Logistic Regression 1

Andrew Ng Notes - 2011

Agenda:

1. Intro Logistic.
2. Step and Sigmoid function
3. Geometric Intuition
4. Loss function
5. Gradient / optimization
6. Scratch implementation.

Use case

"Classical Binary Prediction"

$\hat{y}_{binary}$

Can use upper and lower bound $\boxed{w^T x + w_0}$ $\hat{y} \in \{-1, 1\}$

$\underset{(n \times 1)}{y^{(i)}} = \underset{(n \times m)}{w^T} \cdot \underset{(m \times 1)}{x^{(i)}} + \underset{(n \times 1)}{w_0}$

Classification

$\underset{(n \times 1)}{y^{(i)}} \in \{-1, 1\}$ → "Binary classification"

STEP function.

$z^{(i)} = \underset{(1 \times m)}{w^T x^{(i)}} + \underset{(1 \times 1)}{w_0}$ $\xrightarrow{\sigma(\cdot)}$ $\{-1, 1\}$

"Step function"

$f(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$

$g(\cdot) \to$ step function

$g(\underset{(1 \times m)}{w^T x^{(i)}} + w_0) \to \hat{y}^{(i)}$

Prediction ✓

Sigmoid function (logistic function)

$\hat{y} = \frac{1}{1 + e^{-(w^T x + w_0)}}$

$z^{(i)} \to \infty \quad \hat{y} / y(z^{(i)}) = 1$

$z^{(i)} \to -\infty \quad \hat{y} / y(z^{(i)}) = 0$

$z^{(i)} = 0 \quad \hat{y} / y(z^{(i)}) = 0.5$

$\hat{y} \in \pm 1$

$\hat{y}^{(i)} > \underset{\text{Threshold}}{0.5}$

Affinity measure

$\hat{y} \to g(\underset{\text{Sigmoid}}{\sigma}(w^T x + w_0)) \to (0, 1)$

$\hat{y}^{(i)} \to P(y^{(i)} = 1 \mid x^{(i)}) \in (0, 1)$ Probability

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

$\hat{y} = \begin{cases} 1 & \frac{P(y^{(i)} = 1 \mid x^{(i)})}{P(y^{(i)} = 0 \mid x^{(i)})} > 1 \\ 0 & \frac{P(y^{(i)} = 1 \mid x^{(i)})}{P(y^{(i)} = 0 \mid x^{(i)})} < 1 \end{cases}$

Q: A different threshold.

"Classification metric"

✓ Fp, Recall, F1

Properties of Sigmoid.

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

$\sigma(x^{(i)})$

1. Smooth & differentiable
2. Symmetry Centered around $(0, 0.5)$
3. Bounded from Probability principle.
4. Differentiable Sensitive near Monotonic function (near decision)

Geometric Intuition.

$z^{(i)} > 0$ $\hat{y}^{(i)} = 1$

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

Threshold = 0

Centered

if $z^{(i)} > 0 \longrightarrow$ class 1

$z^{(i)} \to z^{(i)}$

$z^{(i)} = w^T x^{(i)} + w_0 \to 0 < 0$

$z^{(i)} < 0$

$\hat{y} = \sigma(z^{(i)}) < 0.5$

$P(y^{(i)} = 1 \mid x^{(i)}) = 0.25$

$P(y^{(i)} = 0 \mid x^{(i)}) = 0.75$

$\underset{(1 \times m)}{y^{(i)}} \quad \underset{(1 \times m)}{z^{(i)}}$

$z^{(i)} = \underset{(1 \times m)}{w^T x^{(i)}} + \underset{(1 \times 1)}{w_0}$

$\underset{(1 \times m)}{y^{(i)}} \to \underset{(1 \times m)}{z^{(i)}}$

$\sigma(z^{(i)}) > \sigma(z^{(i)})$

$P(y^{(i)} = 1 \mid x^{(i)}) > P(y^{(i)} = 1 \mid x^{(i)})$

$\hat{y} = 0.75$

$P(y = 1 \mid x) = 0.25$

$P(\hat{y}^{(i)} = 1 \mid x^{(i)}) > P(\hat{y}^{(i)} = 1 \mid x^{(i)})$

$\sigma(\underset{\text{Affinity}}{z^{(i)}}) \to (0, 1)$

Probability $\mid x^{(i)}$

$\hat{y} = \begin{cases} 1 & \frac{P(y^{(i)} = 1 \mid x^{(i)})}{P(y^{(i)} = 0 \mid x^{(i)})} > 1 \\ 0 & \frac{P(y^{(i)} = 1 \mid x^{(i)})}{P(y^{(i)} = 0 \mid x^{(i)})} < 1 \end{cases}$

The loss function → Log-loss

Setup: $D = \{ (x^{(1)}, y^{(1)}), \ldots, x^{(m)} \in \mathbb{R}^d, y^{(m)} \in \{-1, 1\} \}$

1. Linear Separator M_θ : $\boxed{w^T x + w_0 = 0}$

2. $x^{(i)} \to z^{(i)} = \underset{(1 \times m)}{w^T x^{(i)}} + \underset{(1 \times 1)}{w_0}$ $\xrightarrow[\sigma(z^{(i)})]{\text{Sigmoid}}$ $(0, 1)$ $\{y^{(i)}, \hat{y}^{(i)}\}$ $\{y, \hat{y}\} \in \{-1, 1\}$

3. $y^{(i)} = 0$ or 1

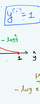
$\hat{y}^{(i)} \in (0, 1)$

Convex $\hat{y}^{(i)}$ & $y^{(i)}$

"$\boxed{\text{Log-loss}}$"

Log-loss

$\boxed{y^{(i)} = 1}$ $\boxed{y^{(i)} = 0}$



$-\log = \frac{\text{Pr}}{\text{Ur}}$

$-\log(1 - x)$

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

Convex since $\hat{y}$ sigmoid $\to$ convex ✓

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

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

Lower $y^{(i)} = 0$: $-\log(1 - \hat{y}^{(i)})$

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

Gradients / optimization.

$\hat{y} = \sigma(\underset{(1 \times m)}{w^T x^{(i)}} + w_0)$

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

argmin $\mathcal{L}$

$\underset{\hat{y}, \hat{z}}{}$

$\frac{\partial \mathcal{L}}{\partial w_0}, \frac{\partial \mathcal{L}}{\partial w_1}, \ldots, \frac{\partial \mathcal{L}}{\partial w_m}$

N.W.

$\hat{y} = \sigma(z) = \sigma(w^T x + w_0)$.

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

$\frac{d}{dz} \frac{1}{1 + e^{-z}} \Rightarrow \sigma(z) (1 - \sigma(z))$

$\boxed{\frac{d\hat{y}}{dz} \Rightarrow \hat{y} [1 - \hat{y}]}$

note: for 2 variables, we have 1 to one $\sum$

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

Log-loss

$\mathcal{L} \in (w_1, w_2, w_3, w_4, \ldots, w_0, y, w)$

$\frac{\partial \mathcal{L}_\theta}{\partial w_j} = \frac{\partial \mathcal{L}_\theta}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w_j}$

$\Rightarrow \frac{y}{\hat{y}} - \frac{1 - y}{1 - \hat{y}} \cdot \hat{y} [1 - \hat{y}] \cdot x_j$

$\Rightarrow y \cdot [1 - \hat{y}] \cdot x_j$

$\frac{\partial \mathcal{L}_\theta}{\partial w_j} = \frac{\partial \mathcal{L}_\theta}{\partial (1 - \hat{y})} \cdot \frac{\partial (1 - \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w_j}$

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

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

$\frac{\partial \mathcal{L}_\theta}{\partial w_j} = \frac{\partial \mathcal{L}_\theta}{\partial w_j} + \frac{\partial \mathcal{L}_\theta}{\partial w_j}$

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

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

$\Rightarrow [y - y\hat{y} - \hat{y} + y\hat{y}] \cdot x_j$

$\frac{\partial \mathcal{L}_\theta}{\partial w_j} \Rightarrow [y - \hat{y}] \cdot x_j$

(Gradient, one)

$\frac{\partial \mathcal{L}_\theta}{\partial w_j} = [y - \hat{y}] \cdot x_j$

$$\boxed{\frac{\partial \mathcal{L}_\theta}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} \left[\hat{y}^{(i)} - y^{(i)} \right] \cdot x_j^{(i)}}$$

note: $\hat{y} = w^T x + w_0$

$\Rightarrow \log \frac{y}{1 - y} \quad \hat{y} = \sigma(w^T x + w_0)$

(regression)

Gradient Descent.

$w_j^{(t+1)} \leftarrow w_j^{(t)} - \eta \cdot \frac{\partial \mathcal{L}}{\partial w_j}$

$w_j^{(t)} \in \mathbb{R}^d$