인공지능과 수학적 배경

Classification

Logistic regression loss

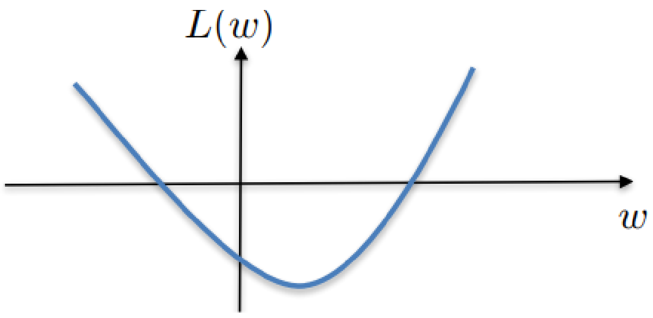
- **Most popular** classification loss is the **logistic regression loss**.
  - **Note: The name "logistic regression" may be confusing** as we deal with the classification task (not the regression task).
- **Definition**:

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(p_w(x_i), y_i)$$

$$\text{with } \ell(p_w(x_i), y_i) = \begin{cases} -\log p_w(x_i) & \text{if} \quad y_i = 1 \\ -\log(1 - p_w(x_i)) & \text{if} \quad y_i = 0 \end{cases}$$

$$\text{and} \quad p_w(x_i) = \frac{1}{1 + e^{-w^T x_i}}$$

- **Convexity: Logistic regression function $L(w)$ is convex** ☺



Classification

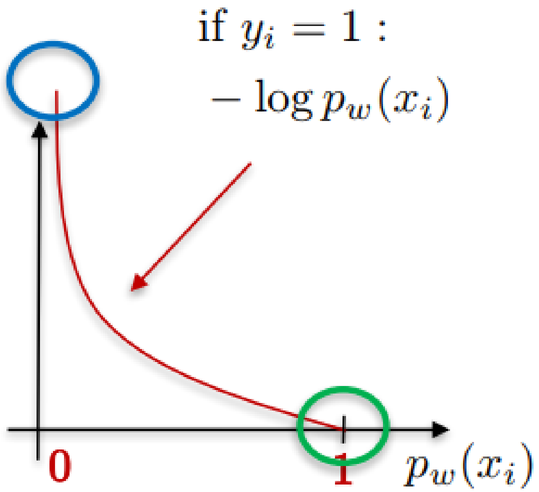
Loss analysis

- **Properties of the logistic regression loss:**
  - **If $y_i = 1$ and the predictive function $p_w(x_i)$ predict 1 (correct), we should have:**

$$\ell(p_w(x_i), y_i) = 0$$

  - **If $y_i = 1$ and the predictive function $p_w(x_i)$ predict 0 (mistake), we should penalize:**

$$\ell(p_w(x_i), y_i) = +\infty$$



Classification

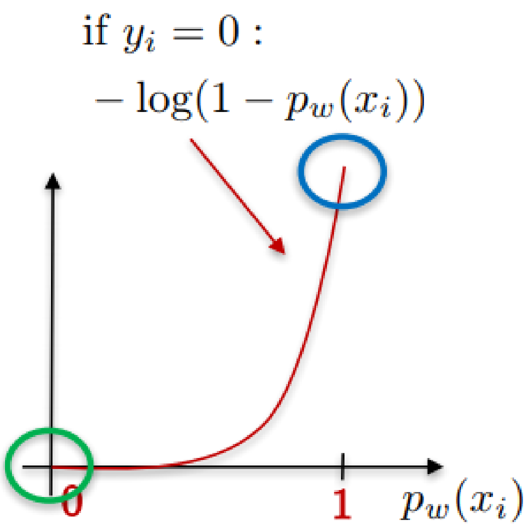
Loss analysis

- **Properties of the logistic regression loss:**
  - **If y_i = 0 and the predictive function $p_w(x_i)$ predict 0 (correct), we should have:**

$$\ell(p_w(x_i), y_i) = 0$$

  - **If y_i= 0 and the predictive function $p_w(x_i)$ predict 1 (mistake), we should penalize:**

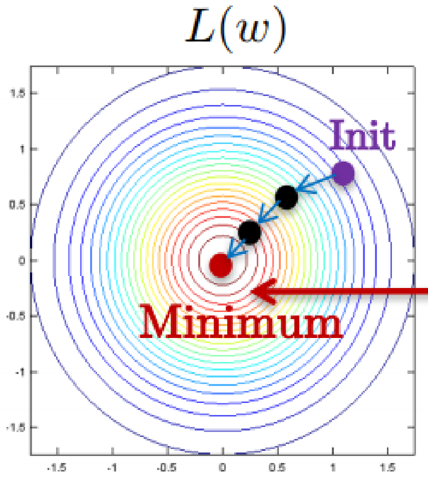
$$\ell(p_w(x_i), y_i) = +\infty$$



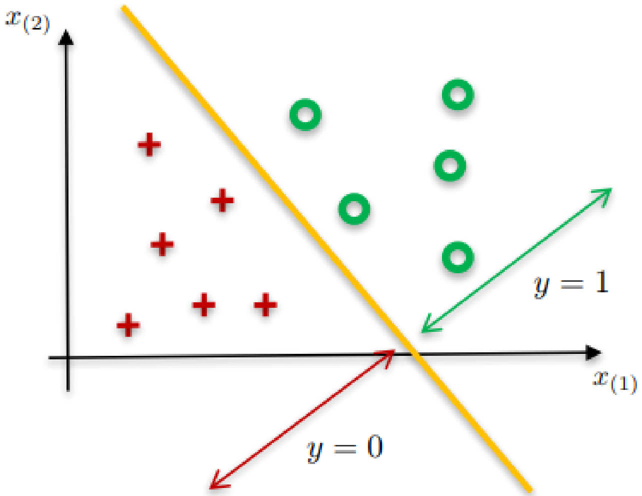
Classification

Gradient descent for logistic regression

- **Prediction function:** $$p_w(x) = \frac{1}{1+e^{-w^T x}}$$
- **Parameters:** $$w = [w_0, w_1, \ldots, w_d]$$
- **Loss function:** $$L(w) = -\frac{1}{n}\sum_{i=1}^{n}\Big(y_i \log p_w(x_i) + (1-y_i)\log(1-p_w(x_i))\Big)$$
- **Optimization:** $$\min_w \; L(w)$$
- **Gradient descent:** $$w_j \leftarrow w_j - \tau \frac{\partial}{\partial w_j} L(w)$$



$$\min_w \; L(w)$$



Classification

Gradient descent for logistic regression

- **Loss:**

$$L(w) = -\frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{y_i \log p_w(x_i)}_{\text{RHS1}} + \underbrace{(1-y_i)\log(1-p_w(x_i))}_{\text{RHS2}}\Big)$$

- **Gradient of RHS1:**

$$\frac{\partial}{\partial w_j}\left[-\frac{1}{n}\sum_{i=1}^{n} y_i \log p_w(x_i)\right] = -\frac{1}{n}\sum_{i=1}^{n} y_i \frac{\partial}{\partial w_j}\left[\log \sigma(w^T x_i)\right]$$

Chain rule

$$= -\frac{1}{n}\sum_{i=1}^{n} y_i \frac{\sigma'}{\sigma}\frac{\partial}{\partial w_j}\left[w^T x_i\right]$$

$$\sigma' = \frac{d\sigma}{d\eta} = (1-\sigma(\eta))\sigma(\eta)$$

$$= -\frac{1}{n}\sum_{i=1}^{n} y_i \frac{\sigma(1-\sigma)}{\sigma} x_{i(j)}$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i(\sigma - 1)x_{i(j)}$$

Chain rule:

$$\frac{\partial}{\partial w}\Big[\underbrace{\log \sigma}_{f}\underbrace{(w^T x)}_{z}\Big] = \underbrace{\frac{\partial f}{\partial z}}\,\underbrace{\frac{\partial z}{\partial w}}$$

$$\frac{\partial \log \sigma(z)}{\partial z} = \frac{\sigma'}{\sigma} \qquad \frac{\partial (w^T x)}{\partial w} = x$$

Classification

Gradient descent for logistic regression

- **Loss:**

$$L(w) = -\frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{y_i \log p_w(x_i)}_{\text{RHS1}} + \underbrace{(1-y_i)\log(1-p_w(x_i))}_{\text{RHS2}}\Big)$$

- **Gradient of RHS2:**

$$\begin{aligned}
\frac{\partial}{\partial w_j}\left[-\frac{1}{n}\sum_{i=1}^{n}(1-y_i)\log(1-p_w(x_i))\right] &= -\frac{1}{n}\sum_{i=1}^{n}(1-y_i)\frac{\partial}{\partial w_j}\left[\log(1-\sigma(w^T x_i))\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}(1-y_i)\frac{(1-\sigma)'}{(1-\sigma)}\frac{\partial}{\partial w_j}\left[w^T x_i\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}(1-y_i)\frac{-\sigma(1-\sigma)}{(1-\sigma)}x_{i(j)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(1-y_i)\sigma x_{i(j)}
\end{aligned}$$

Chain rule

$$\frac{\partial}{\partial z}\left[\log(1-\sigma)\right] = \frac{(1-\sigma)'}{(1-\sigma)}$$

$$(1-\sigma)' = -\sigma' = -\sigma(1-\sigma(\eta))$$

Classification

Gradient descent for logistic regression

- **Loss:**

$$L(w) = -\frac{1}{n}\sum_{i=1}^{n}\Big(y_i \log p_w(x_i) + (1-y_i)\log(1-p_w(x_i))\Big)$$

- **Putting gradients together:**

$$\begin{aligned} w_j &\leftarrow w_j - \tau \frac{\partial}{\partial w_j} L(w) \\ &\leftarrow w_j - \tau \frac{1}{n}\sum_{i=1}^{n} (\sigma(w^T x_i) - y_i) x_{i(j)} \\ &\leftarrow w_j - \tau \frac{1}{n}\sum_{i=1}^{n} (p_w(x_i) - y_i) x_{i(j)} \end{aligned}$$

Classification

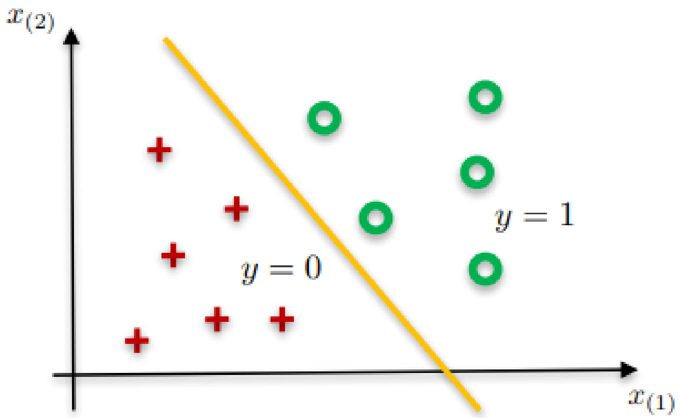
Multi-class problem

- Examples of binary classification tasks:
  - Email: Spam (1) or not spam (0)
  - Online financial transaction: Fraudulent (1) or legitimate (0)

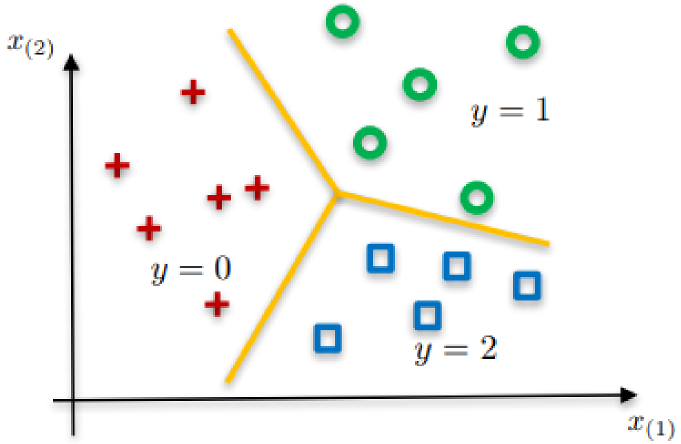
$$y = \{0, 1\}$$ Binary variable

- From binary to multi-class classification:
  - Email: Spam (0), work (1), friends (2), family (3)
  - Medical diseases: Benign (0), malign I (1), malign II (2), malign III (3)

$$y = \{0, 1, 2, ..., K\}$$ Multi-value variable



Binary classification



Multi-variate classification

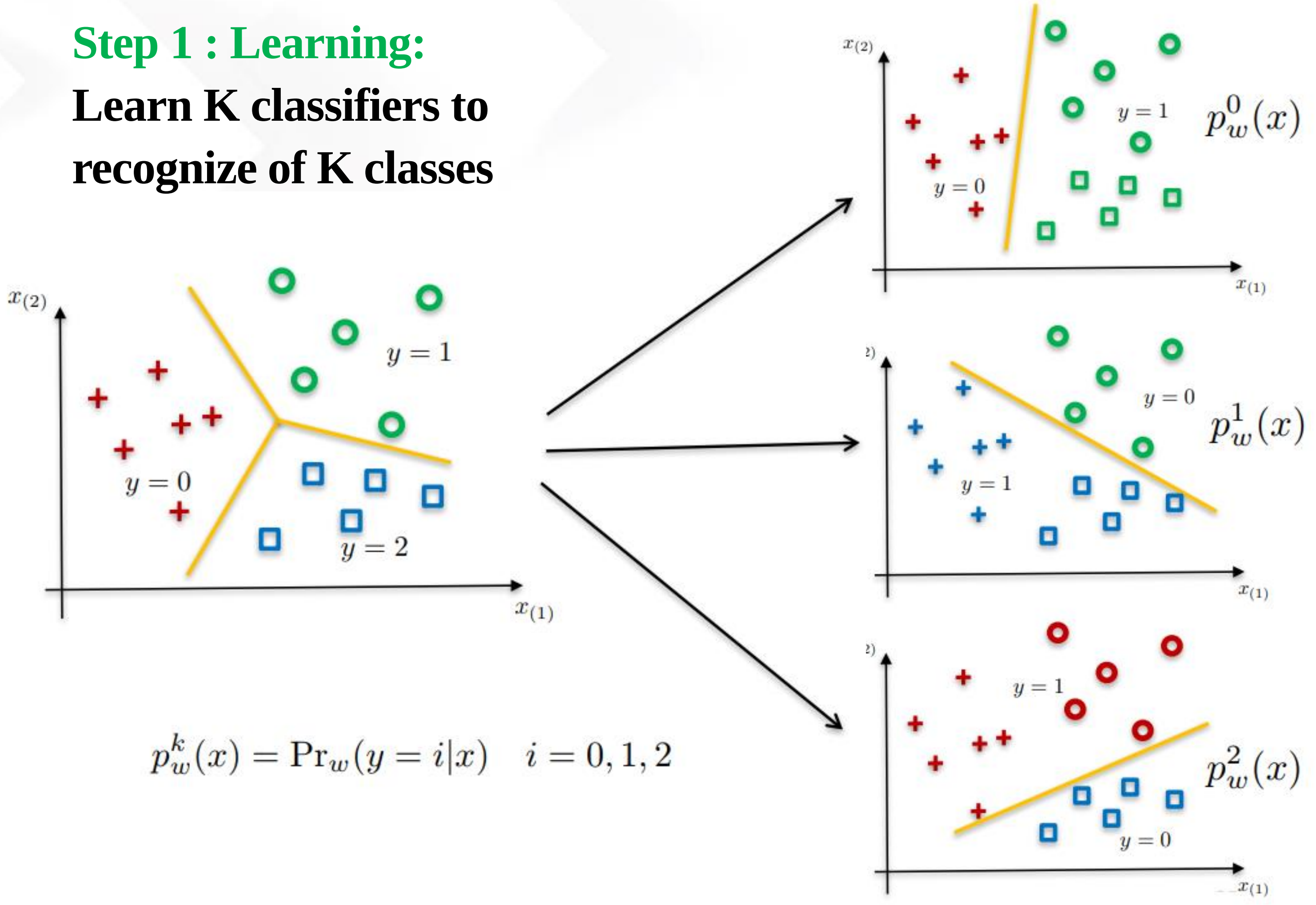
Classification

One-vs-all classification problem

- **Two steps:**

Step 1 : Learning:

Learn K classifiers to recognize of K classes



$$p_w^k(x) = \Pr_w(y = i|x) \quad i = 0, 1, 2$$

One-vs-all classification problem

- **Two steps:**

  Step 2 : Testing: Classify a **new data x** with the class *k* that provides the highest probability:

$$k = \arg\max_c p_w^c(x)$$