Logistic Regression.

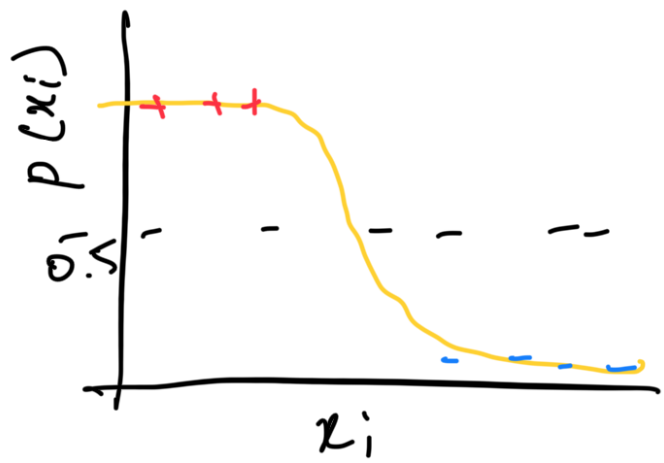
→ classification ; binary cross-entropy loss.

$$\hat{y}_i = \sigma(wx_i + b)$$

↳ sigmoid : $\sigma(z) = \dfrac{1}{1+e^{-z}}$

→ Now loss, here it is not straightforward,

so, lets see how loss came



→ Lets, say we are having $y_i \{1, -1\}$.

→ And $p(x_i) \to \hat{y}_{(i)}$ is the prob of the class.

reg model to

→ We want our $y_i \to$ [illegible] to maximize the observed prob. either be +ve or -ve.

→ The total prob. of data, since each observation is i.i.d

$$\prod (\text{data points})$$

$$\Rightarrow \prod \begin{cases} g(i) & \text{when } y_i = +1 \\ 1-g(i) & \text{when } y_i = -1 \end{cases}$$

$g(i) \to$ prob. of x_i being +1, that's why we use $1-g(i)$ when $x_i \neq +1$, to maximize overall prob.

$$\prod_{i=1}^{N} (g(i))^{\overbrace{1\{y_i=+1\}}^{y_i}} \cdot (1-g(i))^{\overbrace{1\{y_i=-1\}}^{1-y_i}}$$

-> multiplying all maximized prob.

$$\Rightarrow \prod_{i}^{N} \hat{y}_i^{\,y_i} \cdot (1-\hat{y}_i)^{1-y_i}$$

$i=1$

→ the loss is −ve of above term & since multiply long prob ~0 and math. efficiency converting multplication to addition we take log.

$$\text{loss} = -\log[\ldots]$$

$$\boxed{= y_i \cdot \log(\hat{y}) + (1-y_i)\log(1-\hat{y})}$$

final loss ↑

Now, need to find.

$$\frac{\partial L}{\partial w} \quad \& \quad \frac{\partial L}{\partial b}$$

$$\frac{\partial L}{\partial w} = y_i \frac{\partial \log(\hat{y})}{\partial w} + (1-y_i)\frac{\partial}{\partial w}\log(1-\hat{y})$$

lets start with inner,

$$\frac{\partial(\log(\hat{y}_i))}{\partial w} = \frac{1}{\hat{y}} \cdot \frac{\partial \hat{y}_i}{\partial w}$$

σ -

since, $\hat{y}_i = \sigma(z_i) = \dfrac{1}{1+e^{-z_i}}$; $z_i = w x_i + b$

$$\frac{\partial \hat{y}_i}{\partial w} = \frac{\partial \sigma(z_i)}{\partial w} = \sigma'(z_i) \cdot \frac{\partial z_i}{\partial w}$$

$$\sigma'(z_i) = \overbrace{\sigma(z_i)}^{\hat{y}_i}(1-\sigma(z_i)) = \hat{y}_i \cdot (1-\hat{y}_i)$$

$$z_i = w x_i + b \; ; \; \frac{\partial z_i}{\partial w} = x_i$$

$$\frac{\partial z_i}{\partial b} = 1$$

$$\therefore \frac{\partial \hat{y}_i}{\partial w} = \hat{y}_i \cdot (1-\hat{y}_i) \cdot x_i$$

$$\therefore \frac{\partial(\log(\hat{y}_i))}{\partial w} = \frac{1}{\cancel{\hat{y}}} \cdot (\cancel{\hat{y}_i}) \cdot (1-\hat{y}_i) \cdot x_i$$

$$= (1-\hat{y}_i)\, x_i$$

Similarly,

$$-\frac{\partial \hat{y}_i}{\partial w}$$

$$\frac{\partial(\log(1-\hat{y}_i))}{\partial w} = \frac{1}{1-\hat{y}_i}\cdot\frac{\partial(1-\hat{y}_i)}{\partial w}$$

$$= \frac{1}{\cancel{1-\hat{y}}}\cdot(-1)\cdot\hat{y}_i\cdot(\cancel{1-\hat{y}_i})x_i$$

$$= -\hat{y}_i x_i$$

Combining everything.

$$\frac{\partial L}{\partial w} = y_i(1-\hat{y}_i)x_i + (-\hat{y}_i x_i)(1-\hat{y}_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i(y_i-\hat{y}_i)$$

Vectorized,

$$\frac{1}{n}\cdot X^T(y-\hat{y})$$

$\frac{\partial J}{\partial b}$, not calculating, [illegible],

$$\frac{1}{n} 1^T (\hat{y} - y)$$