converting $t = w_0 + w_1 x$ to vector form

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \qquad X_n = \begin{bmatrix} 1 \\ x_n \end{bmatrix}$$

Parameter vector Data vector

$$f(x_n ; w_0, w_1) = w^T x_n = w_0 + w_1 x_n$$

matrix/vector multiplication

Any instance of $w_0 + w_1 x_n$ can be replaced with $w^T x_n$

$$Sq\,loss = \frac{1}{N} \sum (t_n - \underline{\underline{w^T x_n}})^2$$

$$\hookrightarrow \frac{1}{N} (t - Xw)^T (t - Xw)$$

How does this notation work?

first collect all X_n (input) & t_n (output) into vectors

$$X = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_N^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_3 \end{bmatrix}$$

Perform MM on X_W $\quad -t$

$$X_W = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_3 \end{bmatrix} \times \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_N \end{bmatrix} \begin{matrix} -t_1 \\ -t_2 \\ -t_n \end{matrix}$$

Note that this summises the $t_n - w_0 - w_1 x_n$ component and also encapsulates the $\hat{\Sigma}$ part (?) wrong needs matrix for this

to capture the squared and sum part of the loss function MM $(X_W - t)$ with itself

$$t - X_W = t_1 - w_0 - w_1 x_1$$
$$(t - X_W)^2 = (t_1 - w_0 - w_1 x_1)^2$$

↳ $(t - X_W)^T (t - X_W)$

creates a pairwise combination of $(t_n - X w_n)$ against itself which squares it and MM sums it also

Hence $L = \frac{1}{N} (t - X_W)^T (t - X_W)$

Factorize this down:

$$L = \frac{1}{N}(Xw - t)^T(Xw - t)$$

$$= ((Xw)^T - t^T)(Xw - t)$$

$$= \frac{1}{N}(Xw)^T Xw - \frac{1}{N}t^T Xw - \frac{1}{N}(Xw)^T t + \frac{1}{N}t^T t$$

$$= \frac{1}{N}w^T X^T Xw - \frac{2}{N}w^T X^T t + \frac{1}{N}t^T t$$

Differentiating loss in vector/matrix form

We want to work out the value of the w vector (w_0, w_1) at the turning point (min)

take partial Derivative of L with respect to w vector

take PD of L by each element of w & stack into a vector

$$\frac{\partial L}{\partial w} = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \end{bmatrix}$$

recall we calced both of these PD's in the non-matrix example

Start with the multiplied out expression from above

$$L = \frac{1}{N}(w^T X^T X w - 2 w^T X^T t + t^T t)$$

We need multiply out these terms which is a long & involved process of transposing and matrix multiplication

note $t^T t$ term is ignore as doesn't include w

result:

$$\underbrace{w_0^2 + 2 w_0 w_1 \frac{1}{N}\left(\sum x_n\right) + w_1^2 \frac{1}{N}\left(\sum x_n^2\right)}_{w^T X^T X w} + \underbrace{2 w_0 \frac{1}{N}\left(\sum t_n\right) + 2 w_1 \frac{1}{N}\left(\sum x_n t_n\right)}_{2 w^T X^T t}$$

recall $w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

note the Σ's hold shorthand averages that we went over in the previous section that are relevant for calculating linear models, i.e. $\frac{1}{N}\sum x_n = \bar{x} = \text{avg } x$

This allows us to write the expression much more concise

Partial Dij this equation =

$$\frac{\partial L}{\partial w_0} = 2w_0 + 2w_1 \bar{x} - 2\bar{t}$$

$$\frac{\partial L}{\partial w_1} = 2w_0 \bar{x} + 2w_1 \overline{x^2} - 2\overline{xt}$$

Here instead of setting the PD's to 0 and solving to find the optimal w_0, w_1, the author highlights how we can stay in the vectorized land using identities:

recall the vectorized loss function:

$$L = \frac{1}{N}(t - Xw)^T (t - Xw)$$

~~or $L = \frac{1}{2D} * (t - Xw)^T (t - Xw) \|Xw - t\|^2$~~

Expand this expression out using the properties of transposes & matrix multi (long process)

then you take the partial derivative in Matrix form using the identities that the author provids

How To Partial Diff vectors/Matrix

Expanded & Simplified loss function:

$$L = \frac{1}{N} * (\underbrace{t^T t}_{t1} - \underbrace{2w^T X^T t}_{t2} + \underbrace{w^T X^T X w}_{t3})$$

useful identities when differentiating w/ resp to a vec

$F(w)$	$\partial f / \partial w$
$W^T X$	X
$X^T W$	X
$W^T W$	$2W$
$W^T C W$	$2CW$

term 1: $t^T t$, no w vector. so deriv = 0

term 2: $-2w^T X^T t$ rewrite as $2(X^T t)^T w$

- Treat $(X^T t)$ as a constant as it has no reference to $W = x$
- now, $-2 W^T X$
- using identity rule, $w^T x = x$
  why? in calculus deriv of $ax = a$ where a is a constant (Power rule)
- Now, $-2X$

- However this result is not in the format we want. We are trying to optimise w to minimize the loss which depends on X & t
- recall that in step 1 $x = X^T t$. we can sub this back in
- So, $-2x = -2X^T t$

term 3: $w^T X^T X w$

- $X^T X = C$ where C = Symmetric Matrix
- ~~$w^T X^T X w = w^T C w$~~
- $X^T X$ creates a square matrix as $X^T = p \times n$ and $X = n \times p$
- Symetric matrix = its transpose = itself. the proof that something is symetric is confusing so i wont write it here
- Now, $t3 = w^T C w$
- Next identity says $w^T C w = 2Cw$
  the reason why is do with the quality of the symetric matrix which doubles cuz the summation due to the product rule: $uv = u'v + vu'$
- To bring back the terms $2Cw = 2X^T X w$

Putting this differentiating together:

Vector loss: $\frac{1}{N}(t - Xw)^T (t - Xw)$

Expand & Simp: $\frac{1}{N}(t^T t - 2w^T X^T t + w^T X^T Xw)$

Partial Diff w: $\frac{1}{N}(0 - 2X^T t + 2X^T Xw)$

Reorder & Simplify: $\frac{2}{N} X^T Xw - \frac{2}{N} X^T t$

Next, set the PD to zero & solve for w

Steps:
- Multiply both sides by $\frac{2}{N}$
- Isolate $w = + X^T t$ both sides

① $\frac{2}{N} X^T Xw - \frac{2}{N} X^T t = 0$

② $X^T Xw - X^T t = 0$

③ $X^T Xw = X^T t$

Next use inverse matrix to remove $X^T X$.
note dividing isn't defined for matricies

$(X^T Xw)^{-1} X^T Xw$ = Identity Matrix = $\mathbf{I}$

④ $Iw = (X^T X)^{-1} X^T t$

opt value of $\hat{w} = (X^T X)^{-1} X^T t$

Numberical Example using Small Data

n	x_n	t_n	$x_n t_n$	x_n^2
1	1	4.8	4.8	1
2	3	11.3	33.9	9
3	5	17.2	86	25
$\frac{1}{N}\sum_{n=1}^{n}$	3	11.1	41.57	11.67

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{bmatrix} \qquad t = \begin{bmatrix} 4.8 \\ 11.3 \\ 17.2 \end{bmatrix}$$

① Calculate $X^T X$

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 3 & 9 \\ 9 & 35 \end{bmatrix}$$

② calaulate this inverse $(X^T X)^{-1}$

Formula $= \frac{1}{ad-bc}\begin{bmatrix} a & -b \\ -c & a \end{bmatrix}$

$$= \frac{1}{24}\begin{bmatrix} 35 & -9 \\ -9 & 3 \end{bmatrix}$$

③ Multiply inverse by $X^T = (X^T X)^{-1} X^T$

$$\frac{1}{24}\begin{bmatrix} 35 & -9 \\ -9 & 3 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{bmatrix} = \frac{1}{24}\begin{bmatrix} 26 & 8 & -10 \\ -6 & 0 & 6 \end{bmatrix}$$

(4) finally multiply by t: $(X^T X)^{-1} X^T t$

$$\frac{1}{24}\begin{bmatrix} 26 & 8 & -10 \\ -6 & 0 & 6 \end{bmatrix} \times \begin{bmatrix} 4.8 \\ 11.3 \\ 17.2 \end{bmatrix} = \begin{bmatrix} 1.8 \\ 3.1 \end{bmatrix}$$

(5) there the formula is:

$$f(x; w_0, w_1) = 1.8 + 3.1x$$

Note the textbook page 25
Also walks through an example using only mathematical notation

Making predictions

$$t_{new} = \hat{W}^T X_{new} \quad \text{where } \hat{w} = \begin{bmatrix} 1.8 \\ 3.1 \end{bmatrix}$$