

《神经网络与深度学习》



线性模型

<https://nndl.github.io/>

大纲

- ▶ 分类问题示例
- ▶ 线性分类模型
 - ▶ Logistic Regression
 - ▶ Softmax Regression
 - ▶ Perceptron
 - ▶ SVM



分类示例

示例：图像分类

►数据集：CIFAR-10

►60000 张 32x32 色彩图像，共 10 类

►每类 6000 张图像

airplane



automobile



bird



cat



deer



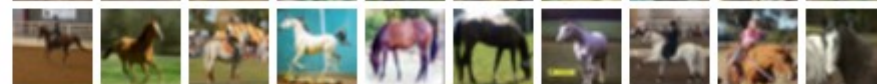
dog



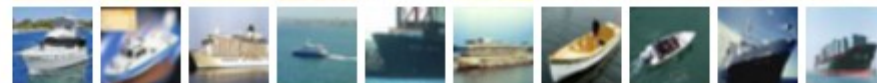
frog



horse



ship



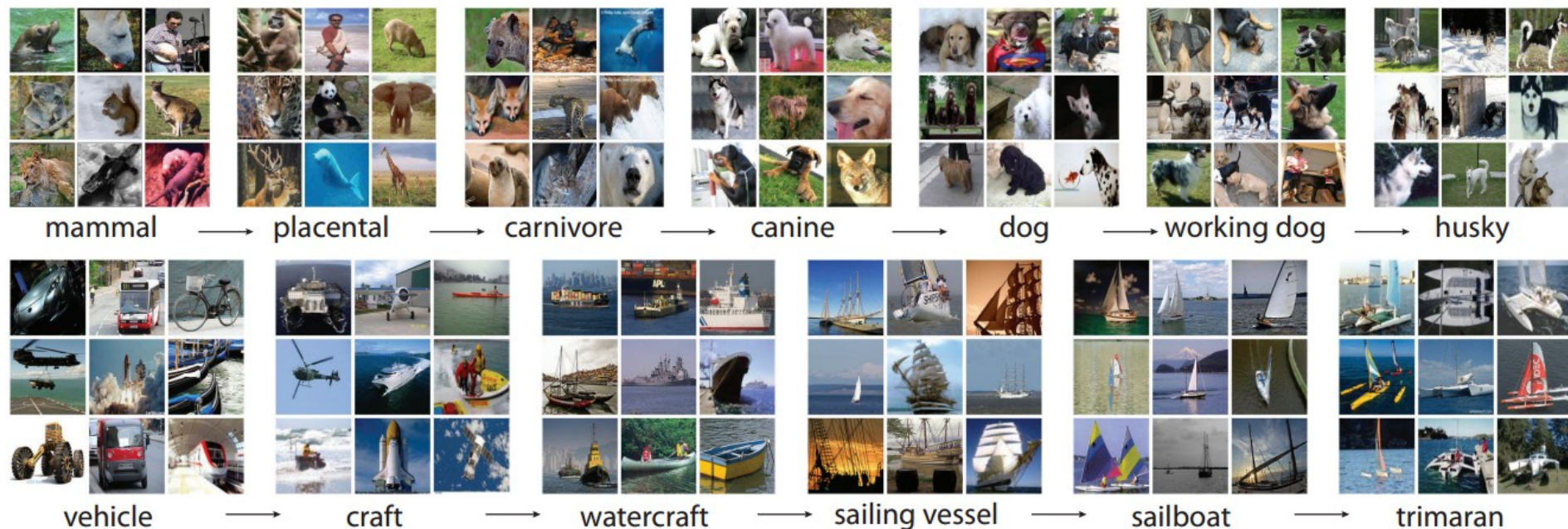
truck



示例：图像分类

►数据集： ImageNet

►14,197,122 images, 21841 synsets



示例：图像分类

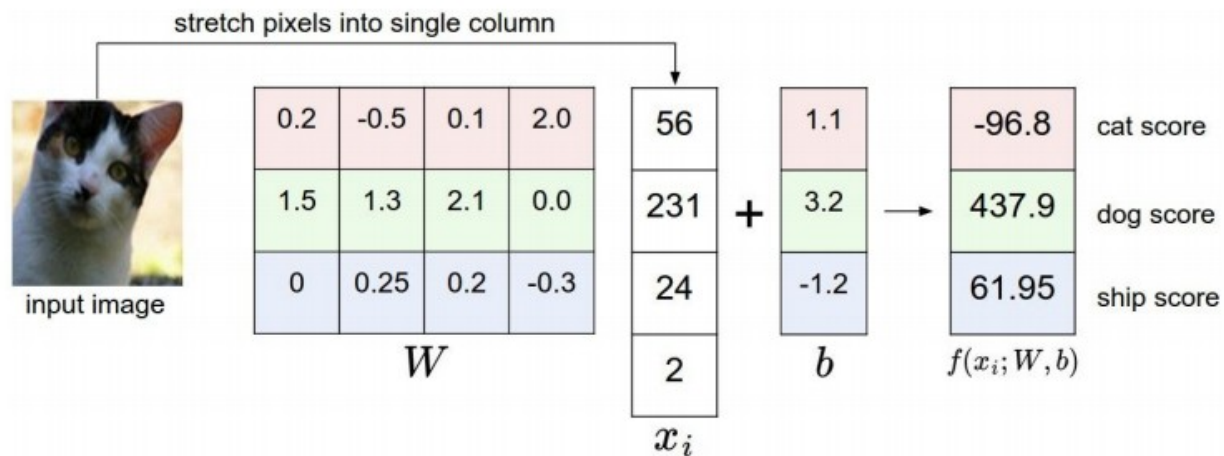
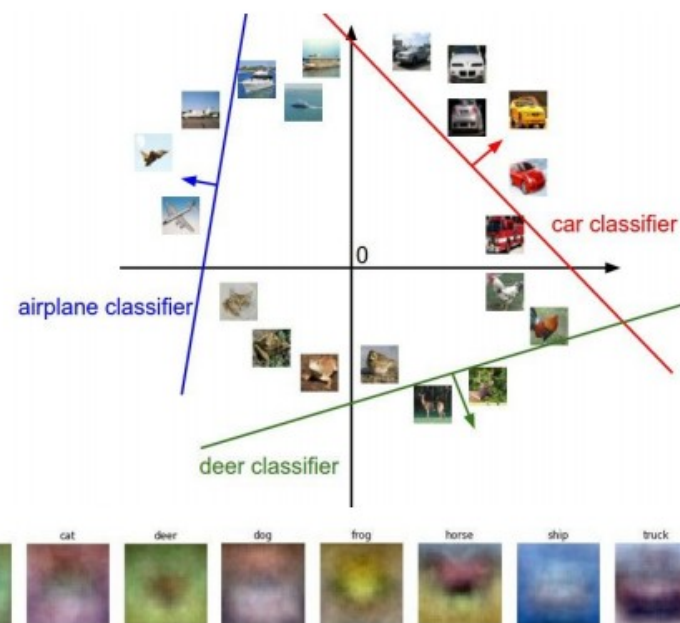


[32x32x3]

array of numbers 0...1
(3072 numbers total)

image parameters
 $f(\mathbf{x}, \mathbf{W})$

10 numbers, indicating
class scores



示例：图像分类、目标检测、实例分割

Classification



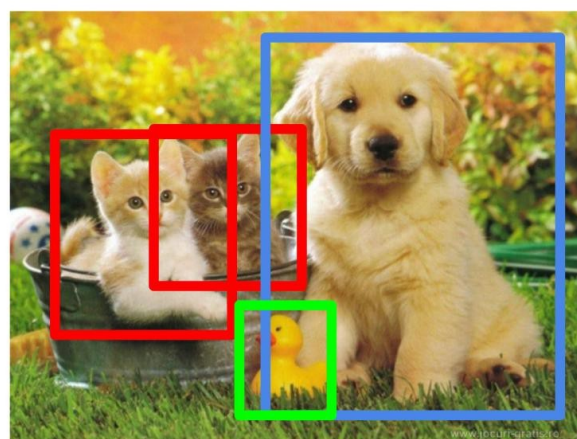
CAT

**Classification
+ Localization**



CAT

Object Detection



CAT, DOG, DUCK

**Instance
Segmentation**



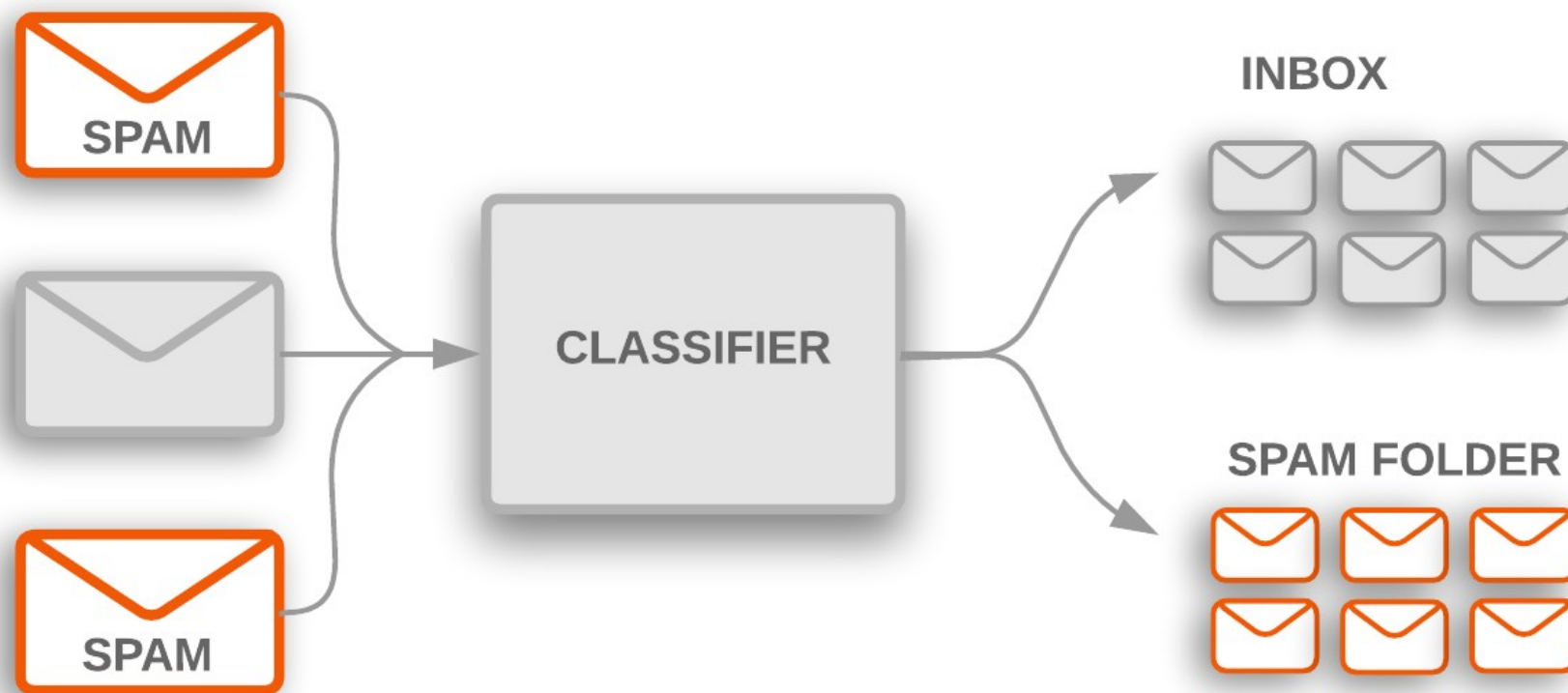
CAT, DOG, DUCK

Single object

Multiple objects

<https://medium.com/zylapp/review-of-deep-learning-algorithms-for-object-detection-c1f3d437b852>

示例：垃圾邮件过滤

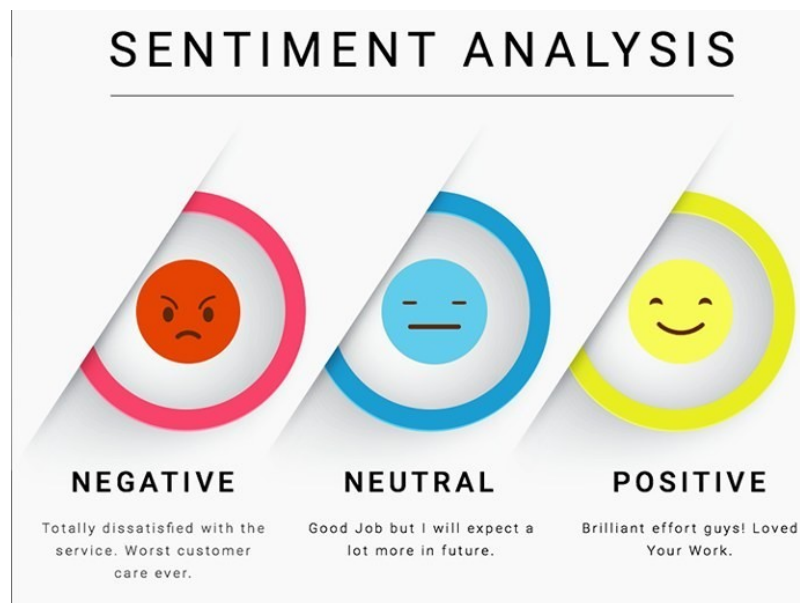


示例：文档归类



<https://towardsdatascience.com/automated-text-classification-using-machine-learning-3df4f4f9570b>

示例：情感分类



Review (X)

Rating (Y)

"This movie is fantastic! I really like it because it is so good!"



"Not to my taste, will skip and watch another movie"



"This movie really sucks! Can I get my money back please?"



示例：文本分类

- ▶ 将样本 x 从文本形式转为向量形式
- ▶ 词袋模型（Bag-of-Words，BoW）模型

the dog is on the table

0	0	1	1	0	1	1	1
are	cat	dog	is	now	on	table	the

比如两个文本“我 喜欢 读书”和“我 讨厌 读书”中共有“我”、“喜欢”、“讨厌”、“读书”四个词，它们的BoW表示分别为

$$\mathbf{v}_1 = [1 \ 1 \ 0 \ 1]^T,$$

$$\mathbf{v}_2 = [1 \ 0 \ 1 \ 1]^T.$$

示例：文本情感分类

根据文本内容来判断文本的相应类别

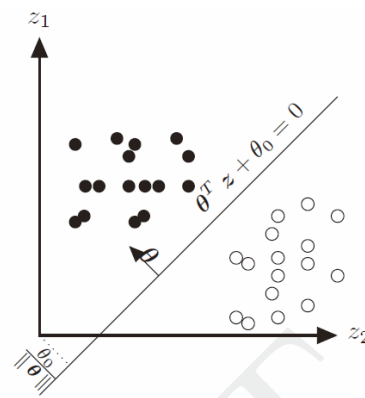
D_1 : “我喜欢读书”

D_2 : “我讨厌读书”

	我	喜欢	讨厌	读书
D_1	1	1	0	1
D_2	1	0	1	1

+

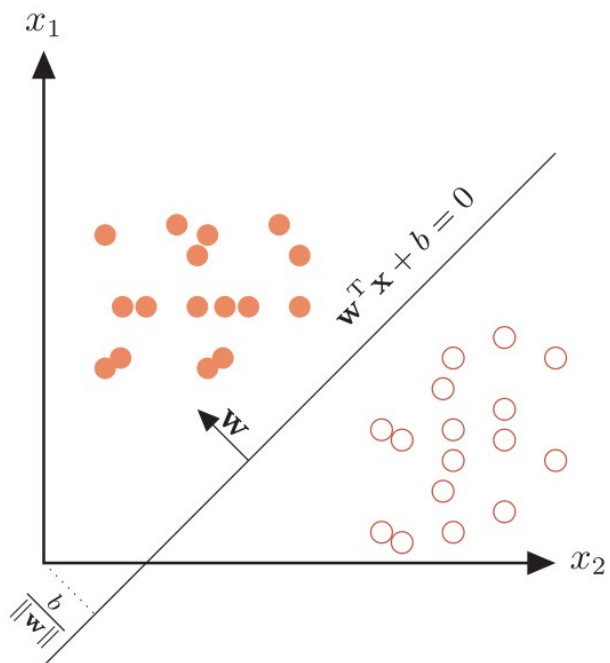
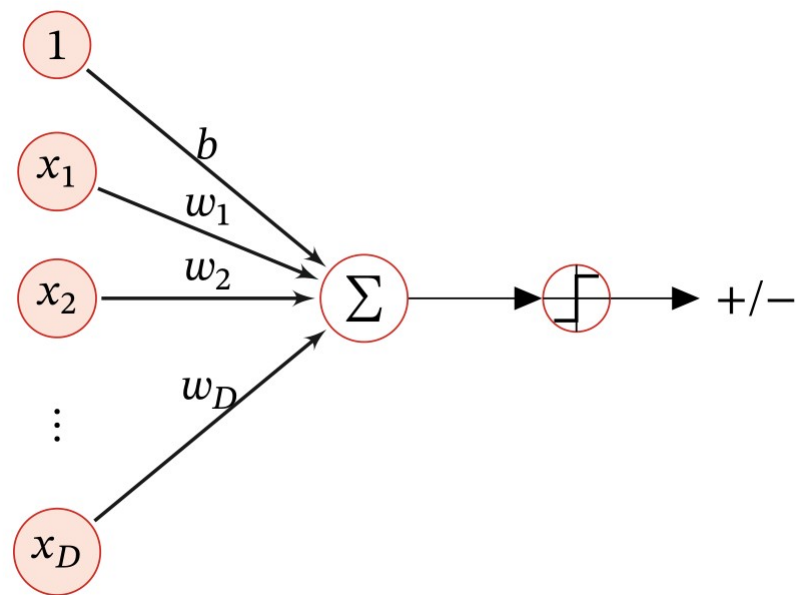
-





线性分类模型

线性模型



线性模型

- ▶ Logistic 回归
- ▶ Softmax 回归
- ▶ 感知器
- ▶ 支持向量机



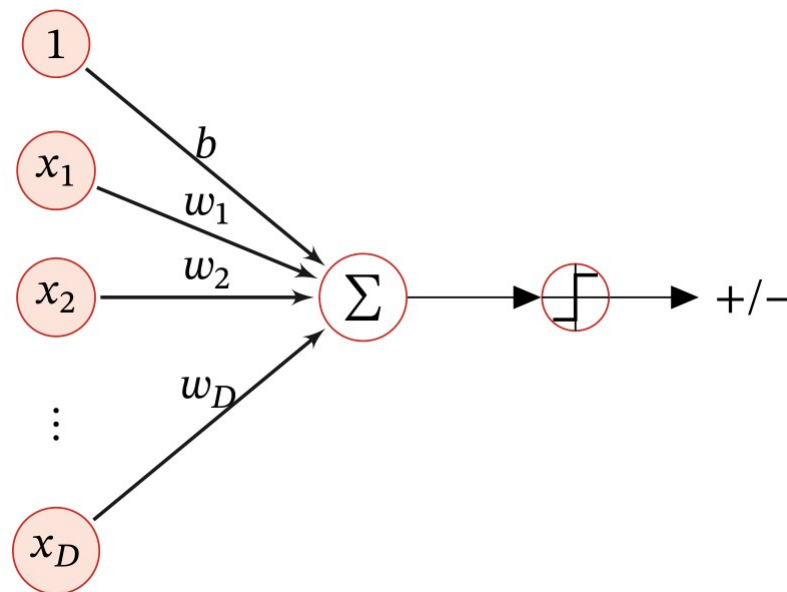
Logistic Regression

Logistic Regression

► 模型

$$g(f(\mathbf{x}; \mathbf{w})) = \begin{cases} 1 & \text{if } f(\mathbf{x}; \mathbf{w}) > 0 \\ 0 & \text{if } f(\mathbf{x}; \mathbf{w}) < 0 \end{cases}$$

损失函数？



分类问题

▶ 将分类问题看作条件概率估计问题

- ▶ 引入非线性函数 g 来预测类别标签的条件概率 $p(y = c|x)$ 。
- ▶ 以二分类为例，

$$p(y = 1|\mathbf{x}) = g(f(\mathbf{x}; \mathbf{w}))$$

- ▶ 函数 f : 线性函数
- ▶ 函数 g : 把线性函数的值域从实数区间“挤压”到了 $(0,1)$ 之间，可以用来表示概率。

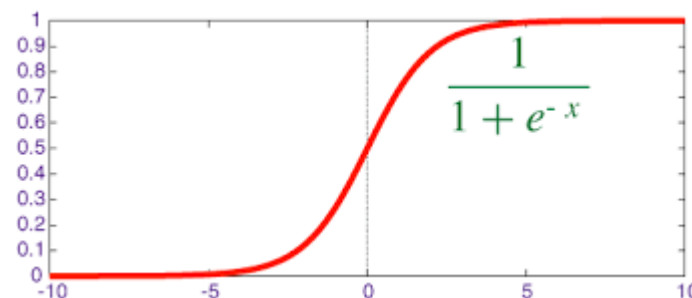
如何构造函数 g ?

Logistic 函数与回归

► Logistic 函数

除了logistic, 还有什么函数 $f: \mathbb{R} \rightarrow (0,1)$?

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$



► Logistic 回归

$$\begin{aligned} p(y = 1|\mathbf{x}) &= \sigma(\mathbf{w}^\top \mathbf{x}) \\ &\triangleq \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})} \end{aligned}$$

学习准则

► 模型预测条件概率 $p_{\theta}(y|\mathbf{x})$

$$p_{\theta}(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x})$$

► 真实条件概率 $p_r(y|\mathbf{x})$

► 对于一个样本 $(\mathbf{x}, y^*)$, 其真实条件概率为

$$\begin{aligned} p_r(y = 1|\mathbf{x}) &= y^* \\ p_r(y = 0|\mathbf{x}) &= 1 - y^* \end{aligned}$$

如何衡量两个条件分布的差异？

熵 (Entropy)

►在信息论中，熵用来衡量一个随机事件的不确定性。

►自信息 (Self Information)

$$I(x) = -\log(p(x))$$

►熵

$$H(X) = \mathbb{E}_X[I(x)]$$

$$= \mathbb{E}_X[-\log p(x)]$$

$$= - \sum_{x \in \mathcal{X}} p(x) \log p(x)$$

►熵越高，则随机变量的信息越多；

►熵越低，则随机变量的信息越少。

►在对分布 $q(y)$ 的符号进行编码时，熵 $I(q)$ 也是理论上最优的平均编码长度，这种编码方式称为熵编码 (Entropy Encoding)

交叉熵 (Cross Entropy)

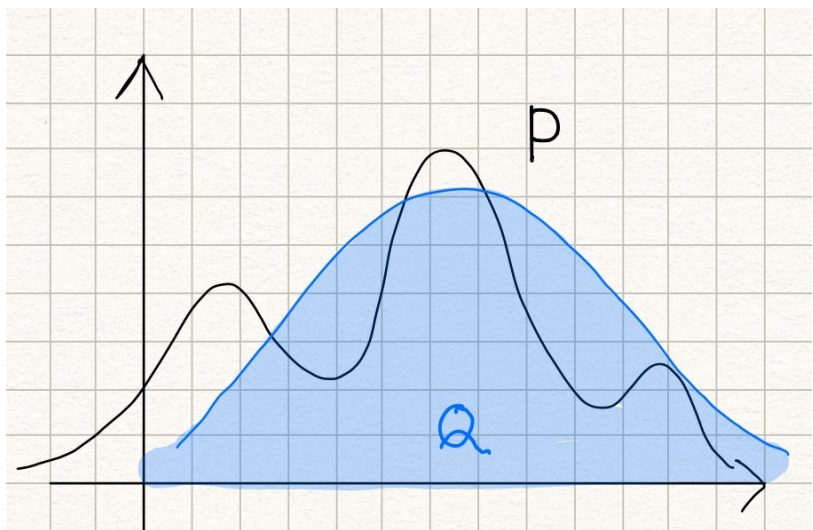
- ▶ 交叉熵是按照概率分布 q 的最优编码对真实分布为 p 的信息进行编码的长度。

$$\begin{aligned} H(p, q) &= \mathbb{E}_p[-\log q(x)] \\ &= -\sum_x p(x) \log q(x) \end{aligned}$$

- ▶ 在给定 q 的情况下，如果 p 和 q 越接近，交叉熵越小；
- ▶ 如果 p 和 q 越远，交叉熵就越大。

KL 散度 (Kullback-Leibler Divergence)

- ▶ **KL 散度是用概率分布 q 来近似 p 时所造成的信息损失量。**
- ▶ KL 散度是按照概率分布 q 的最优编码对真实分布为 p 的信息进行编码，其平均编码长度（即交叉熵） $H(p, q)$ 和 p 的最优平均编码长度（即熵） $H(p)$ 之间的差异。



$$\begin{aligned} \text{KL}(p, q) &= H(p, q) - H(p) \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} \end{aligned}$$

交叉熵损失

$$D_{KL}(p_r(y|x)||p_\theta(y|x)) = \sum_{y=0}^1 p_r(y|x) \log \frac{p_r(y|x)}{p_\theta(y|x)}$$

KL 散度

$$\propto - \sum_{y=0}^1 p_r(y|x) \log p_\theta(y|x)$$

交叉熵损失

y^* 为 x 的真实标签

$$= -I(y^* = 1) \log p_\theta(y = 1|x) - I(y^* = 0) \log p_\theta(y = 0|x)$$

$$= -y^* \log p_\theta(y = 1|x) - (1 - y^*) \log p_\theta(y = 0|x)$$

$$= -\log p_\theta(y^*|x)$$

负对数似然

梯度下降

- ▶ 交叉熵损失函数，模型在训练集的风险函数为

$$\mathcal{R}(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^N \left(y^{(i)} \log \left(\sigma(\mathbf{w}^T \mathbf{x}^{(i)}) \right) + (1 - y^{(i)}) \log \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) \right) \right).$$

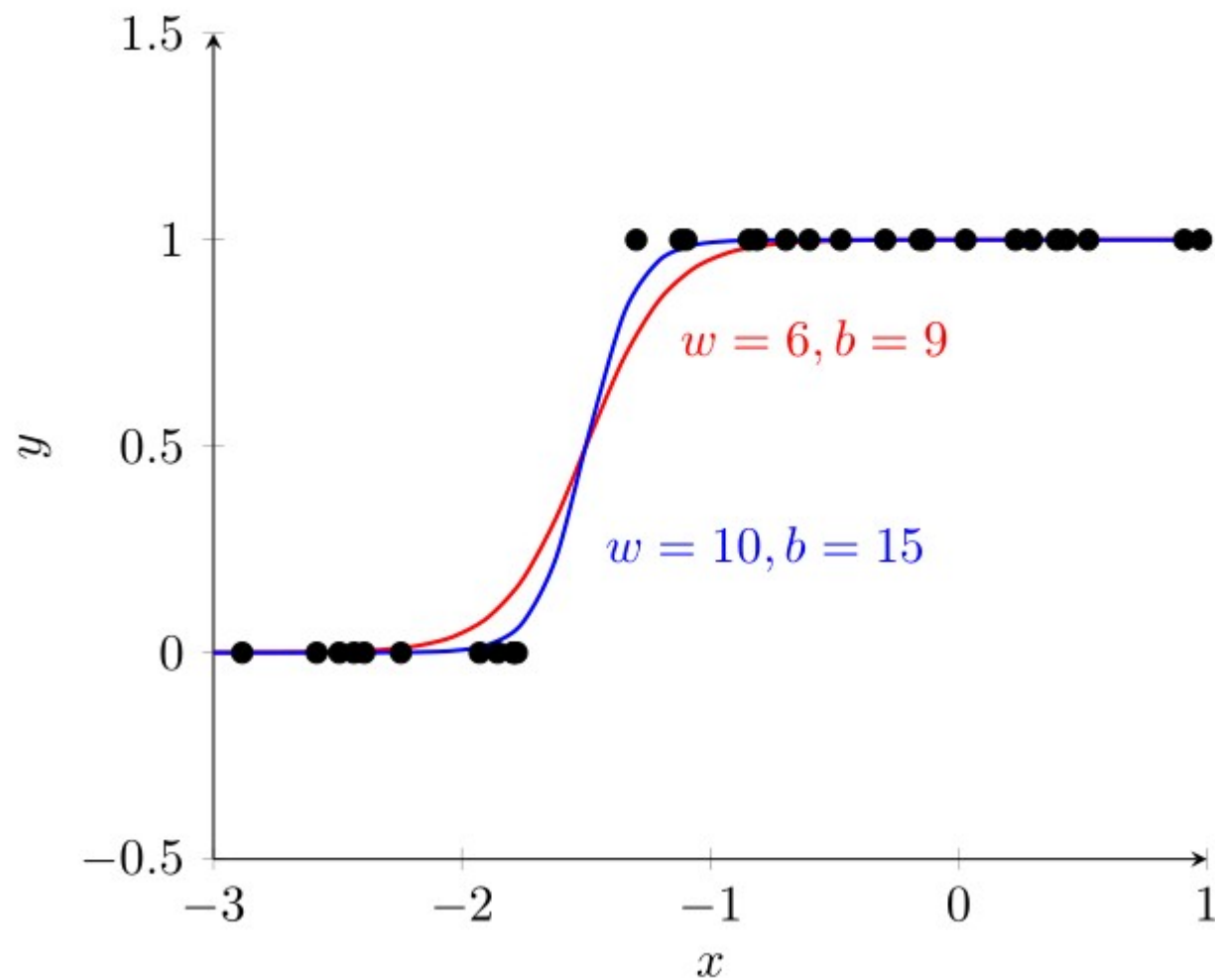
- ▶ 梯度为

$$\frac{\partial \mathcal{R}(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{N} \sum_{i=1}^N \left(\mathbf{x}^{(i)} \cdot \left(\sigma(\mathbf{w}^T \mathbf{x}^{(i)}) - y^{(i)} \right) \right)$$

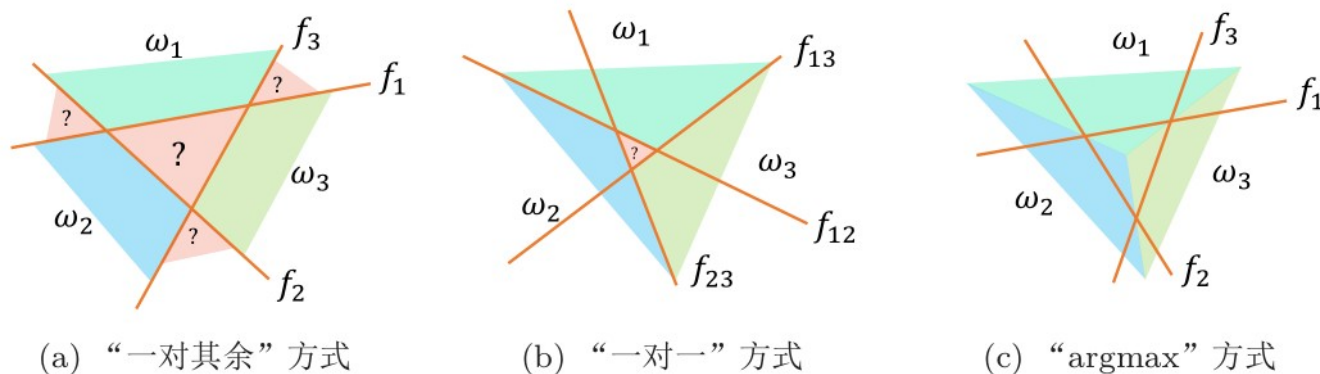
推导过程

$$\begin{aligned}\frac{\partial \mathcal{R}(\mathbf{w})}{\partial \mathbf{w}} &= -\frac{1}{N} \sum_{n=1}^N \left(y^{(n)} \frac{\hat{y}^{(n)}(1 - \hat{y}^{(n)})}{\hat{y}^{(n)}} \mathbf{x}^{(n)} - (1 - y^{(n)}) \frac{\hat{y}^{(n)}(1 - \hat{y}^{(n)})}{1 - \hat{y}^{(n)}} \mathbf{x}^{(n)} \right) \\ &= -\frac{1}{N} \sum_{n=1}^N \left(y^{(n)}(1 - \hat{y}^{(n)}) \mathbf{x}^{(n)} - (1 - y^{(n)}) \hat{y}^{(n)} \mathbf{x}^{(n)} \right) \\ &= -\frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)} (y^{(n)} - \hat{y}^{(n)}).\end{aligned}$$

Logistic 回归



多分类 (Multi-class Classification)



“*argmax*” 方式：这是一种改进的“一对其余”方式，共需要 C 个判别函数

$$f_c(\mathbf{x}; \mathbf{w}_c) = \mathbf{w}_c^T \mathbf{x} + b_c, \quad c = [1, \dots, C] \quad (3.10)$$

如果存在类别 c , 对于所有的其他类别 $\tilde{c} (\tilde{c} \neq c)$ 都满足 $f_c(\mathbf{x}; \mathbf{w}_c) > f_{\tilde{c}}(\mathbf{x}; \mathbf{w}_{\tilde{c}})$, 那么 $\mathbf{x}$ 属于类别 c 。即

$$y = \arg \max_{c=1}^C f_c(\mathbf{x}; \mathbf{w}_c). \quad (3.11)$$

Softmax 回归

► 多分类问题

$$y = \arg \max_{c=1}^C f_c(\mathbf{x}; \mathbf{w}_c)$$

► Softmax 函数

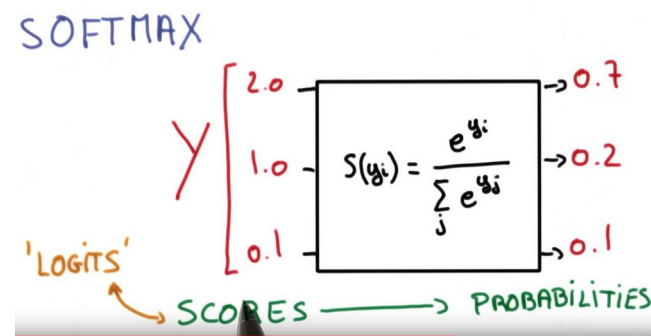
$$\text{softmax}(x_k) = \frac{\exp(x_k)}{\sum_{i=1}^K \exp(x_i)}$$



Softmax 回归

- 利用 softmax 函数，目标类别 $y = c$ 的条件概率为：

$$\begin{aligned} P(y = c|\mathbf{x}) &= \text{softmax}(\mathbf{w}_c^\top \mathbf{x}) \\ &= \frac{\exp(\mathbf{w}_c^\top \mathbf{x})}{\sum_{i=1}^C \exp(\mathbf{w}_i^\top \mathbf{x})} \end{aligned}$$



交叉熵损失

► KL 散度

$$D_{\text{kl}}(p_r(y|x) || p_{\theta}(y|x)) = \sum_{y=1}^C p_r(y|x) \log \frac{p_r(y|x)}{p_{\theta}(y|x)}$$

$$\propto - \sum_{y=1}^C p_r(y|x) \log p_{\theta}(y|x) \quad \text{交叉熵损失}$$

$$y^* \text{ 为 } x \text{ 的真实标签} \quad = -\log p_{\theta}(y^*|x) \quad \text{负对数似然}$$

参数学习

► 模型： Softmax 回归

► 学习准则： 交叉熵

$$\mathcal{R}(W) = -\frac{1}{N} \sum_{n=1}^N (\mathbf{y}^{(n)})^T \log \hat{\mathbf{y}}^{(n)}$$

► 优化： 梯度下降

$$\frac{\partial \mathcal{R}(W)}{\partial W} = -\frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)} \left(\mathbf{y}^{(n)} - \hat{\mathbf{y}}^{(n)} \right)^T$$

交叉熵损失函数

► 负对数似然损失函数

$$\mathcal{L}(\mathbf{y}, f(\mathbf{x}, \theta)) = - \sum_{c=1}^C y_c \log f_c(\mathbf{x}, \theta)$$

► 对于一个三类分类问题，类别为 $[0,0,1]$ ，预测类别概率为 $[0.3,0.3,0.4]$ ，则

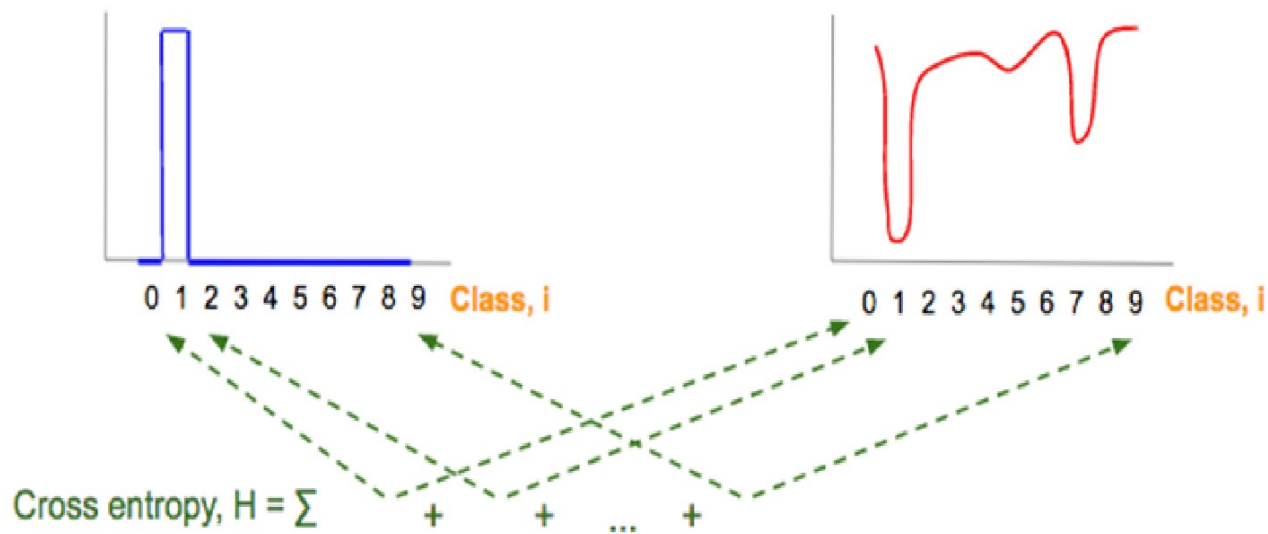
$$\begin{aligned}\mathcal{L}(\theta) &= -(0 \times \log(0.3) + 0 \times \log(0.3) + 1 \times \log(0.4)) \\ &= -\log(0.4).\end{aligned}$$

交叉熵损失

$$-\sum_{y=1}^c p_r(y|x) \log p_{\theta}(y|x)$$

真实概率 $p_r(y|x)$

预测概率的负对数 $-\log p_{\theta}(y|x)$





感知器

感知器

Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN¹

F. ROSENBLATT
Cornell Aeronautical Laboratory

HAVING told you about the giant digital computer known as I.B.M. 704 and how it has been taught to play a fairly creditable game of chess, we'd like to tell you about an even more remarkable machine, the perceptron, which, as its name implies, is capable of what amounts to original thought. The first perceptron has yet to be built,

The New Yorker, December 6, 1958 P. 44

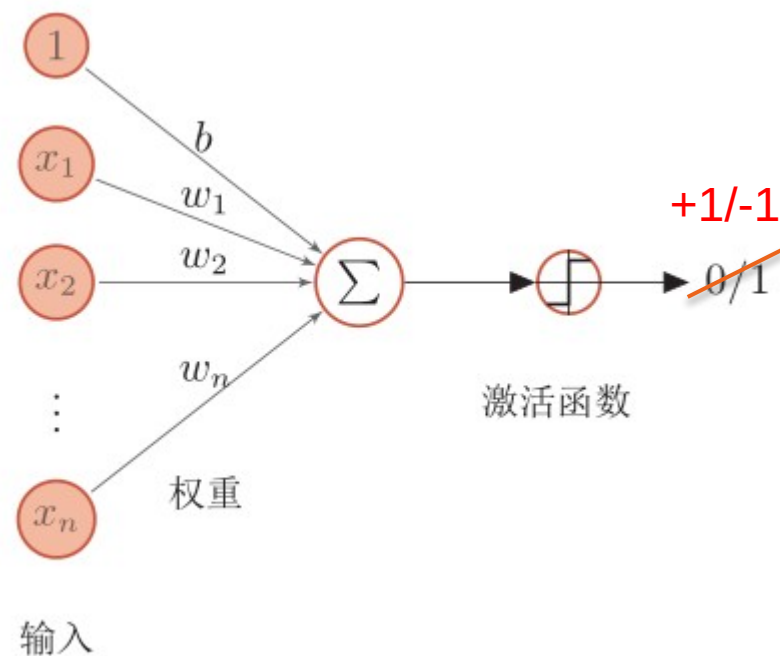


The IBM 704 computer

感知器

- ▶ 模拟生物神经元行为的机器，有与生物神经元相对应的部件，如权重（突触）、偏置（阈值）及激活函数（细胞体），输出为 +1 或 -1。

$$\hat{y} = \begin{cases} +1 & \text{当 } \mathbf{w}^T \mathbf{x} > 0 \\ -1 & \text{当 } \mathbf{w}^T \mathbf{x} \leq 0 \end{cases},$$



感知器

▶ 学习算法

- ▶ 一种错误驱动的在线学习算法：
- ▶ 先初始化一个权重向量 $\mathbf{w} \leftarrow \mathbf{0}$ （通常是全零向量）；
- ▶ 每次分错一个样本 $(\mathbf{x}, y)$ 时，即

$$y\mathbf{w}^T\mathbf{x} < 0$$

- ▶ 用这个样本来更新权重

$$\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$$

- ▶ 根据感知器的学习策略，可以反推出感知器的损失函数为

$$\mathcal{L}(\mathbf{w}; \mathbf{x}, y) = \max(0, -y\mathbf{w}^T\mathbf{x})$$

感知器的学习过程

算法 3.1 两类感知器的参数学习算法

输入: 训练集 $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$, 最大迭代次数 T

```
1 初始化:  $\mathbf{w}_0 \leftarrow 0, k \leftarrow 0, t \leftarrow 0$ ;  
2 repeat  
3   对训练集  $\mathcal{D}$  中的样本随机排序;  
4   for  $n = 1 \cdots N$  do  
5     选取一个样本  $(\mathbf{x}^{(n)}, y^{(n)})$ ;  
6     if  $\mathbf{w}_k^\top (y^{(n)} \mathbf{x}^{(n)}) \leq 0$  then  
7        $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y^{(n)} \mathbf{x}^{(n)}$ ;  
8        $k \leftarrow k + 1$ ;  
9     end  
10     $t \leftarrow t + 1$ ;  
11    if  $t = T$  then break;  
12  end  
13 until  $t = T$ ;  
    输出:  $\mathbf{w}_k$ 
```

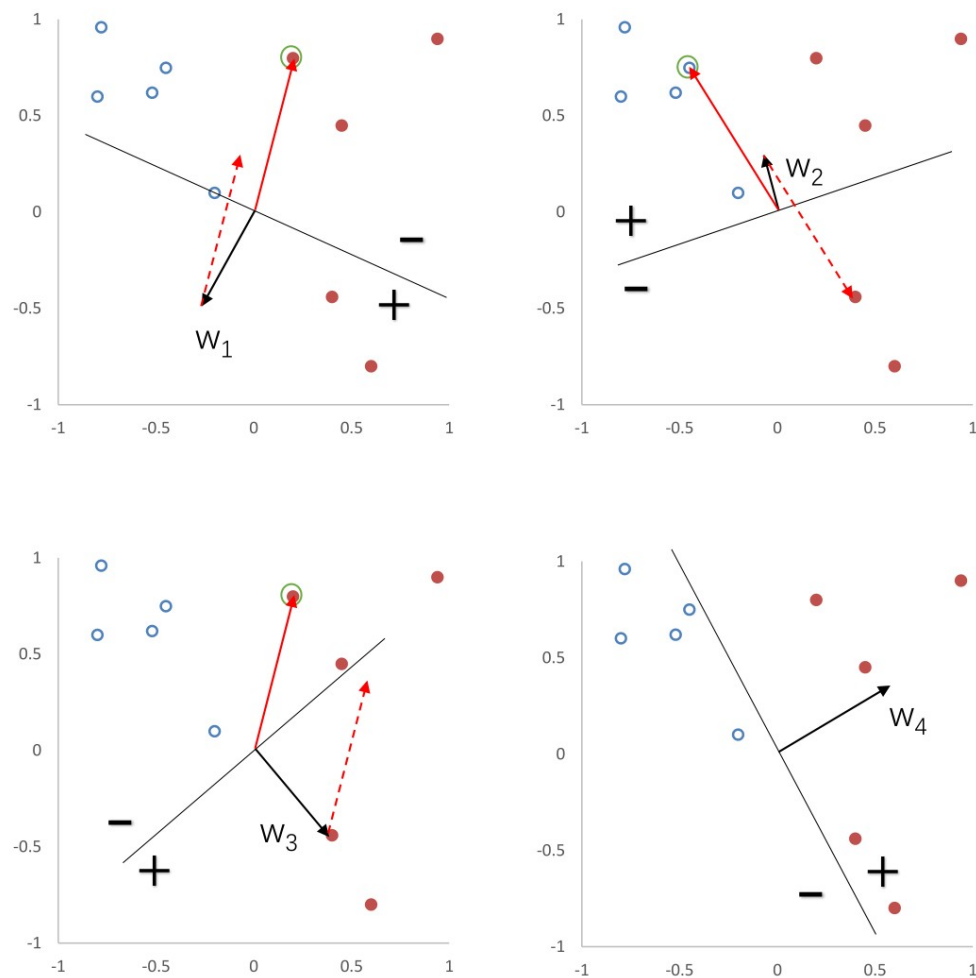
// 达到最大迭代次数

表示分错

对比 Logistic 回归的更新方式:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \alpha \frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)} (y^{(n)} - \hat{y}_{\mathbf{w}_t}^{(n)})$$

感知器参数学习的更新过程



收敛性

当数据集是两类线性可分时,对于训练集 $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$, 其中 $\mathbf{x}^{(n)}$ 为样本的增广特征向量, $y^{(n)} \in \{-1, 1\}$, 那么存在一个正的常数 $\gamma (\gamma > 0)$ 和权重向量 $\mathbf{w}^*$, 并且 $\|\mathbf{w}^*\| = 1$, 对所有 n 都满足 $(\mathbf{w}^*)^\top (y^{(n)} \mathbf{x}^{(n)}) \geq \gamma$. 我们可以证明如下定理.

定理 3.1 – 感知器收敛性: 给定训练集 $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$, 令 R 是训练集中最大的特征向量的模, 即

$$R = \max_n \|\mathbf{x}^{(n)}\|.$$

如果训练集 $\mathcal{D}$ 线性可分, 两类感知器的参数学习算法 3.1 的权重更新次数不超过 $\frac{R^2}{\gamma^2}$.

证明过程

- ▶ 感知器的权重向量的更新方式为

$$\mathbf{w}_k = \mathbf{w}_{k-1} + y^{(k)} \mathbf{x}^{(k)}$$

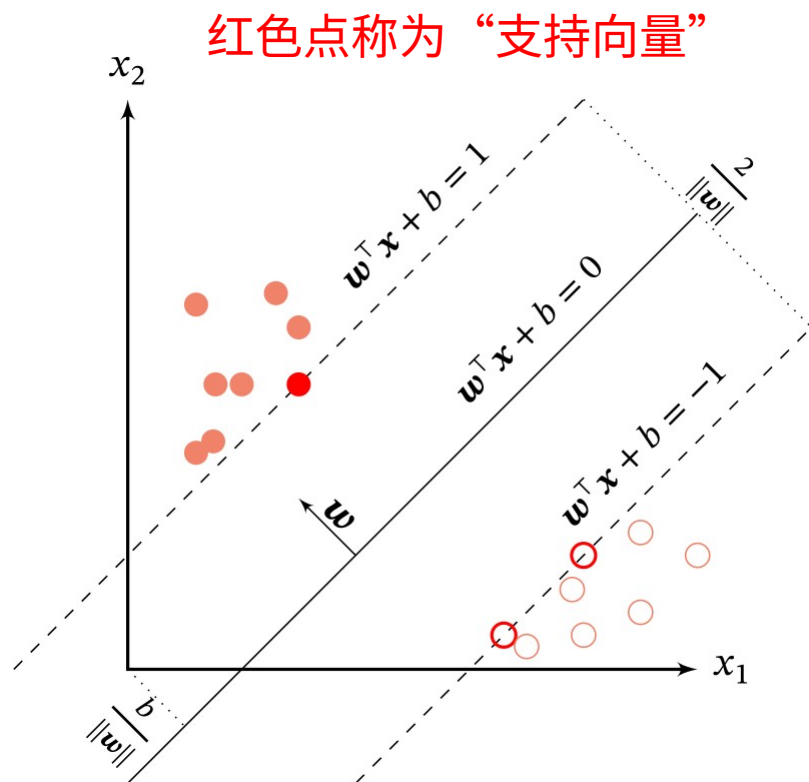
- ▶ 在第 K 次更新时感知器的权重向量为

$$\mathbf{w}_K = \sum_{k=1}^K y^{(k)} \mathbf{x}^{(k)}$$

- ▶ $\|\mathbf{w}_K\|^2$ 的上界

- ▶ $\|\mathbf{w}_K\|^2$ 的下界

支持向量机（Support Vector Machine，SVM）



支持向量机的目标是寻找一个超平面 $(\mathbf{w}^*, b^*)$ 使得 γ 最大, 即

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \gamma \\ \text{s.t.} \quad & \frac{y^{(n)}(\mathbf{w}^\top \mathbf{x}^{(n)} + b)}{\|\mathbf{w}\|} \geq \gamma, \forall n \in \{1, \dots, N\} \end{aligned}$$



$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \frac{1}{\|\mathbf{w}\|^2} \\ \text{s.t.} \quad & y^{(n)}(\mathbf{w}^\top \mathbf{x}^{(n)} + b) \geq 1, \forall n \in \{1, \dots, N\} \end{aligned}$$

数据集 $\mathcal{D}$ 中每个样本 $\mathbf{x}^{(n)}$ 到分割超平面的距离为:

$$\gamma^{(n)} = \frac{\|\mathbf{w}^\top \mathbf{x}^{(n)} + b\|}{\|\mathbf{w}\|} = \frac{y^{(n)}(\mathbf{w}^\top \mathbf{x}^{(n)} + b)}{\|\mathbf{w}\|}.$$



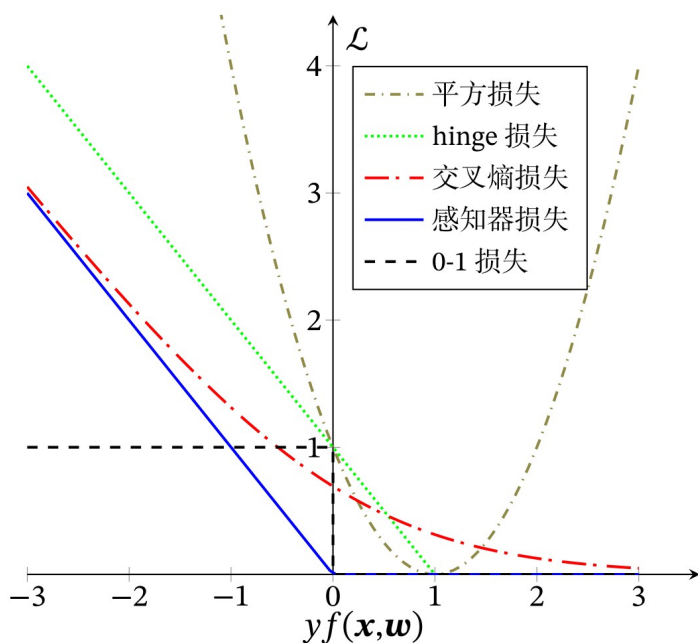
总结

线性分类模型小结

线性模型	激活函数	损失函数	优化方法
线性回归	-	$(y - \mathbf{w}^\top \mathbf{x})^2$	最小二乘、梯度下降
Logistic 回归	$\sigma(\mathbf{w}^\top \mathbf{x})$	$\mathbf{y} \log \sigma(\mathbf{w}^\top \mathbf{x})$	梯度下降
Softmax 回归	$\text{softmax}(\mathbf{W}^\top \mathbf{x})$	$\mathbf{y} \log \text{softmax}(\mathbf{W}^\top \mathbf{x})$	梯度下降
感知器	$\text{sgn}(\mathbf{w}^\top \mathbf{x})$	$\max(0, -y\mathbf{w}^\top \mathbf{x})$	随机梯度下降
支持向量机	$\text{sgn}(\mathbf{w}^\top \mathbf{x})$	$\max(0, 1 - y\mathbf{w}^\top \mathbf{x})$	二次规划、SMO 等

不同损失函数的对比

为了比较这些损失函数，我们统一定义类别标签 $y \in \{+1, -1\}$ ，并定义 $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x} + b$ 。这样对于样本 $(\mathbf{x}, y)$ ，若 $yf(\mathbf{x}; \mathbf{w}) > 0$ ，则分类正确，相反则分类错误。这样为了方便比较这些模型，我们可以将它们的损失函数都表述为定义在 $yf(\mathbf{x}; \mathbf{w})$ 上的函数。



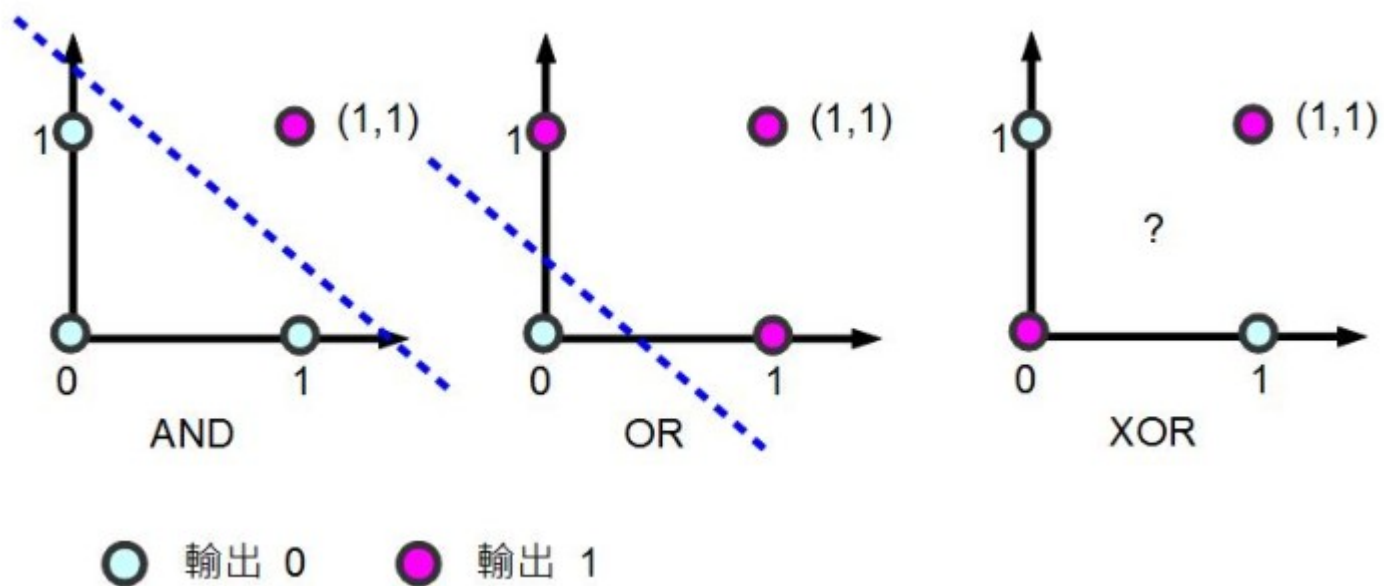
$$\mathcal{L}_{LR} = \log(1 + \exp(-yf(\mathbf{x}; \mathbf{w})))$$

$$\mathcal{L}_{hinge} = \max(0, 1 - yf(\mathbf{x}; \mathbf{w}))$$

$$\mathcal{L}_p = \max(0, -yf(\mathbf{x}; \mathbf{w}))$$

$$\mathcal{L}_{squared} = (1 - yf(\mathbf{x}; \mathbf{w}))^2$$

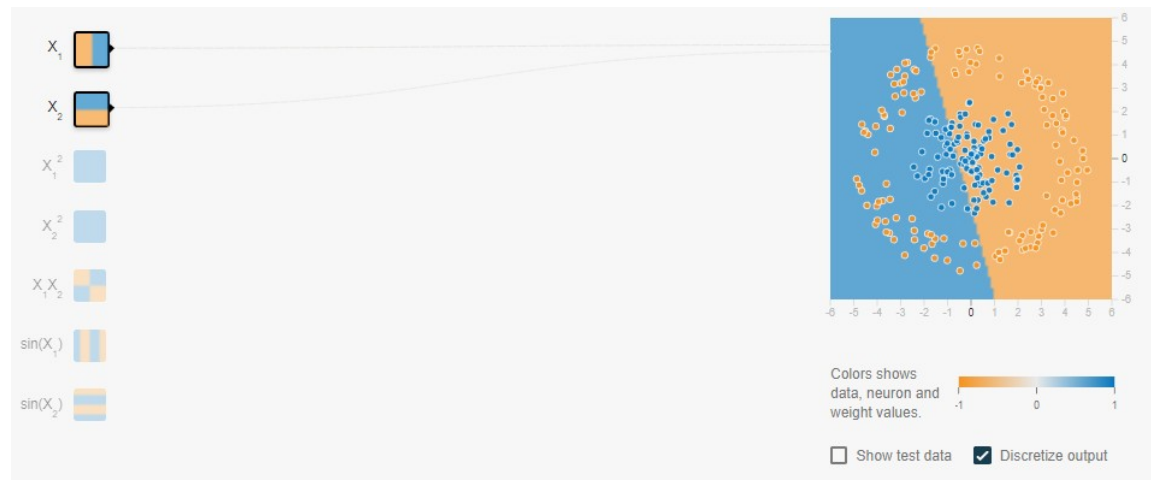
XOR 问题



特征

<http://playground.tensorflow.org>

如何处理非线性可分问题？



课后作业

▶ 掌握知识点

- ▶ Logistic 回归
- ▶ Softmax 回归
- ▶ 损失函数

▶ 编程练习

- ▶ chap3 softmax regression

谢 谢

<https://nndl.github.io/>