Hinge Loss for Binary Classifiers

专门处理二分类

核心思想：不仅要正确分类，还要确保分类结果有足够的"置信度"

（分类边界有足够的间隔）

Objectives

- introduce disadvantage of squared error for classification

- introduce hinge loss cost function

- characteristics of hinge loss

Squared error "loss" can be problematic

Classifier design

$$\min_{\underline{w}} \ell(\underline{w}; \underline{A}, \underline{d}) + \lambda r(\underline{w})$$

← regularizer

loss function

Penalty for complexity of the model

Squared error loss

$$\ell(\underline{w}; \underline{A}, \underline{d}) = \|\underline{A}\underline{w} - \underline{d}\|_2^2$$

Example: dwarf planet vs. planet

object		Ceres	Eris	Pluto	Mercury	Earth	Jupiter
x_i	radius ($\times 10^6$ m)	1.0	2.3	2.4	4.9	12.8	143.0
d_i	label	-1	-1	-1	1	1	1

加一列1是偏置项想要拟合一条直线，这样x是二维

$$\underline{A} = \begin{bmatrix} 1 & 1 \\ 2.3 & 1 \\ 2.4 & 1 \\ 4.9 & 1 \\ 12.8 & 1 \\ 143 & 1 \end{bmatrix}, \quad \underline{d} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

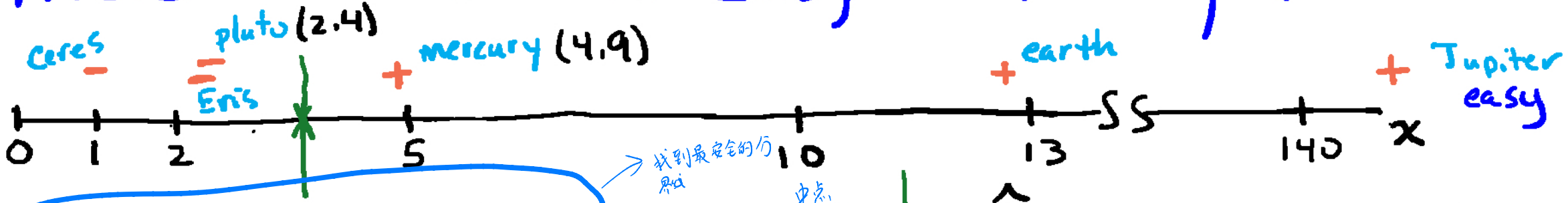
$$\underline{w}_{LS} = (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d} \approx 0.01 \begin{bmatrix} 1 \\ -28 \end{bmatrix}$$

dwarf: $x_i < 28$ (earth!)

planet: $x_i > 28$ (平方误差对异常值很敏感)

squared error → poor classification

Avoid loss due to "easy-to-classify" data



找到最安全的分界

max margin classifier: midpoint (中点)

$$\hat{d} = \text{sign}(x - 3.65)$$

Margin $4.9 - 2.4 = 2.5$ (class separation)

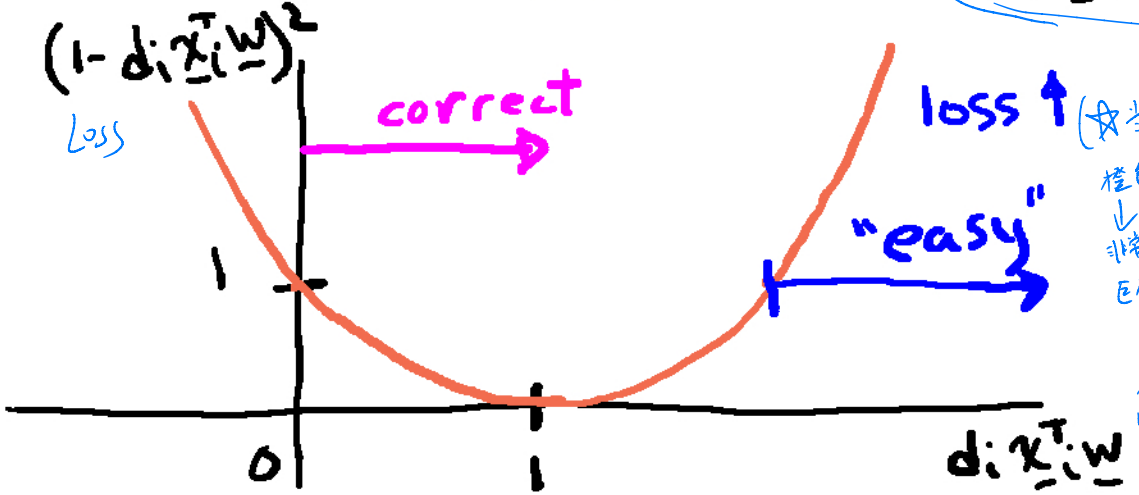
Squared error loss: $\|Aw - d\|_2^2 = \sum_{i=1}^{N} (d_i - x_i^T w)^2 = \sum_i (1 - d_i x_i^T w)^2$

Since d is ± 1, we can rotate the triangle like this

correct classification: $d_i x_i^T w > 0$ → Because $d_i x_i^T w = 1 \cdot x_i^T w$ means correct, and the loss $1 - d_i x_i^T w$ is small

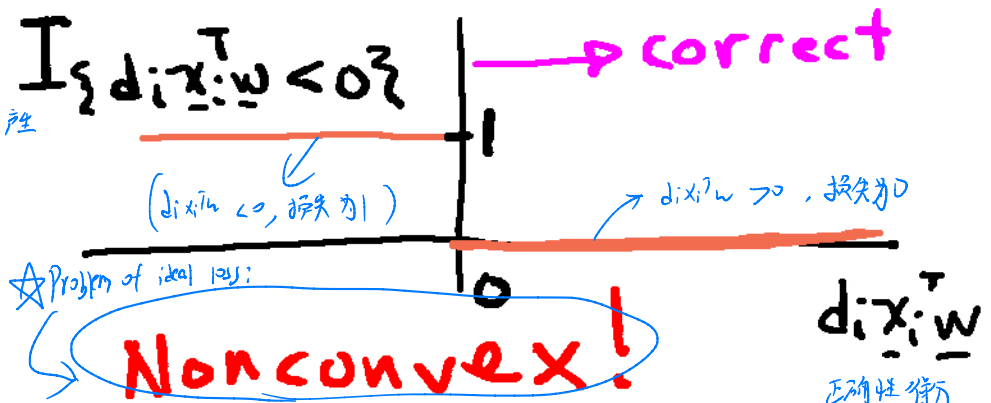
$(d_i = \pm 1)$

Ideal loss



(当 $d_i x_i^T w$ 越来越大，橙色抛物线开始向上走，非常容易被分类的点也会产生巨大的损失)

正确性得分



Problem of ideal loss:

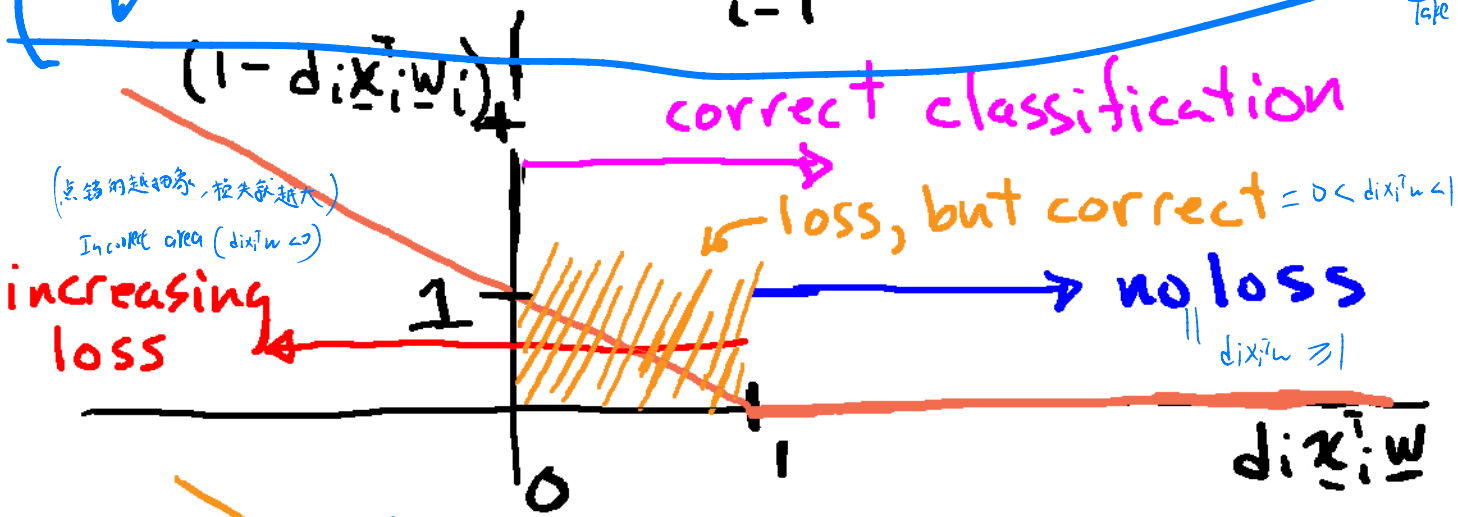
Nonconvex!

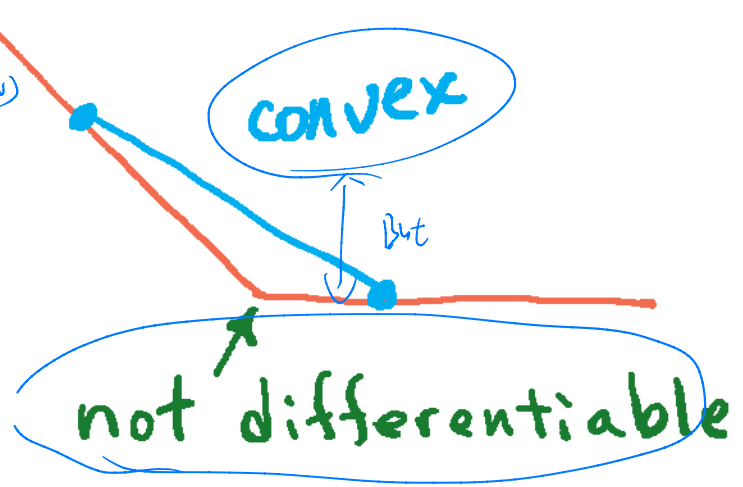
Hinge loss is convex and has no loss for easy to classify data

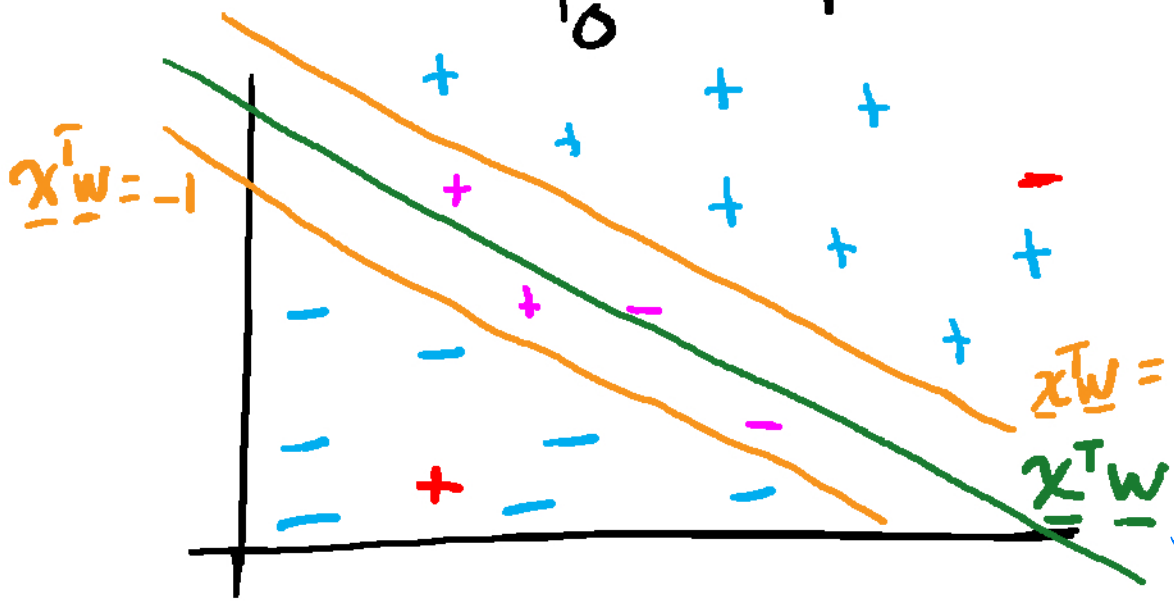
$$\ell(\underline{w}; \underline{A}, \underline{d}) = \sum_{i=1}^{N} (1 - d_i \underline{x}_i^T \underline{w})_+$$

☆ $L_H(x) = \max(0, 1 - d_i f(w))$

Take the positive part $(\alpha)_+ = \begin{cases} \alpha & \alpha > 0 \\ 0 & \alpha \leq 0 \end{cases}$







+, − no hinge loss → 太对了

+, − small hinge loss → 离决策边界太近，不够自信，给小loss (penalty)

+, − large hinge loss → 完全不对

Hinge loss better approximates ideal: number of misclassifications

Ideal, Hinge loss approximates this

Iterative algorithms required for finding minimum hinge loss classifier

convex, but we cannot solve in closed form