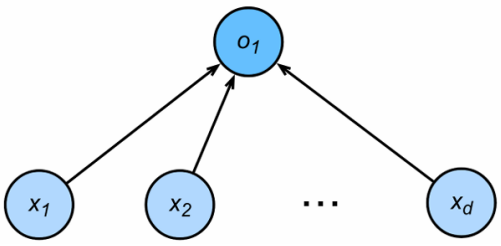
感知机

- 课件：part-0_12.pdf
- 60年代发明的，叫“感知”实际和神经网络相关
- 给定输入**x**，权重**w**，和偏移b，感知机输出

$$o = \sigma\left(\langle \mathbf{w}, \mathbf{x} \rangle + b\right) \quad \sigma(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$



理解为线性回归模型的输出运用σ函数，本质是二分类

- 注意此处σ函数的值也可以不是0和1，如设为1和-1，即二分类
  - 联系回归模型输出**实数**

    感知机的输出是离散的类而不是实数
  - Softmax回归输出概率

- **训练感知机**
  - 与以往不同，w的初始值不是随机值而是确定的0

$$\begin{aligned} &\textbf{initialize } w = 0 \text{ and } b = 0 \\ &\textbf{repeat} \\ &\quad \textbf{if } y_i \left[\langle w, x_i \rangle + b\right] \leq 0 \textbf{ then} \\ &\quad\quad w \leftarrow w + y_i x_i \text{ and } b \leftarrow b + y_i \\ &\quad \textbf{end if} \\ &\textbf{until all classified correctly} \end{aligned}$$

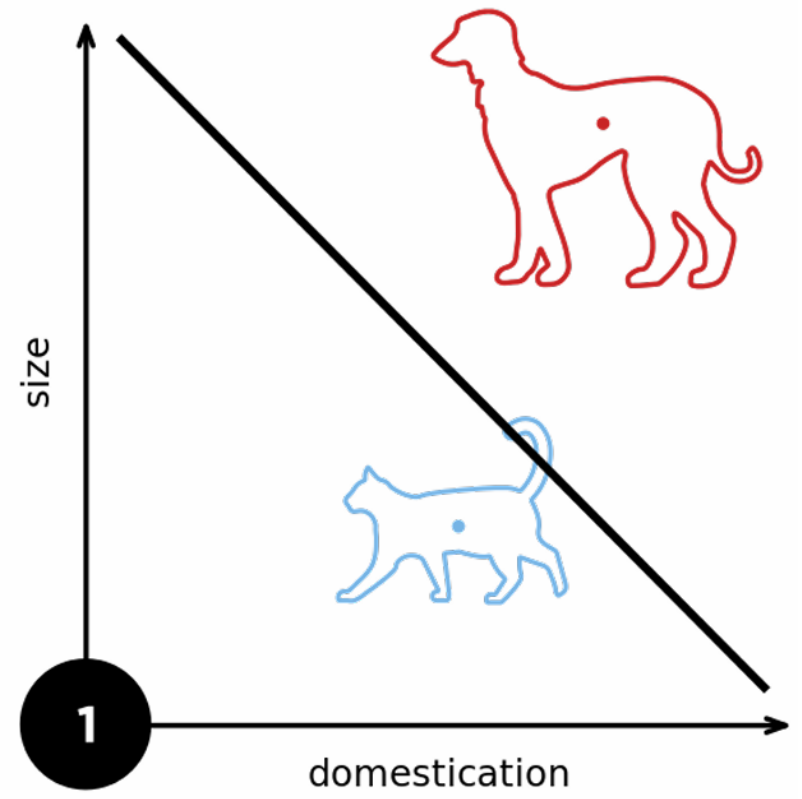
由于输出只能是正类或负类，yi·[<w,xi>+b]<=0意味着真实值和预测值符号不同，即预测错误。此时选择更新权重和偏置，直到所有样本均被正确预测

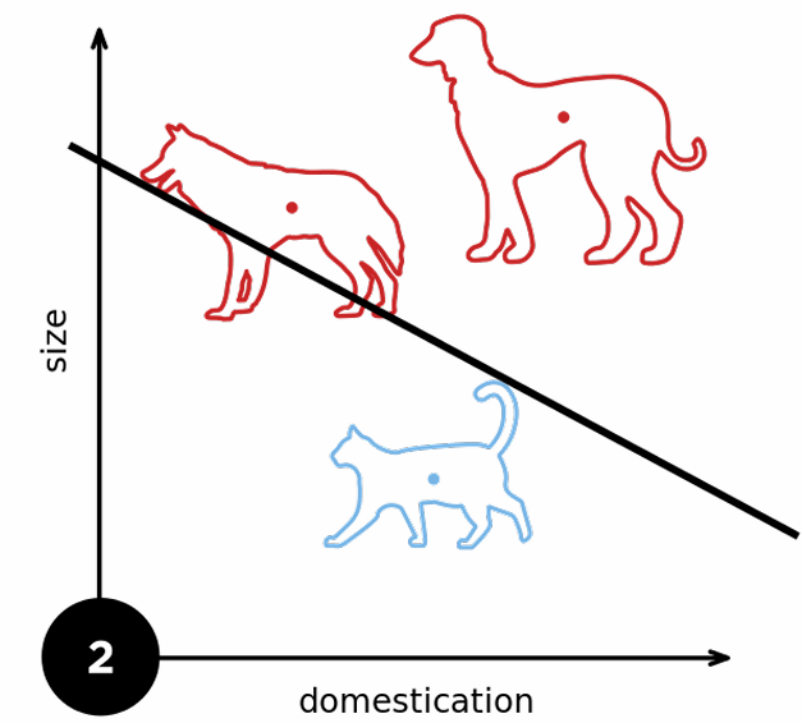
  - 等价于使用批量大小为1的梯度下降，即每一次只用一个样本做梯度更新，注意不是随机梯度（随机梯度下降（stochastic gradient descent，SGD） - 知乎）下降，这和感知机的本质（【机器学习】感知机原理详解_感知机算法原理-CSDN博客）有关，它本身就是一直迭代固定值直到分类正确。并使用如下的损失函数：

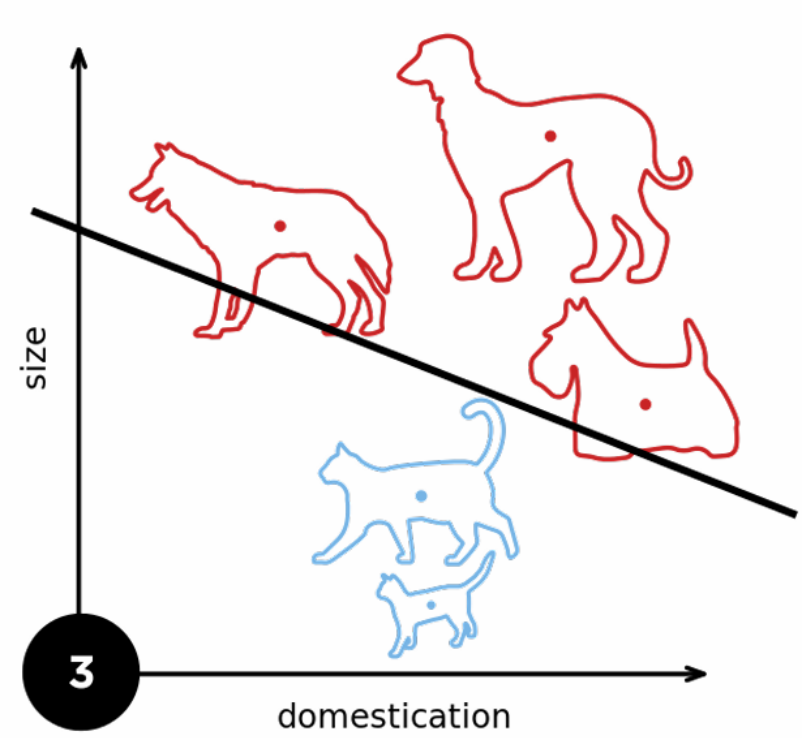
$$\ell(y, \mathbf{x}, \mathbf{w}) = \max(0, -y\langle \mathbf{w}, \mathbf{x} \rangle)$$

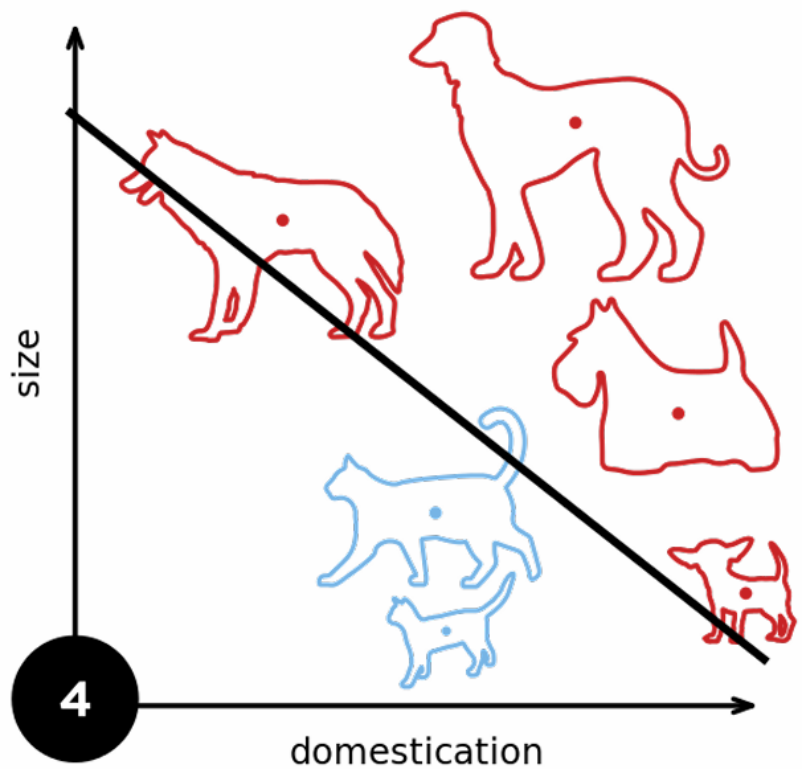
这个损失函数对应上图中的if语句，y<w,x>是y与<w,x>的乘积

- 用图像表示，每一次更新权重和偏置都是对分隔线的斜率和截距进行更新。迭代的停止条件是能够对所有的类都分类正确。









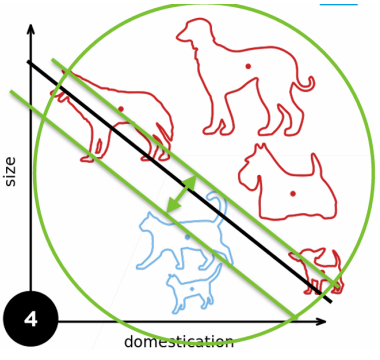
- **收敛定理**

  - 数据在半径 r 内
  - 余量 ρ 分类两类

  $$y(\mathbf{x}^{\top}\mathbf{w}+b)\geq\rho$$

  对于 $\|\mathbf{w}\|^2+b^2\leq 1$

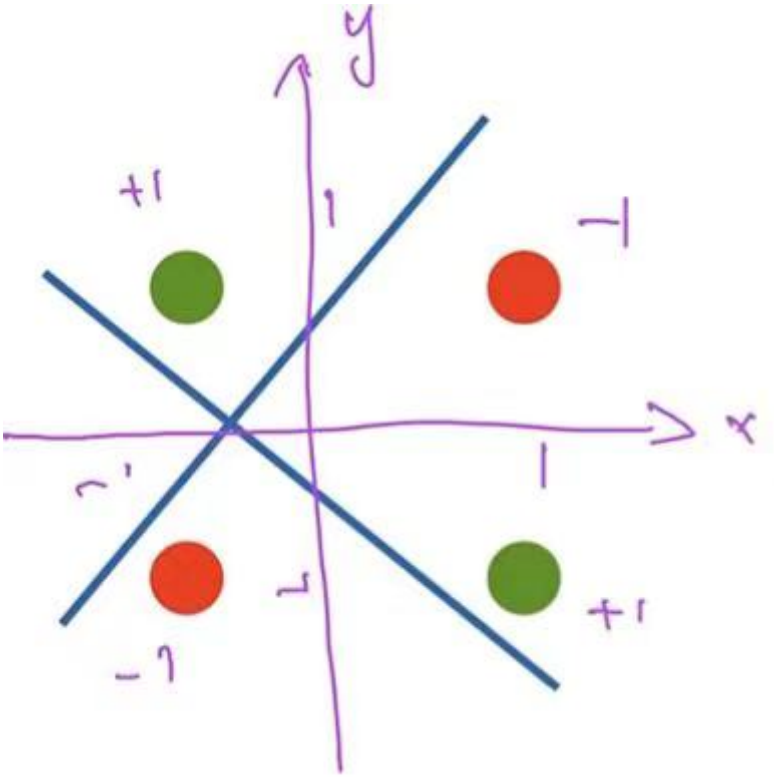
  - 感知机保证在 $\frac{r^2+1}{\rho^2}$ 步后收敛



  其证明不要求掌握

- 对所有数据都分类正确即y·(<**x**,**w**>+b)>0且有一定余量(ρ>0)。在图中的意义是分类线有移动的空间。
- r反映的数据分布区域的大小
- 最后一点的理解：数据分布区域越大，需要走的步数越多；数据够好就能分得很开，ρ更大，也能少走几步更快收敛。

- **XOR问题**
  - 感知机不能拟合XOR函数，只能产生线性分割面



由Minsky和Papert在1969年提出，导致AI的第一个寒冬
多层感知机解决了这个问题

- **总结**
  - 感知机是一个二分类模型，是最早的AI模型之一

    输出1|0,-1|1
  - 其求解算法等价于使用批量大小为1的梯度下降（不是随机梯度下降）
  - 它不能拟合XOR函数