Classification

Classification

- Binary → Sigmoid
- Multiclass → Softmax

$$p(Y=1|x) = \sigma(w^T x)$$
$$p(Y=0|x) = 1 - \sigma(w^T x)$$
$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$p(Y=c|x) = \text{Softmax}(Wx)$$
$$\text{softmax}(z)_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

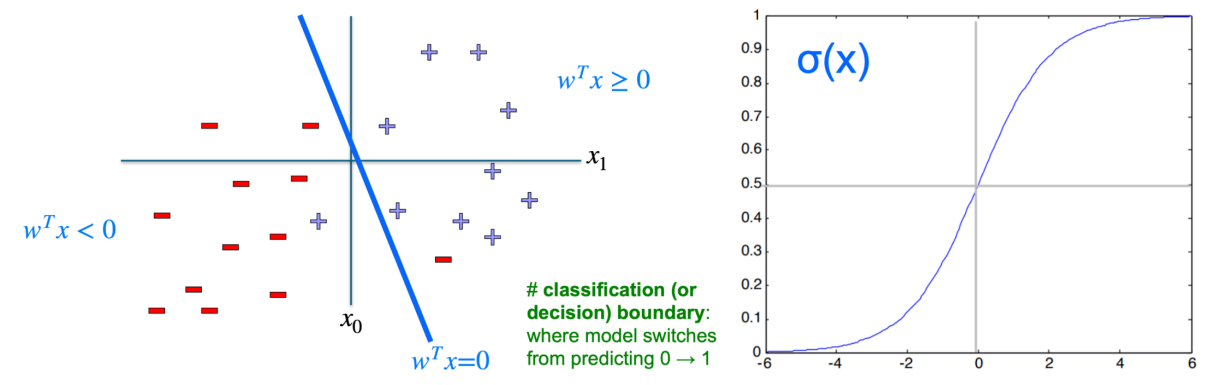
1. Binary Logistic Regression

Goal: predict label y from features x

Probabilistic Model:

$$p(Y=1|x) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$$
$$p(Y=0|x) = 1 - \sigma(w^T x) = 1 - \frac{1}{1+e^{-w^T x}} = \frac{e^{-w^T x}}{1+e^{-w^T x}}$$

Output: $\begin{cases} 1 & \text{if } p(Y=1|x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$



(a) Derive a unified loss: given $y \in \{-1, 1\}$

Show that $P_w(y|x) = \frac{1}{1+\exp(-y w^T x)}$

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

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

Identity $\quad 1 - \sigma(z) = \sigma(-z)$

When $y = +1$, $P_w(+1 \mid x) = \frac{1}{1+e^{-w^T x}}$

$y = -1$, $P_w(-1 \mid x) = \frac{1}{1+e^{w^T x}}$

(b) Given dataset of iid $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$.

derive max likelihood estimation for logistic loss:

$$\hat{w}_{MLE} = \operatorname{argmin}_w \sum_{i=1}^n \log\left(1 + \exp(-y_i w^T x_i)\right)$$

We have: $P_w(y_i \mid x_i) = \frac{1}{1+\exp(-y_i w^T x_i)}$

$$P_w(y \mid x) = \prod_i \frac{1}{1+\exp(-y_i w^T x_i)}$$

$$\log P_w(y \mid x) = \sum_i -\log\left(1+\exp(-y_i w^T x_i)\right)$$

$$\hat{w}_{MLE} = \operatorname{argmax}_w \sum_i -\log\left(1+\exp(-y_i w^T x_i)\right)$$

$$= \operatorname{argmin}_w \sum_i \log\left(1+\exp(-y_i w^T x_i)\right)$$

(c) Suppose we have a dataset that is linearly separable.

Does objective from part (b) converge to finite w?

Look at objective: $\sum_i \log\left(1+\exp(-y_i w^T x_i)\right)$

Data linearly separable: $\exists\, w^*$ s.t. $y_i w^{*T} x_i > 0 \;\forall i$.

But we can multiply w^* by $C > 0 \Rightarrow -y_i w^T x_i \to \infty \Rightarrow \exp(-y_i w^T x_i) \to 0$.

2. Multiclass Logistic Regression

$$P(Y = c_i \mid x) = \frac{e^{w_i^T x}}{\sum_{j=1}^{k} e^{w_j^T x}}$$

$w_i^T x$: logit

$\sum_{j=1}^{k} e^{w_j^T x}$: normalization

(a) Show that when $k = 2$, reduces to binary logistic regression.

Let $c_1 \Rightarrow y = 1$, show that $P(Y = c_1 \mid x) = \sigma(w^T x)$ for some w.

Plugging $k = 2$ into softmax:

$$P(Y = c_1 \mid x) = \frac{e^{w_1^T x}}{e^{w_1^T x} + e^{w_2^T x}}$$

$$= \frac{1}{1 + e^{w_2^T x - w_1^T x}}$$

$$= \frac{1}{1 + e^{(w_2 - w_1)^T x}}$$

$$= \sigma((w_2 - w_1)^T x)$$

(b) Suppose a 3-class classifier produces logits $w_1^T x = \log 3$, $w_2^T x = \log 5$, $w_3^T x = \log 1$ for some input x. Which class is predicted?

$$P(Y = c_1 \mid x) = \frac{e^{\log 3}}{e^{\log 3} + e^{\log 5} + e^{\log 1}}$$

$$= \frac{3}{3 + 5 + 1}$$

$$= \frac{1}{3}$$

$$P(Y = c_2 \mid x) = \frac{5}{9}$$

$$P(Y = c_3 \mid x) = \frac{1}{9}$$

c_2 has the largest probability.

3. Stochastic Gradient Descent

Consider the minimization problem

$$\min_w \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Full gradient descent :

$$w^{(t+1)} = w^{(t)} - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla_w f_i(w^{(t)})$$

Cost : $O(dn)$

Stochastic gradient descent :

$$I_t \sim \text{Uni}\{1, 2, \cdots, n\}$$

$$w^{(t+1)} = w^{(t)} - \eta \cdot \nabla_w f_{I_t}(w^{(t)})$$

Cost : $O(d)$

Problem : too noisy

Minibatch GD :

$$I_t \subseteq \{1, 2, \cdots, n\}, \quad |I_t| = b \ll n.$$

