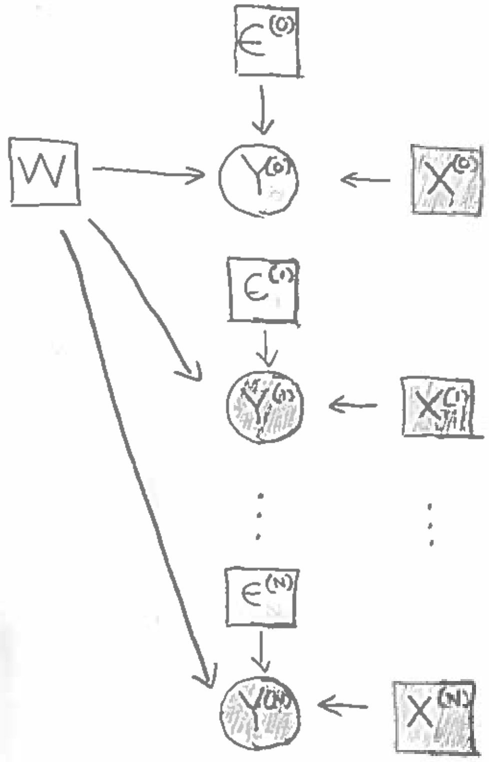
LINEAR REGRESSION: MLE

① Recall "ordinary linear regression":



where: $P_\epsilon(\epsilon^{(n)}) \sim \text{Normal}(0, \sigma^2) \quad \forall n \in \{0, \ldots, N\}$

$y^{(n)} \leftarrow w^T x^{(n)} + \epsilon^{(n)}$

② Also recall that one way to estimate the value of the unobserved response variable $Y^{(0)}$ is through <u>maximum likelihood estimation</u> (MLE):

(a) compute $\hat{w} = \underset{w}{\operatorname{argmax}} \prod_{n=1}^{N} P(y^{(n)} | w, x^{(n)})$

(b) compute $\hat{y}^{(0)} = \underset{y^{(0)}}{\operatorname{argmax}} P(y^{(0)} | \hat{w}, x^{(0)})$

LINEAR REGRESSION : MLE

③ The second step is not too bad:

$$P(y^{(0)} | \hat{w}, x^{(0)}) = \int P(y^{(0)}, \epsilon^{(0)} | \hat{w}, x^{(0)})\, d\epsilon^{(0)}$$

[Total Probability]

$$= \int P(\epsilon^{(0)} | \hat{w}, x^{(0)}) P(y^{(0)} | \epsilon^{(0)}, \hat{w}, x^{(0)})\, d\epsilon^{(0)}$$

[Chain Rule]

$$= \int P(\epsilon^{(0)}) P(y^{(0)} | \epsilon^{(0)}, \hat{w}, x^{(0)})\, d\epsilon^{(0)}$$

[d-sep.]

$$= P(\epsilon^{(0)} = y^{(0)} - \hat{w}^T x^{(0)})$$

[this is the only value of $\epsilon^{(0)}$ s.t. $P(y^{(0)} | \epsilon^{(0)}, \hat{w}, x^{(0)}) \neq 0$]

$$= P_\epsilon(y^{(0)} - \hat{w}^T x^{(0)})$$

Therefore:

$$\hat{y}^{(0)} = \underset{y^{(0)}}{\operatorname{argmax}}\, P(y^{(0)} | \hat{w}, x^{(0)})$$

$$= \underset{y^{(0)}}{\operatorname{argmax}}\, P_\epsilon(y^{(0)} - \hat{w}^T x^{(0)})$$

Since $P_\epsilon \sim \text{Normal}(0, \sigma^2)$, therefore $P_\epsilon(y^{(0)} - \hat{w}^T x^{(0)})$ is maximized when $y^{(0)} - \hat{w}^T x^{(0)} = 0$, thus:

$$\boxed{\hat{y}^{(0)} = \hat{w}^T x^{(0)}}$$

LINEAR REGRESSION: MLE

4) How do we compute the first step? First let's turn those annoying products into friendly sums:

$$\operatorname*{argmax}_{w} \prod_{n=1}^{N} P(y^{(n)} \mid w, x^{(n)})$$

$$= \operatorname*{argmax}_{w} \log \prod_{n=1}^{N} P(y^{(n)} \mid w, x^{(n)})$$

$$= \operatorname*{argmax}_{w} \underbrace{\sum_{n=1}^{N} \log P(y^{(n)} \mid w, x^{(n)})}_{\text{let's call this } \ell(w)}$$

5) Next, let's do some manipulations of $\ell(w)$.

$$\ell(w) = \sum_{n=1}^{N} \log P(y^{(n)} \mid w, x^{(n)})$$

$$= \sum_{n=1}^{N} \log P_{\epsilon}(y^{(n)} - \hat{w}^{T} x^{(n)}) \qquad [\text{from } 3]$$

$$= \sum_{n=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(\frac{-1}{2\sigma^2} (y^{(n)} - w^{T} x^{(n)})^2 \right) \right]$$

$$= \sum_{n=1}^{N} \log \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} + \log \exp\left(\frac{-1}{2\sigma^2} (y^{(n)} - w^{T} x^{(n)})^2 \right)$$

$$= \sum_{n=1}^{N} \frac{-1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} (y^{(n)} - w^{T} x^{(n)})^2$$

$$= \left(\sum_{n=1}^{N} \frac{-1}{2} \log 2\pi\sigma^2 \right) - \left(\sum_{n=1}^{N} \frac{1}{2\sigma^2} (y^{(n)} - w^{T} x^{(n)})^2 \right)$$

$$= \frac{-N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y^{(n)} - w^{T} x^{(n)})^2$$

6) Thus:

$$\underset{w}{\operatorname{argmax}}\ \ell(w) = \underset{w}{\operatorname{argmax}}\left(\frac{-N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y^{(n)} - w^T x^{(n)}\right)^2\right)$$

$$= \underset{w}{\operatorname{argmax}}\ -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y^{(n)} - w^T x^{(n)}\right)^2$$

$$= \underset{w}{\operatorname{argmax}}\ -\sum_{n=1}^{N}\left(y^{(n)} - w^T x^{(n)}\right)^2$$

7) We can express $\sum_{n=1}^{N}\left(y^{(n)} - w^T x^{(n)}\right)^2$ without the explicit summation by resorting to vector dot product:

$$\sum_{n=1}^{N}\left(y^{(n)} - w^T x^{(n)}\right)^2 = \begin{bmatrix} y^{(1)} - w^T x^{(1)} \\ \vdots \\ y^{(N)} - w^T x^{(N)} \end{bmatrix}^T \begin{bmatrix} y^{(1)} - w^T x^{(1)} \\ \vdots \\ y^{(N)} - w^T x^{(N)} \end{bmatrix}$$

and noticing that:

$$\begin{bmatrix} y^{(1)} - w^T x^{(1)} \\ \vdots \\ y^{(N)} - w^T x^{(N)} \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix} - \begin{bmatrix} w^T x^{(1)} \\ \vdots \\ w^T x^{(N)} \end{bmatrix}$$

$$= y - \begin{bmatrix} w_1 x_1^{(1)} + \dots + w_D x_D^{(1)} \\ \\ w_1 x_1^{(N)} + \dots + w_D x_D^{(N)} \end{bmatrix}$$

$$= y - \begin{bmatrix} x_1^{(1)} & \cdots & x_D^{(1)} \\ \\ x_1^{(N)} & \cdots & x_D^{(N)} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_D \end{bmatrix}$$

$$= y - Xw$$

⑧ So
$$\underset{w}{\operatorname{argmax}}\; \ell(w) = \underset{w}{\operatorname{argmax}} -(y-Xw)^T(y-Xw)$$
$$= \underset{w}{\operatorname{argmax}} -\cancel{y^Ty} + y^TXw + (Xw)^Ty - (Xw)^TXw$$
$$= \underset{w}{\operatorname{argmax}} ((Xw)^Ty)^T + (w^TX^T)y - (w^TX^T)Xw$$
[since $(AB)^T = B^TA^T$]
$$= \underset{w}{\operatorname{argmax}} (w^TX^Ty)^T + w^TX^Ty - w^TX^TXw$$
[since $(AB)^T = B^TA^T$]

Notice that w^TX^Ty is a 1×1 matrix (i.e. $(1\times D)\cdot(D\times N)\cdot(N\times 1)$), so $(w^TX^Ty)^T = w^TX^Ty$. That gives us:
$$\underset{w}{\operatorname{argmax}}\; \ell(w) = \underset{w}{\operatorname{argmax}}\; 2w^TX^Ty - w^TX^TXw$$

---

⑨ At this point, we're pretty close. We've shown (over ④–⑧) that the point estimate $\hat{w}$ of our weight vector is:
$$\hat{w} = \underset{w}{\operatorname{argmax}} \prod_{n=1}^{N} P(y^{(n)} \mid w, x^{(n)})$$
$$= \underset{w}{\operatorname{argmax}}\; 2w^TX^Ty - w^TX^TXw$$
$$= \underset{w}{\operatorname{argmin}}\; w^TX^TXw - 2wX^Ty$$
So the loss function for ordinary linear regression is:
$$L_{lin}(w) = w^TX^TXw - 2wX^Ty$$

LINEAR REGRESSION: MLE

(10) We can use the identities

$$\boxed{\frac{\partial}{\partial a} a^T b = \frac{\partial}{\partial a} b^T a = b} \quad \text{and} \quad \boxed{\frac{\partial}{\partial a} a^T X a = (X + X^T) a}$$

to compute the gradient of $L_{lin}(w)$

$$\frac{\partial}{\partial w}\left(-2w^T X^T y + w^T X^T X w\right)$$

$$= \frac{\partial}{\partial w} -2w^T X^T y + \frac{\partial}{\partial w} w^T X^T X w$$

$$= -2X^T y + \frac{\partial}{\partial w} w^T (X^T X) w \qquad \left[\text{b/c } \frac{\partial}{\partial a} a^T b = b\right]$$

$$= -2X^T y + \left(X^T X + (X^T X)^T\right) w \qquad \left[\text{b/c } \frac{\partial}{\partial a} a^T X a = (X + X^T) a\right]$$

$$= -2X^T y + 2X^T X w \qquad \left[\text{b/c } (AB)^T = B^T A^T\right]$$

---

(11) We can then compute $\underset{w}{\operatorname{argmax}} \prod_{n=1}^{N} P(y_n \mid w, x_n)$ by finding when the gradient equals zero:

$$-2X^T y + 2X^T X w = 0$$

$$\Rightarrow \quad X^T X w = X^T y$$

$$\Rightarrow \quad w = (X^T X)^{-1} X^T y$$

So we have our answer:

$$\boxed{\underset{w}{\operatorname{argmin}}\ L_{lin}(w) = (X^T X)^{-1} X^T y}$$

(12) So we can now go back to (2) and make our MLE algorithm more concrete:

i.e.

(a) compute $\hat{w} = \arg\max_{w} \prod_{n=1}^{N} P(y_n \mid w, x_n)$ | $\hat{w} \leftarrow (X^T X)^{-1} X^T y$

(b) compute $\hat{y}_0 = \arg\max_{y_0} P(y_0 \mid \hat{w}, x_0)$ | $\hat{y}_0 \leftarrow \hat{w}^T x_0$