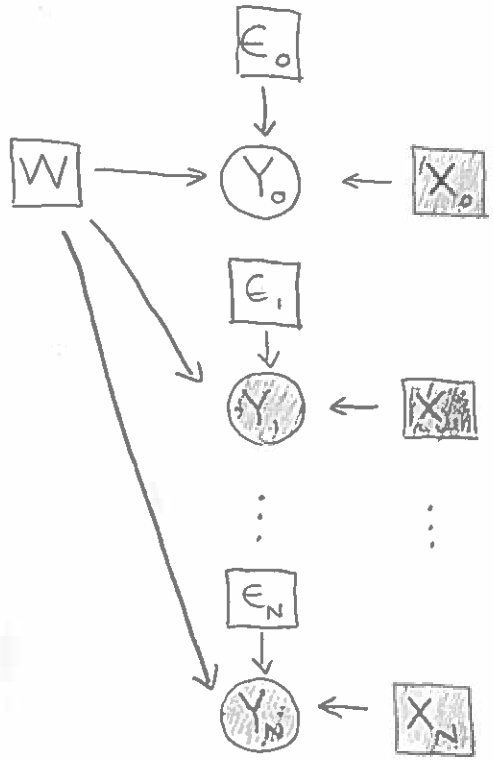
LINEAR REGRESSION: MLE

① Recall "ordinary linear regression":



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

$y_n \leftarrow w^T x_n + \epsilon_n$

② Also recall that one way to estimate the value of the unobserved response variable Y_0 is through maximum likelihood estimation (MLE):

(a) compute $\hat{w} = \underset{w}{\arg\max} \prod_{n=1}^{N} P(y_n | w, x_n)$

(b) compute $\hat{y}_0 = \underset{y_0}{\arg\max} P(y_0 | \hat{w}, x_0)$

LINEAR REGRESSION : MLE

③ The second step is not too bad:

$$P(y_0 \mid \hat{w}, x_0) = \int P(y_0, \epsilon_0 \mid \hat{w}, x_0)\, d\epsilon_0$$

[Total Probability]

$$= \int P(\epsilon_0 \mid \hat{w}, x_0) P(y_0 \mid \epsilon_0, \hat{w}, x_0)\, d\epsilon_0$$

[Chain Rule]

$$= \int P(\epsilon_0) P(y_0 \mid \epsilon_0, \hat{w}, x_0)\, d\epsilon_0$$

[d-sep.]

$$= P(\epsilon_0 = y_0 - \hat{w}^T x_0)$$

[this is the only value of ϵ_0 s.t. $P(y_0 \mid \epsilon_0, \hat{w}, x_0) \neq 0$]

$$= P_\epsilon(y_0 - \hat{w}^T x_0)$$

Therefore:

$$\hat{y}_0 = \operatorname*{argmax}_{y_0} P(y_0 \mid \hat{w}, x_0)$$

$$= \operatorname*{argmax}_{y_0} P_\epsilon(y_0 - \hat{w}^T x_0)$$

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:

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

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

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

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

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

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

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

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

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

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

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

LINEAR REGRESSION: MLE

⑥ Thus:

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

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

$$= \underset{w}{\operatorname{argmax}}\ -\sum_{n=1}^{N}(y_n - w^T x_n)^2$$

---

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

$$\sum_{n=1}^{N}(y_n - w^T x_n)^2 = \begin{bmatrix} y_1 - w^T x_1 \\ \vdots \\ y_N - w^T x_N \end{bmatrix}^T \begin{bmatrix} y_1 - w^T x_1 \\ \vdots \\ y_N - w^T x_N \end{bmatrix}$$

and noticing that:

$$\begin{bmatrix} y_1 - w^T x_1 \\ \vdots \\ y_N - w^T x_N \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} w^T x_1 \\ \vdots \\ w^T x_N \end{bmatrix}$$

$$= y - \begin{bmatrix} w_1 x_1[1] + \dots + w_D x_1[D] \\ \vdots \\ w_1 x_N[1] + \dots + w_D x_N[D] \end{bmatrix}$$

$$= y - \begin{bmatrix} x_1[1] & \cdots & x_1[D] \\ \vdots & & \vdots \\ x_N[1] & \cdots & x_N[D] \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_D \end{bmatrix}$$

$$= y - Xw$$

LINEAR REGRESSION: MLE

(8) So

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

$$= \underset{w}{\operatorname{argmax}} \cancel{-y^T y} + y^T Xw + (Xw)^T y - (Xw)^T Xw$$

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

[since $(AB)^T = B^T A^T$]

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

[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:

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

---

(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:

$$\hat{w} = \underset{w}{\operatorname{argmax}} \prod_{n=1}^{N} P(y_n | w, x_n)$$

$$= \underset{w}{\operatorname{argmax}}\ 2w^T X^T y - w^T X^T Xw$$

$$= \underset{w}{\operatorname{argmin}}\ w^T X^T Xw - 2wX^T y$$

So the loss function for ordinary linear regression is:

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

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

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

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.

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

$\hat{y}_0 \leftarrow \hat{w}^T x_0$