L2 正则化以及反对抗能力

L2正则化是什么

- 在标准损失后面添加penalty
- 作用：
  - 惩罚权重网络，防止过度拟合
  - 还可以做什么？

$$C = C_0 + \frac{\lambda}{2n}\sum_w w^2,$$

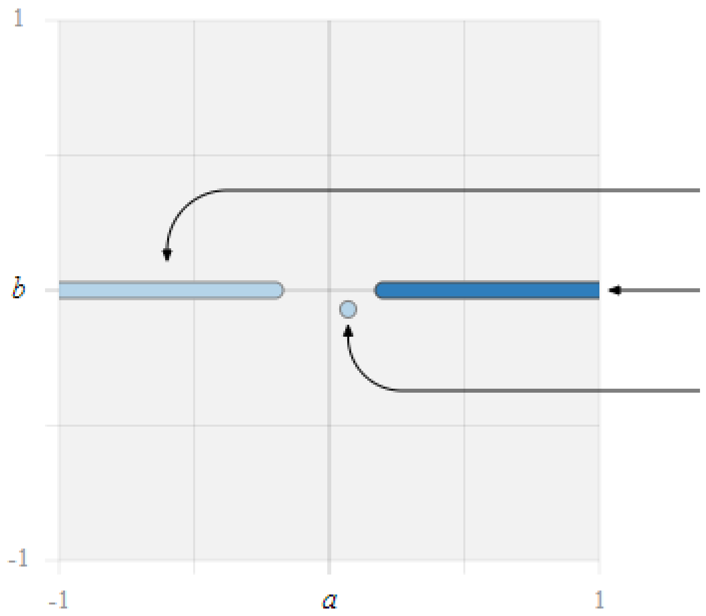
$$\frac{\partial C}{\partial w} = \frac{\partial C_0}{\partial w} + \frac{\lambda}{n} w$$

$$\frac{\partial C}{\partial b} = \frac{\partial C_0}{\partial b}.$$

$$w \rightarrow w - \eta \frac{\partial C_0}{\partial w} - \frac{\eta\lambda}{n} w$$

$$= \left(1 - \frac{\eta\lambda}{n}\right) w - \eta \frac{\partial C_0}{\partial w}.$$

L2和对抗性，以及推理



Training set consisting of:

A large number of images in class I

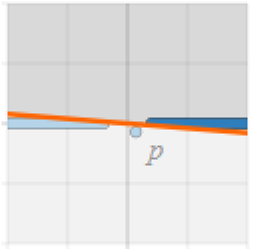
A large number of images in class J

The data point p wrongly labelled in I

Without L2 regularization:

The classification boundary is strongly tilted.

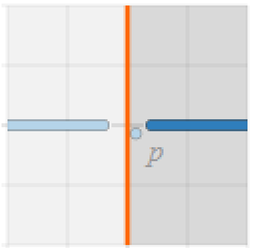
Most of the leeway available to fit the training data resides in the tilting angle of the boundary. Here, the data point p can be classified correctly, but the classifier obtained is then vulnerable to adversarial examples.



With L2 regularization:

The classification boundary is not tilted.

L2 regularization reduces overfitting by allowing some training samples to be misclassified. When enough regularization is used, the data point p is ignored and the classifier obtained is robust to adversarial examples.



几个定义

$$s(\boldsymbol{x}) := \boldsymbol{w}\cdot\boldsymbol{x} + b$$

$$\boldsymbol{x} \text{ is classified in } \left| \begin{array}{l} I \text{ if } s(\boldsymbol{x}) \leq 0 \\ J \text{ if } s(\boldsymbol{x}) \geq 0 \end{array} \right.$$

$$R(\boldsymbol{w}, b) := \frac{1}{n} \sum_{(\boldsymbol{x},y)\in T} f(y\, s(\boldsymbol{x}))$$

$$d(\boldsymbol{x}) := \hat{\boldsymbol{w}}\cdot\boldsymbol{x} + b' \qquad \text{where} \qquad \hat{\boldsymbol{w}} := \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \quad b' := \frac{b}{\|\boldsymbol{w}\|}$$

$$\text{and} \quad s(\boldsymbol{x}) = \|\boldsymbol{w}\|\, d(\boldsymbol{x})$$

Hence, the norm $\|\boldsymbol{w}\|$ can be interpreted as a scaling parameter for the loss function in the expression of the empirical risk:

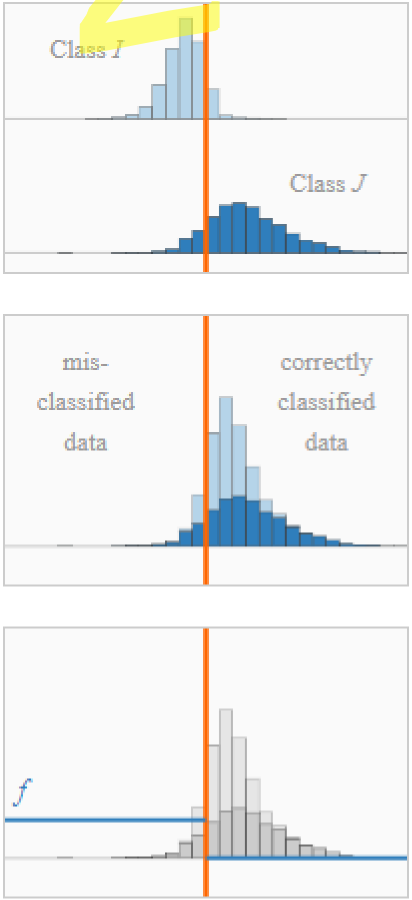
$$R(\boldsymbol{w}, b) = \frac{1}{n} \sum_{(\boldsymbol{x},y)\in T} f(\underbrace{\|\boldsymbol{w}\|}_{\text{scaling parameter for } f} \times y\, d(\boldsymbol{x}))$$

Let us define the *scaled loss function* $f_{\|\boldsymbol{w}\|} : z \to f(\|\boldsymbol{w}\| \times z)$.

$s(\boldsymbol{x})$

$y\, s(\boldsymbol{x})$

$f(y\, s(\boldsymbol{x}))$



If we plot the histograms of the raw scores over the training set, we typically get two clusters of opposite signs

Multiplying by the label allows us to distinguish the correctly classified data from the misclassified data

We can then attribute a penalty to each training point $\boldsymbol{x}$ by applying a *loss function* to $y\, s(\boldsymbol{x})$

WHEN || W || IS LARGE

$$f_{\|w\|}(y\,d(x)) \underset{\|w\|\to+\infty}{\approx} \|w\| \max(-y\,d(x), 0)$$

enience, we name the set of misclassified data:

$$M := \{(x, y) \in T \mid y\,d(x) \leq 0\}$$

an then write the empirical risk as:

$$R(w, b) \underset{\|w\|\to+\infty}{\approx} \|w\| \left(\frac{1}{n} \sum_{(x,y)\in M} (-\,y\,d(x))\right)$$

ession contains a term which we call the *error distance*:

$$d_{\text{err}} := \frac{1}{n} \sum_{(x,y)\in M} (-\,y\,d(x))$$

ive and can be interpreted as the average distance by which
ning sample is misclassified by $\mathcal{C}$ (with a null contribution for
ctly classified data). It is related—although not exactly
nt—to the training error.[3]

e have:

$$\text{minimize: } R(w, b) \underset{\|w\|\to+\infty}{\Longleftrightarrow} \text{minimize: } d_{\text{err}}$$

$$d(x) := \hat{w}\cdot x + b' \qquad \text{where} \qquad \hat{w} := \frac{w}{\|w\|} \qquad b' := \frac{b}{\|w\|}$$

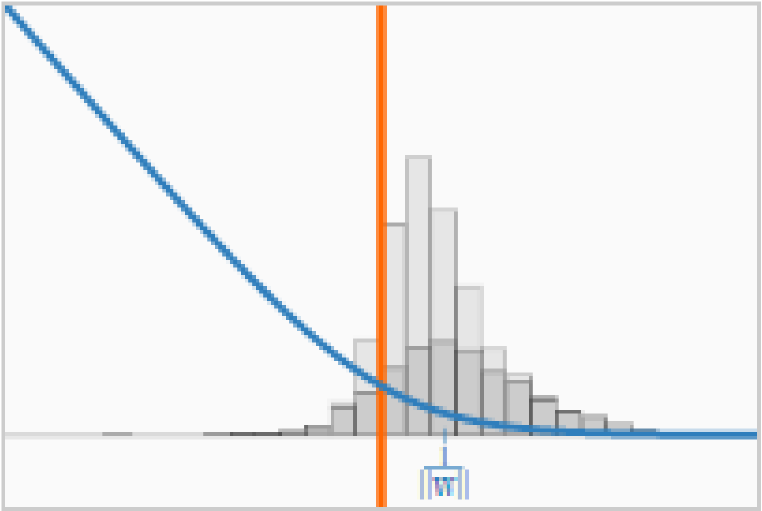
$$\text{and} \quad s(x) = \|w\|\, d(x)$$

Hence, the norm $\|w\|$ can be interpreted as a scaling parameter for the loss function in the expression of the empirical risk:

$$R(w, b) = \frac{1}{n} \sum_{(x,y)\in T} f(\,\underbrace{\|w\|}_{\text{scaling parameter for } f} \times\, y\,d(x))$$

Let us define the *scaled loss function* $f_{\|w\|} : z \to f(\|w\| \times z)$.

softplus loss



WHEN ||W|| IS SMALL

More precisely, both losses satisfy:[4]

$$f_{\|\boldsymbol{w}\|}(y\,d(\boldsymbol{x})) \underset{\|\boldsymbol{w}\|\to 0}{\approx} \alpha - \beta\,\|\boldsymbol{w}\|\,y\,d(\boldsymbol{x})$$

for some positive values α and β.

We can then write the empirical risk as:

$$R(\boldsymbol{w},b) \underset{\|\boldsymbol{w}\|\to 0}{\approx} \alpha - \beta\,\|\boldsymbol{w}\| \left(\frac{1}{n}\sum_{(\boldsymbol{x},y)\in T} y\,d(\boldsymbol{x})\right)$$

This expression contains a term which we call the *adversarial distance*:

$$d_{\text{adv}} := \frac{1}{n}\sum_{(\boldsymbol{x},y)\in T} y\,d(\boldsymbol{x})$$

It is the mean distance between the images in T and the classification boundary $\mathcal{C}$ (with a negative contribution for the misclassified images). It can be viewed as a measure of robustness to adversarial perturbations: when d_{adv} is high, the number of misclassified images is limited and the correctly classified images are far from $\mathcal{C}$.

Finally we have:

$$\text{minimize: } R(\boldsymbol{w},b) \underset{\|\boldsymbol{w}\|\to 0}{\iff} \text{maximize: } d_{\text{adv}}$$

$$d(\boldsymbol{x}) := \hat{\boldsymbol{w}}\cdot\boldsymbol{x} + b' \qquad \text{where} \qquad \hat{\boldsymbol{w}} := \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \qquad b' := \frac{b}{\|\boldsymbol{w}\|}$$

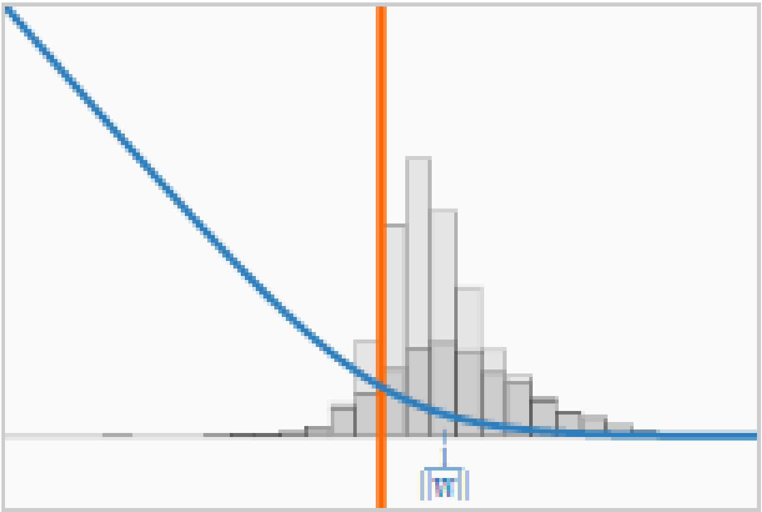
$$\text{and} \quad s(\boldsymbol{x}) = \|\boldsymbol{w}\|\,d(\boldsymbol{x})$$

Hence, the norm $\|\boldsymbol{w}\|$ can be interpreted as a scaling parameter for the loss function in the expression of the empirical risk:

$$R(\boldsymbol{w},b) = \frac{1}{n}\sum_{(\boldsymbol{x},y)\in T} f(\underbrace{\|\boldsymbol{w}\|}_{\text{scaling parameter for } f} \times\, y\,d(\boldsymbol{x}))$$

Let us define the *scaled loss function* $f_{\|\boldsymbol{w}\|} : z \to f(\|\boldsymbol{w}\| \times z)$.

softplus loss



- λ值越大，||W||越小，分类效果越差
  - 同时惩罚了正确分类，但是保证了正确分类离分类边界更远
- λ值越小，||W||越大，分类效果越好
  - 仅仅只惩罚错误分类

$$L(\boldsymbol{w}, b) := \underbrace{R(\boldsymbol{w}, b)}_{\text{empirical risk}} + \underbrace{\lambda \|\boldsymbol{w}\|^2}_{\text{L2 regularization}}$$

APPENDIX

- https://thomas-tanay.github.io/post--L2-regularization/