Logistic Regression-1

Friday, August 16, 2019 10:00 AM

Agenda :

→ log reg Intro.

→ Sig / Sigmoid function

→ Geometric Intuition.

→ Loss function

→ Gradients / optimization.

→ Implementation

→ Hyperparameters

---

* Intro

Lin. reg → Regression → linear model

Classification → Classification

① $\hat{y} = \boxed{w^T x + b}$

② loss = MSE $\frac{1}{2}(y^{(i)} - \hat{y}^{(i)})^2$

* Logistic Regression.

Binary Classification

→ Supervised Alg.

→ Classification Alg.

$D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$; $x^{(i)} \in \mathbb{R}^n$; $y^{(i)} \in \{0, 1\}$

$y^{(i)} \in \{0, 1\}$

→ Binary model : hypothesis $\boxed{w^T x + b}$

$(-\infty, \infty)$

$x^{(i)} \in \mathbb{R}^n$, $w^T x + b$ → $\{0, 1\}$

Sensitivity to points

$w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$

$\hat{y} = \boxed{w^T x + b}$

$(-\infty, \infty)$

[Not bounded]

[not bounded]

$z^{(i)} \longrightarrow \frac{1}{1+e^{-z}} \longrightarrow \frac{1}{1+e^{-(w^T x + b)}} \longrightarrow \{0, 1\}$ or $\{-1, +1\}$

$\frac{1}{1+e^{-z}}$ → [0, 1]

[not bounded]

if $z^{(i)} \to \infty$ $\frac{1}{1+0} = 1$

if $z^{(i)} \to -\infty$ $\frac{1}{1+\infty} \to 0$

* Logistic function
Sigmoid function

$g(z)$

$z^{(i)} \to \infty$: $g(z^{(i)}) \to 1$

$(e^{-\infty} \to 0)$

$z^{(i)} \to -\infty$: $g(z^{(i)}) \to 0$

$z^{(i)} \to 0$: $g(z^{(i)}) \to 0.5$

$\hat{y} \Rightarrow g(w^T x + b) = (0, 1)$

$\hat{y} \to 0$

$y^{(i)} \in \{0, 1\}$

$\boxed{\hat{y}^{(i)} \in (0, 1)}$

$\hat{y}^{(i)} \to 1$ $\hat{y}^{(i)} \to 0$

$\hat{y}^{(i)} \to 0.5$

Can be thought as as Probability.

$\hat{y}^{(i)} = P(y^{(i)} = 1 | x^{(i)})$ → Conditional Prob.

$P(y^{(i)} = 0 | x^{(i)}) = 1 - \hat{y}^{(i)}$

$\frac{e^{-z}}{1+e^{-z}}$ for the above model

$\xrightarrow{\text{let}}$ $\frac{1}{1+e^{-(w^T x + b)}}$ $\xrightarrow{\sigma(z^{(i)})}$ $\boxed{\hat{y}^{(i)}}$

$(-\infty, \infty)$ $(0, 1)$

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

$P(y=1|x) \to 0.5$

Decision Boundary

$\to 0.5$

$\to 0$

$\hat{y}(x_{new})$

$\hat{y}(x_{old})$

$\hat{y}_{x_1}$

$\hat{y}_{x_2}$

→ Linear Function

$g(z^{(i)}) = \text{Sigmoid}(z^{(i)}) \Rightarrow \frac{1}{1+e^{-z}}$

① Smooth & diff.

② Squashing function
$(-\infty, \infty) \to (0, 1)$

③ Derived from probability theorem

④ monotonic function (Non decreasing)

* Geometric Intuition

w, x, m

$x^{(i)}, w^T x + b$

$\sigma(z^{(i)})$ $\boxed{\hat{y}^{(i)}}$

$z^{(i)}, w^T x + b$

$\{x^{(i)}, y^{(i)}\} \to D$

$w^T x + b$

$\hat{y}^{(i)} \to 0$

$\{x^{(i)}, y^{(i)}\} \in D$

$\hat{y}^{(i)} \to 1$

$\{x^{(i)}, y^{(i)}\} \in D$

$w^T x + b = 0$

Separating hyperplane

$\boxed{\hat{y}^{(i)} = 1}$

$\hat{y}^{(i)} \to 0$

$\hat{y}^{(i)} \to 1$

$\hat{y}^{(i)} \to 0.5$

$\hat{y}^{(i)} \to 0.5$

$w^T x + b \to \infty$

$\hat{y}^{(i)} \to 1$

$\boxed{P(y^{(i)} = 1 | x^{(i)}) > P(y^{(i)} = 0 | x^{(i)})}$

$0.98 \quad > \quad 0.65$

$w^T x + b \to -\infty$

$\hat{y}^{(i)} \to 0$

$\boxed{P(y^{(i)} = 0 | x^{(i)}) > P(y^{(i)} = 1 | x^{(i)})}$

$0.98 \quad > \quad 0.65$

$\hat{y}^{(i)} \to 0.5$

$w^T x + b \to 0$

$\hat{y}^{(i)} \to 0.5$

$\hat{y}^{(i)} = 1$

$\hat{y}^{(i)} = 0$

$P(y^{(i)} = 1 | x^{(i)}) < P(y^{(i)} = 0 | x^{(i)})$

---

* Loss func + Sigloss

Inputs : $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$

① Linear Separator $w^T x + b = 0$

② $w^T x + b = 0, z \in (-\infty, \infty)$ → Squashing $\sigma(z)$ → $(0, 1)$ $\hat{y}^{(i)}$ $P(y=1|x)$

③ $\begin{cases} \hat{y}^{(i)} = 0 \text{ or } 1 \\ \hat{y}^{(i)} \in (0, 1) \end{cases}$ Compare $y^{(i)}$ Vs $\hat{y}^{(i)}$

Log-Loss

$\hat{y}^{(i)} = 1$

$\log(\hat{y}^{(i)})$

$\hat{y}^{(i)} = 0$

$\log(1-\hat{y}^{(i)})$

$\text{log-Loss} = \begin{cases} -\log(\hat{y}^{(i)}) & \text{if } y^{(i)} = 1 \\ -\log(1-\hat{y}^{(i)}) & \text{if } y^{(i)} = 0 \end{cases}$

$\text{Log-Loss} = -y^{(i)} \log(\hat{y}^{(i)}) - (1-y^{(i)}) \log(1-\hat{y}^{(i)})$

when $y^{(i)} = 1$ $\Rightarrow$ $-\log(\hat{y}^{(i)})$

when $y^{(i)} = 0$ $\Rightarrow$ $-\log(1-\hat{y}^{(i)})$

$\boxed{\text{loss} = -\frac{1}{m}\sum_{i=1}^{m} [y^{(i)} \log(\hat{y}^{(i)}) + (1-y^{(i)}) \log(1-\hat{y}^{(i)})]}$

Why log-loss ? and not MSE ?

→ if you use MSE for Binary classification, it will be a non-convex function.

MSE

(non-convex)

(log-loss)

(MSE)

Linear Reg. use MSE

Loss

$L = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(\hat{y}^{(i)}) + (1-y^{(i)}) \log(1-\hat{y}^{(i)})$

$\boxed{\hat{y} = \sigma(w^T x)}$

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

Log

Note : Non-linear Func

Can it work in our reg ?

Yes.

Why ? → Loss do not convert to non convex.

Inputs

Train. $-\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(\hat{y}^{(i)}) + (1-y^{(i)}) \log(1-\hat{y}^{(i)})$

$\hat{y} \Rightarrow \boxed{\sigma(z)} \Rightarrow \sigma(w^T x + b)$

$\sigma'(z) \Rightarrow \frac{d\sigma(z)}{dz} \Rightarrow \sigma(z)[1 - \sigma(z)]$

$\sigma(z)(1-\sigma(z)) = \hat{y}(1-\hat{y})$

$\boxed{\frac{d\hat{y}}{dz} = \hat{y}(1-\hat{y})}$

NOTE : For 1 data point

$L = \boxed{-[y \log(\hat{y}) + (1-y) \log(1-\hat{y})]}$

$\frac{\partial L}{\partial w_j} \Rightarrow \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w_j}$

$\Rightarrow \frac{y}{\cancel{\hat{y}}} \cdot \cancel{\hat{y}}(1-\hat{y}) \cdot x_j$

$\Rightarrow y(1-\hat{y}) \cdot x_j$

$\frac{\partial L}{\partial w_j} \Rightarrow \frac{\partial L}{\partial (1-\hat{y})} \cdot \frac{\partial (1-\hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w_j}$

$\Rightarrow \frac{1-y}{\cancel{1-\hat{y}}} \cdot (-1) \times \cancel{\hat{y}(1-\hat{y})} \times x_j$

$\frac{\partial(1-\hat{y})}{\partial \hat{y}} = -1$

$\frac{\partial z}{\partial w} = x$

$\frac{\partial \hat{y}}{\partial z} = \cancel{\hat{y}(1-\hat{y})}$

$\Rightarrow \frac{1-y}{\cancel{1-\hat{y}}} \cdot (-1) \times \hat{y} \cdot x_j$

$\Rightarrow \frac{\partial L}{\partial w_j} \Rightarrow y \cdot (1-\hat{y}) \cdot x_j - (1-y)\hat{y} \cdot x_j$

$\Rightarrow x_j(y \cdot (1-\hat{y}) - (1-y)\hat{y})$

$\Rightarrow x_j(y - y\hat{y} - \hat{y} + y\hat{y})$

$\Rightarrow (y - \hat{y}) x_j$

$\frac{\partial L}{\partial w_j} \Rightarrow (y - \hat{y}) x_j$

$\boxed{\frac{\partial L}{\partial w_j} = (\hat{y} - y) x_j}$

→ Intuition : +ve, you change ∂L

for all examples (i = 1 to m)

$\boxed{\frac{\partial L}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) x_j^{(i)}}$

Gradient Descent

$w_j^{new} = w_j^{old} - \eta \cdot \frac{\partial L}{\partial w_j}$

* Big Picture

① Hypothesis $(\hat{y})$: $\sigma(w^T x + b)$ ✓

② Error func : log-loss ✓

③ Gradient : Stoch. Linear Reg.

④ Gradient Descent : Stoch. Linear Reg.

$y^{(i)}$

$\boxed{y_{pred} = 0}$

$\hat{y}_{pred}$

$\hat{y} = 0.0001$

$\hat{y} = 0.1$ not bad (?)

$\hat{y} = 0.5$ → $\sigma(z)$

$\hat{y} = 0.9$ → $\sigma(z)$

$y^{(i)}, \hat{y}^{(i)}$:

	Predicted	Logistic
0	0.0001	$\sigma(z)$
1	0.1	$\sigma(z)$
0	0.9	$\sigma(z)$
1	0.0001	$\sigma(z)$

for 0.0001 → 0.1 → not too worse & $\boxed{\log(-\ln(\text{zero}))}$

Finally :

$\log(\text{zero})$

$\log(\log(\text{zero}))$