Logistic Regression

30 October 2021 09:52 AM

$$X = \begin{bmatrix} | & | & | & & | \\ x^{(1)} & x^{(2)} & x^{(3)} & \dots & x^{(m)} \\ | & | & | & & | \end{bmatrix}$$ n features, m examples. $n \times m$. → diff. from linear reg.

$$Y = \begin{bmatrix} y^{(1)} & y^{(2)} & y^{(3)} & \dots & y^{(m)} \end{bmatrix} \quad 1 \times m$$

Logistic Regression:

example x, we want $\hat{y} = P(y=1 \mid x)$. (prediction.)

☑ → CAT ($y=1$)
→ NOT CAT ($y=0$)

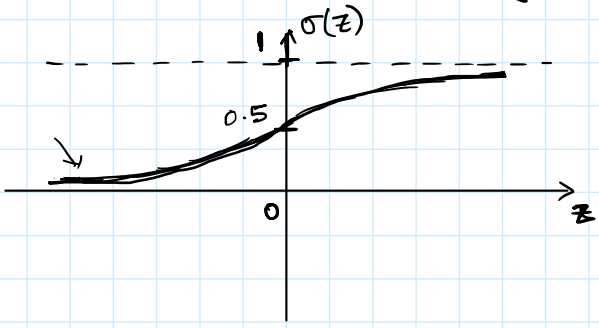
in linear reg., $h(x) = \theta_0 + \theta_1 x$ Here, define 2 parameters w, b $w \in \mathbb{R}^n$, $b \in \mathbb{R}$.

$\hat{y} = \boxed{w^T x + b}$ ← APPLY FN. ($w^T x + b$: one example)

($1 \times n$) ($n \times 1$)

we need $0 \le \hat{y} \le 1$. use SIGMOID

$$\hat{y} = \sigma(\underbrace{w^T x + b}_{z}). \qquad \sigma(z) = \frac{1}{1+e^{-z}}.$$



z very large, $z \to \infty$, $\sigma(z) \to 1$
z very small $z \to -\infty$, $\sigma(z) \to 0$

$\theta^T x$: θ_0 → we add a new feature $x_0 = 1$; θ_1; θ_2; ⋮

✗ keep b separate

$\hat{y} = \sigma(\underbrace{w^T x + b}_{z})$ COST function. $L(\hat{y}, y) = -(y \log \hat{y} + (1-y)\log(1-\hat{y}))$

CROSSENTROPY LOSS. how diff. $\hat{y}$ is from y.

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)}).$$

$J(w,b)$: function of parameters.

Now we have: • hypothesis → $\hat{y} = \sigma(w^T x + b)$
• cost → $J(w,b)$.

GOAL: minimize J. → gradient descent

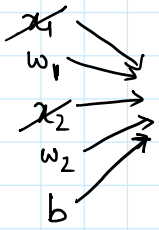
Repeat until convergence: $w = w - \alpha \cdot \frac{\partial J}{\partial w}$ $\qquad dw = \frac{\partial J}{\partial w} = \frac{\partial L}{\partial w}$, $db = \frac{\partial L}{\partial b}$

$$b = b - \alpha \frac{\partial J}{\partial b}.$$

(α: LEARNING RATE)

Set $n=2$. $z = w^T x + b$

$= w_1 x_1 + w_2 x_2 + b$ $\quad \hat{y} = \sigma(z) \rightarrow$ predictions.



$z = w_1 x_1 + w_2 x_2 + b \rightarrow \hat{y} = \sigma(z) \rightarrow L(\hat{y}, y)$. computation graph.

$da = \frac{\partial L}{\partial a} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial a}$ $\quad d\hat{y} = \frac{\partial L}{\partial \hat{y}}$

$L = -(y \log \hat{y} + (1-y) \log(1-\hat{y}))$.

$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}$

$= -\left(\frac{y}{\hat{y}} - \frac{(1-y)}{1-\hat{y}}\right) = -\frac{y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}$

$\frac{\partial \hat{y}}{\partial z} = \sigma'(z)$.

$= \hat{y}(1-\hat{y})$.

$\frac{e^{-z}}{(1+e^{-z})^2}$ $\quad \left(\frac{1}{1+e^{-z}}\right) \cdot \frac{1+e^{-z}-1}{(1+e^{-z})} = \frac{1}{(1+e^{-z})}\left(1 - \frac{1}{1+e^{-z}}\right)$

$= \hat{y}(1-\hat{y})$.

$dz = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z}$

$= \left(-\frac{y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}\right) \cdot \hat{y}(1-\hat{y})$.

$= -y(1-\hat{y}) + (1-y)\hat{y}$

$= -y + y\hat{y} + \hat{y} - y\hat{y} = \hat{y} - y$.

$\frac{\partial L}{\partial z} = dz = \hat{y} - y$. $\quad \frac{\partial z}{\partial w_1} = x_1$, $\frac{\partial z}{\partial w_2} = x_2$

$\frac{\partial z}{\partial b} = 1$.

$\frac{\partial L}{\partial w_1} = dw_1 = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial w_1} = x_1 dz = x_1(\hat{y} - y)$.

$dw_2 = x_2 dz = x_2(\hat{y} - y)$.

$db = dz = \hat{y} - y$.

FOR ONE STEP OF GRADIENT DESCENT: $w_1 = w_1 - \alpha dw_1$, $w_2 = w_2 - \alpha dw_2$

$b = b - \alpha db$.

$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix}$ $\quad n \times 1$

$z^{(1)} = w^T x^{(1)} + b$, $\hat{y}^{(1)} = \sigma(z^{(1)})$

$z^{(2)} = w^T x^{(2)} + b$ $\quad \hat{y}^{(2)} = \sigma(z^{(2)})$

$z^{(3)} = w^T x^{(3)} + b$. $\quad \hat{y}^{(3)} = \sigma(z^{(3)})$.

$Z = w^T X + b$

$(1,n)\ (n,m) \rightarrow (1,m)$

$[z^{(1)} \ z^{(2)} \ z^{(3)} \ \dots]$.

add same b.

$dz = [dz^{(1)} \ dz^{(2)} \dots]$.

$(1,m)$

$dz = \hat{y} - y$. $\quad dw_1 = x_1 dz$

$dw_2 = x_2 dz$.

$\hat{Y} = \sigma(Z)$. $\quad dZ = \hat{Y} - Y$.

$dw = X \cdot dz^T \cdot \frac{1}{m}$

ALL EXAMPLES, using J.

$A = \sigma(Z)$. $dZ = A - Y$.

$\underline{dw} = X \cdot dZ^T \cdot \frac{1}{m}$

$(n,m)\ (m,1) \Rightarrow (n,1)$.

$db = \frac{1}{m}\sum_{i=1}^{m} dz^{(i)} = \frac{1}{m}\, np.sum(dZ)$

ALL EXAMPLES, using J.

$J = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)})$.

$\frac{\partial J}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} dw^{(i)} = \frac{1}{m} \cdot \underline{dw}$

$\frac{1}{m}\sum db = \frac{1}{m}\sum dz^{(i)}$

dZ ✓, dw, db.

$w = w - \alpha\, dw$, $b = b - \alpha\, db$. → one step of GD.

$(n,1) \quad (n,1)$

$\underline{J} = -\frac{1}{m}\left(\underline{Y} \cdot (\underline{np.log(\hat{Y})})^T + (1-Y)\,(\underline{np.log(1-\hat{Y})})^T\right)$.