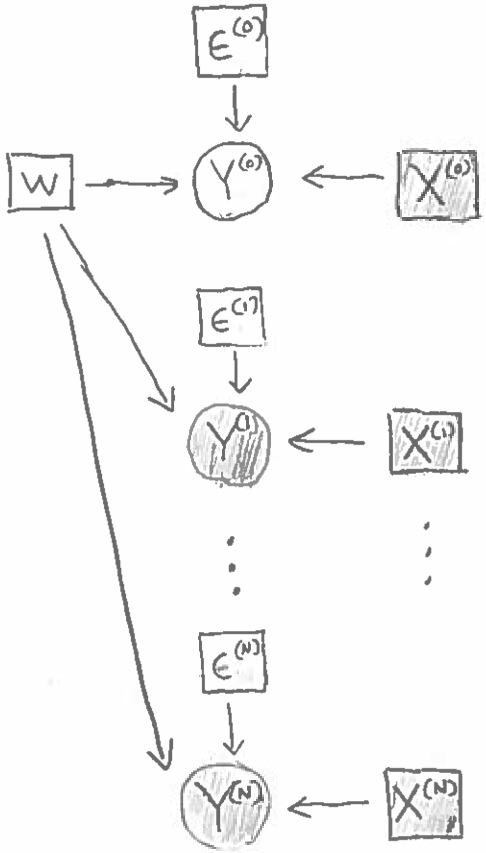
LOGISTIC REGRESSION: MLE

① Recall "logistic regression":



where: $\epsilon^{(n)} \sim \text{Constant}(0, 1) \quad \forall n \in \{0, \ldots, N\}$

$$y^{(n)} \leftarrow \mathbb{1}_{\epsilon^{(n)} < (1+\exp(-w^T x^{(n)}))^{-1}}$$

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

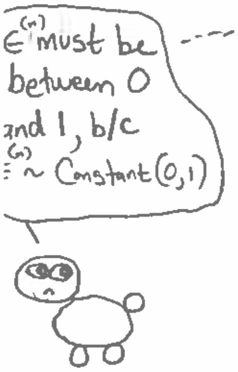
(b) compute $\hat{y}^{(0)} = \underset{y^{(0)}}{\text{argmax}} \; P(y^{(0)} | \hat{w}, x^{(0)})$

Logistic Regression: MLE

③ To compute the second step, observe:

$$P(Y^{(n)}=1 \mid w, x^{(n)})$$

$$= \int_0^1 P(Y^{(n)}=1, \epsilon^{(n)} \mid w, x^{(n)})\, d\epsilon^{(n)} \qquad \text{[total probability]}$$



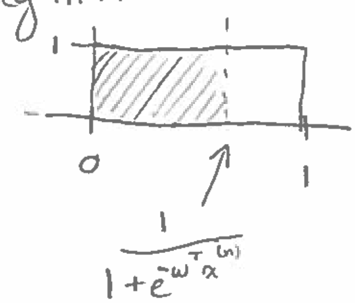
$$= \int_0^1 P(Y^{(n)}=1 \mid w, x^{(n)}, \epsilon^{(n)})\, P(\epsilon^{(n)} \mid w, x^{(n)})\, d\epsilon^{(n)} \qquad \text{[Chain Rule]}$$

$$= \int_0^1 P(Y^{(n)}=1 \mid w, x^{(n)}, \epsilon^{(n)})\, P(\epsilon^{(n)})\, d\epsilon^{(n)} \qquad \text{[d-separation]}$$

$$= \int_0^1 P\left(\epsilon^{(n)} < \frac{1}{1+e^{-w^T x^{(n)}}}\right) P(\epsilon^{(n)})\, d\epsilon^{(n)} \qquad \left[\text{b/c } y^{(n)} \leftarrow \mathbb{1}_{\epsilon^{(n)} < (1+\exp(-w^T x^{(n)}))^{-1}}\right]$$

$$= \int_0^{\frac{1}{1+e^{-w^T x^{(n)}}}} P(\epsilon^{(n)})\, d\epsilon^{(n)} \qquad \left[\text{everywhere else, } P\left(\epsilon^{(n)} < \frac{1}{1+e^{-w^T x^{(n)}}}\right) = 0\right]$$

$$= \frac{1}{1+e^{-w^T x^{(n)}}} \qquad \text{[Constant distribution integration]}$$



Thus:

$$P(Y^{(n)}=0 \mid w, x^{(n)}) = 1 - \frac{1}{1+e^{-w^T x^{(n)}}} = \frac{1+e^{-w^T x^{(n)}}}{1+e^{-w^T x^{(n)}}} - \frac{1}{1+e^{-w^T x^{(n)}}}$$

$$= \frac{e^{-w^T x^{(n)}}}{1+e^{-w^T x^{(n)}}}$$

Logistic Regression: MLE

④ Or, more compactly:

$$P(y^{(n)} | w, x^{(n)}) = \frac{e^{-(1-y^{(n)}) w^T x^{(n)}}}{1 + e^{-w^T x^{(n)}}}$$

$$= \begin{cases} \dfrac{e^{-w^T x^{(n)}}}{1 + e^{-w^T x^{(n)}}} & \text{if } y^{(n)} = 0 \\ \dfrac{1}{1 + e^{w^T x^{(n)}}} & \text{if } y^{(n)} = 1 \end{cases}$$

---

⑤ This allows us to express ②(b) as:

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

$$\boxed{= \underset{y^{(o)} \in \{0,1\}}{\operatorname{argmax}} \; \frac{e^{-(1-y^{(o)}) w^T x^{(o)}}}{1 + e^{-w^T x^{(o)}}}}$$

Logistic Regression: MLE

⑥ To compute 2(a), we start with some simplifications:

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

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

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

$$= \underset{w}{\operatorname{argmax}} \sum_{n=1}^{N} \log \frac{e^{-(1-y^{(n)})w^T x^{(n)}}}{1 + e^{-w^T x^{(n)}}} \qquad \text{[from ④]}$$

$$= \underset{w}{\operatorname{argmax}} \sum_{n=1}^{N} \log e^{-(1-y^{(n)})w^T x^{(n)}} - \log(1 + e^{-w^T x^{(n)}})$$

$$= \underset{w}{\operatorname{argmax}} \sum_{n=1}^{N} -(1-y^{(n)})w^T x^{(n)} + \log \frac{1}{1 + e^{-w^T x^{(n)}}}$$

$$= \underset{w}{\operatorname{argmin}} \sum_{n=1}^{N} (1-y^{(n)})w^T x^{(n)} - \log \frac{1}{1 + e^{-w^T x^{(n)}}}$$

So for logistic regression, our loss function is:

$$L_{\text{logistic}}(w) = \sum_{n=1}^{N} (1-y^{(n)})w^T x^{(n)} - \log \frac{1}{1+e^{-w^T x^{(n)}}}$$

LOGISTIC REGRESSION: MLE

7) To compute the gradient of $L_{logistic}(w)$, we'll first prove the following lemma:

> Lemma: If $\sigma(a) = \frac{1}{1+e^{-a}}$, then:
>
> $$\frac{d}{da}\sigma(a) = \sigma(a)(1-\sigma(a))$$

Proof:

$$\frac{d}{da}\sigma(a) = \frac{-1}{(1+e^{-a})^2} \cdot e^{-a} \cdot -1$$

$$= \frac{e^{-a}}{(1+e^{-a})^2}$$

$$= \left(\frac{1}{1+e^{-a}}\right)\left(\frac{e^{-a}}{1+e^{-a}}\right)$$

$$= \sigma(a)\left(\frac{1+e^{-a}-1}{1+e^{-a}}\right)$$

$$= \sigma(a)\left(\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right)$$

$$= \sigma(a)(1-\sigma(a)) \quad \boxtimes$$

(8) So the gradient of the loss function is:

$$\frac{d}{dw} L_{logistic}(w)$$

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

$$= \sum_{n=1}^{N} (1-y^{(n)}) \frac{d}{dw} w^T x^{(n)} - \frac{1}{\sigma(w^T x^{(n)})} \frac{d}{dw} \sigma(w^T x^{(n)})$$

$$= \sum_{n=1}^{N} (1-y^{(n)}) \frac{d}{dw}(w^T x^{(n)}) - \frac{1}{\sigma(w^T x^{(n)})} \sigma(w^T x^{(n)})(1-\sigma(w^T x^{(n)})) \frac{d}{dw} w^T x^{(n)}$$

[from Lemma]

$$= \sum_{n=1}^{N} (1-y^{(n)}) \frac{d}{dw}(w^T x^{(n)}) - (1-\sigma(w^T x^{(n)})) \frac{d}{dw}(w^T x^{(n)})$$

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

[b/c $\frac{d}{dw} w^T x = x$]

$$= \sum_{n=1}^{N} (1-y^{(n)} - 1 + \sigma(w^T x^{(n)})) x^{(n)}$$

$$\boxed{= \sum_{n=1}^{N} (\sigma(w^T x^{(n)}) - y^{(n)}) x^{(n)}}$$

LOGISTIC REGRESSION: MLE

⑨ This can be expressed even more compactly in terms of the evidence matrix X and response vector y:

$$\frac{d}{dw} L_{\text{logistic}}(w) = \sum_{n=1}^{N} \left(\sigma(w^T x^{(n)}) - y^{(n)}\right) x^{(n)}$$

$$= X^T(\sigma(Xw) - y)$$

Exercise: Show $X^T(\sigma(Xw)-y) = \sum_{n=1}^{N} \left(\sigma(w^T x^{(n)}) - y^{(n)}\right) x^{(n)}$

LOGISTIC REGRESSION: MLE

(10) As usual, there isn't a known way to solve directly for $\frac{d}{dw} L_{logistic}(w) = 0$, however we are free to use gradient descent.

LOGISTIC REGRESSION $(X, y, x^{(0)})$:

- Compute point estimate $\hat{w} = \text{GRADDESCENT}\left(L_{logistic}\right)$
- Compute prediction $\hat{y}^{(0)} = \underset{y^{(0)} \in \{0,1\}}{\operatorname{argmax}} \dfrac{e^{-(1-y^{(0)})w^T x^{(0)}}}{1 + e^{-w^T x^{(0)}}}$
- return $\hat{y}^{(0)}$