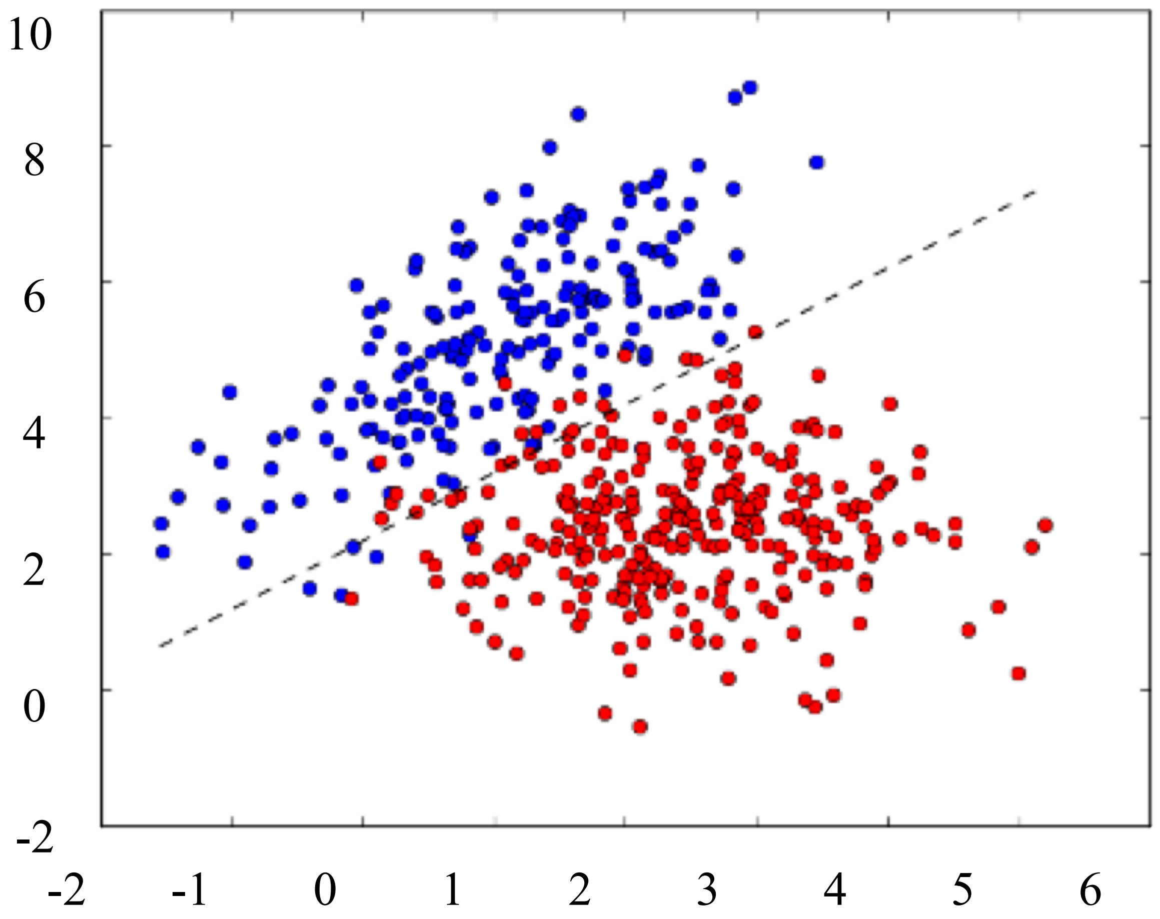
Linear model for classification

Binary classification ($y \in \{-1, 1\}$):

$$a(x) = \text{sign}(w^T x)$$

Number of parameters: d ($w \in \mathbb{R}^d$)

Linear model for classification example



Linear model for classification

Multi-class classification ($y \in \{1, \dots, K\}$):

$$a(x) = \arg \max_{k \in \{1, \dots, K\}} (w_k^T x)$$

Number of parameters: K*d ($w_k \in \mathbb{R}^d$)

Example:

$z = (7, -7.5, 10)$ — scores

$a(x) = 3$

Classification loss

Classification accuracy:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) = y_i]$$

- Not differentiable
- Doesn’t assess model confidence

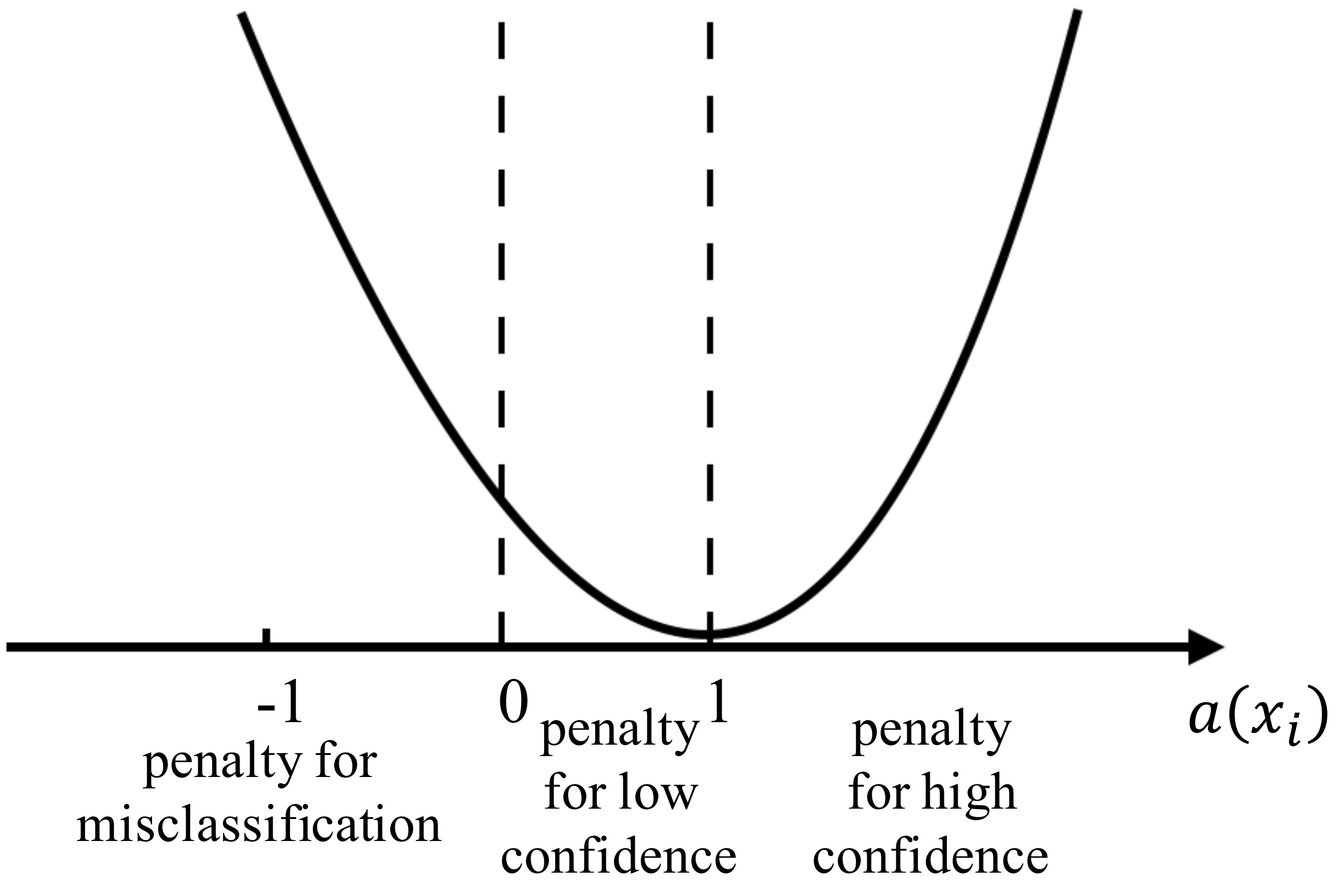
[P] — Iverson bracket:

$$[P] = \begin{cases} 1, & P \text{ is true} \\ 0, & P \text{ is false} \end{cases}$$

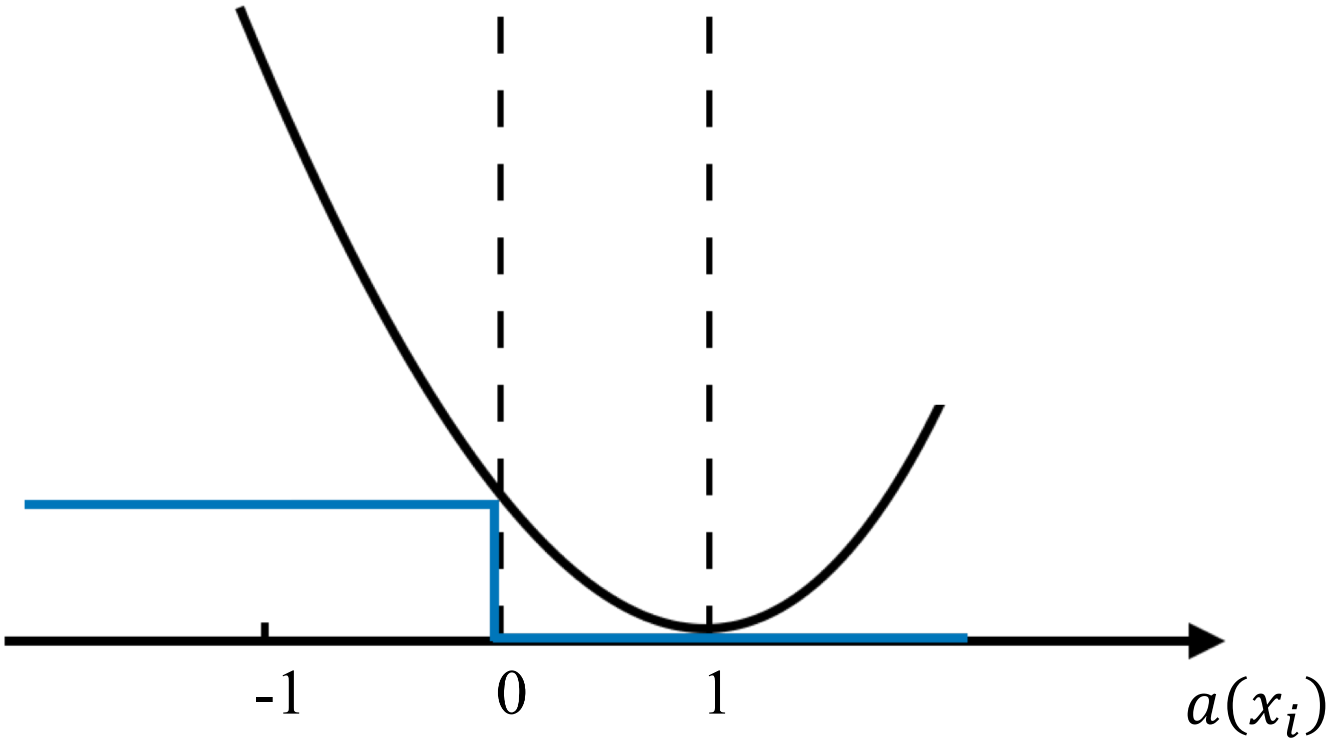
Classification loss

Consider an example x_i such that $y_i = 1$

Squared loss: $(w^T x_i - 1)^2$



Classification loss



Class probabilities

Class scores (**logits**) from a linear model:

$$z = (w_1^T x, \dots, w_K^T x)$$

$$\downarrow$$

$$(e^{z_1}, \dots, e^{z_K})$$

$$\downarrow$$

$$\sigma(z) = \left(\frac{e^{z_1}}{\sum_{k=1}^{K} e^{z_k}}, \dots, \frac{e^{z_K}}{\sum_{k=1}^{K} e^{z_k}} \right)$$

(softmax transform)

Softmax

$$\sigma(z) = \left(\frac{e^{z_1}}{\sum_{k=1}^{K} e^{z_k}}, \ldots, \frac{e^{z_K}}{\sum_{k=1}^{K} e^{z_k}} \right)$$

Example:

$$z = (7, -7.5, 10)$$

$$\sigma(z) \approx (0.05, 0, 0.95)$$

Loss function

Predicted class probabilities (model output):

$$\sigma(z) = \left(\frac{e^{z_1}}{\sum_{k=1}^{K} e^{z_k}}, \dots, \frac{e^{z_K}}{\sum_{k=1}^{K} e^{z_k}}\right)$$

Target values for class probabilities:

$$p = ([y = 1], \dots, [y = K])$$

Similarity between z and p can be measured by the cross-entropy:

$$-\sum_{k=1}^{K} [y = k] \log \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} = -\log \frac{e^{z_y}}{\sum_{j=1}^{K} e^{z_j}}$$

Cross-entropy examples

Suppose $K = 3$ and $y = 1$:

- $-1 * \log 1 - 0 * \log 0 - 0 * \log 0 = 0$
- $-1 * \log 0.5 - 0 * \log 0.25 - 0 * \log 0.25 \approx 0.693$
- $-1 * \log 0 - 0 * \log 1 - 0 * \log 0 = +\infty$

Cross-entropy for classification

Cross-entropy is differentiable and can be used as a loss function:

$$L(w, b) = -\sum_{i=1}^{\ell} \sum_{k=1}^{K} [y_i = k] \log \frac{e^{w_k^T x_i}}{\sum_{j=1}^{K} e^{w_j^T x_i}}$$

$$= -\sum_{i=1}^{\ell} \log \frac{e^{w_{y_i}^T x_i}}{\sum_{j=1}^{K} e^{w_j^T x_i}} \to \min_{w}$$

Summary

- Linear models can be easily generalized for classification tasks
- There are lots of loss functions for classification
- Cross-entropy is one of the most popular