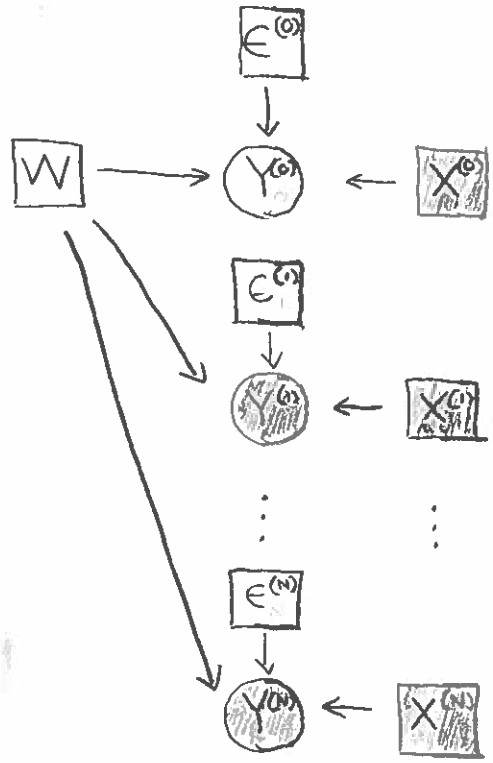
Linear Regression: MLE

① Recall "ordinary linear regression":



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

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

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

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

LINEAR REGRESSION: MLE

③ The second step is not too bad:

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

[Total Probability]

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

[Chain Rule]

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

[d-sep.]

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

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

$$= \overset{P_\epsilon}{\varphi}(y^{(0)} - x^{(0)}\hat{\omega})$$

Therefore:

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

$$= \underset{y^{(0)}}{\operatorname{argmax}}\, \varphi(y^{(0)} - x^{(0)}\hat{\omega})$$

Since $\varphi \propto \text{Normal}(0,\sigma^2)$, therefore $\varphi(y^{(0)} - x^{(0)}\hat{\omega})$ is maximized when $y^{(0)} - x^{(0)}\hat{\omega} = 0$, thus:

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

Linear Regression: MLE

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

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

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

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

---

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

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

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

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

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

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

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

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

Linear Regression: MLE

⑥ Thus:

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

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

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

---

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

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

and noticing that:

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

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

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

$$= y - Xw$$

LINEAR REGRESSION : MLE

(8) So
$$\underset{w}{\operatorname{argmax}}\ \ell(w) = \underset{w}{\operatorname{argmax}} -(y - Xw)^T(y - Xw)$$
$$= \underset{w}{\operatorname{argmax}} -\cancel{y^T y} + y^T X w + (Xw)^T y - (Xw)^T X w$$
$$= \underset{w}{\operatorname{argmax}} \left((Xw)^T y\right)^T + (w^T X^T) y - (w^T X^T) X w$$
[since $(AB)^T = B^T A^T$]
$$= \underset{w}{\operatorname{argmax}} (w^T X^T y)^T + w^T X^T y - w^T X^T X w$$
[since $(AB)^T = B^T A^T$]

Notice that $w^T X^T y$ is a 1×1 matrix (i.e. $(1 \times D) \cdot (D \times N) \cdot (N \times 1)$), so $(w^T X^T y)^T = w^T X^T y$. That gives us:

$$\underset{w}{\operatorname{argmax}}\ \ell(w) = \underset{w}{\operatorname{argmax}}\ 2w^T X^T y - w^T X^T X w$$

---

(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:

$$\hat{w} = \underset{w}{\operatorname{argmax}} \prod_{n=1}^{N} P(y^{(n)} | w, x^{(n)})$$
$$= \underset{w}{\operatorname{argmax}}\ 2w^T X^T y - w^T X^T X w$$
$$= \underset{w}{\operatorname{argmin}}\ w^T X^T X w - 2w^T X^T y$$

So the loss function for ordinary linear regression is:

$$L_{lin}(w) = w^T X^T X w - 2w^T X^T y$$

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 $\underset{w}{\text{argmax}} \prod_{n=1}^{N} P(y_n \mid w, x_n)$ 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:

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

LINEAR REGRESSION: MLE

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

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

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

i.e.

$\hat{w} \leftarrow (X^T X)^{-1} X^T y$

$\hat{y}^{(0)} \leftarrow x^{(0)} \hat{w}$