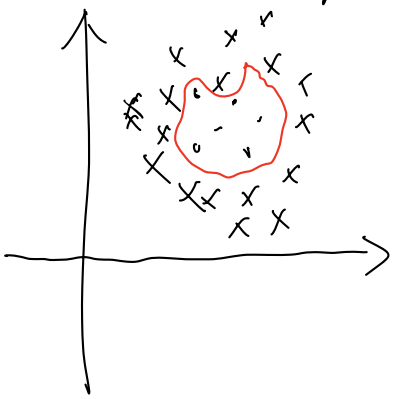
① Lecture2中的线性变化可以解决部分问题，但有很大缺陷。

物理世界很多都是线性问题，比如平或运算、分类



红色曲线划分两类数据，但曲线本身是非线性

② 引入非线性激活函数 $h_\theta(x)=\theta^T\phi(x)$，$\theta\in R^{d\times k}$，$\phi: R^n\to R^d$。

ϕ 用于把 x 组合成非线性的隐藏的特征

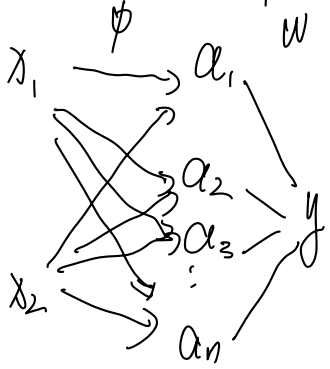
比如上面的曲线分类，$\phi(x_1,x_2)=(x_1-3)^2+(x_2-4)^2$，$\phi: R^2\to R^1$

问题变成寻找合适的 ϕ，映射出隐藏特征。这是古典机器学习：

人工识别数据特征，设计 ϕ，解决分类问题

但是，这不具备通用性。场景之间 ϕ 差异是巨大的

③ 无法提前选择合适的 ϕ，那么只好全上了！



$a_1=x_1^2$,
$a_2=x_2^2$
$a_3=\sin x_1$
$a_4=\cos x_1$
$a_5=$ 其它非线性函数

预先准备常用的非线性函数，由后面的 W 来线性拟合，通过学习 W 来选择不同的权重

④ 有了 feature mapping，后面的层怎么办？

$$\begin{cases} z_1 \xrightarrow{W_1} z_2 \xrightarrow{W_2} \cdots \xrightarrow{W_L} z_{L+1} \\ z_{i+1} = G_i(z_i W_i), \longrightarrow \begin{cases} i = 1, 2, \cdots, L \\ z_i = X \\ h_\theta(x) = z_{L+1} \\ z_i \in R^{m \times n_i}, W_i \in R^{n_i \times n_{i+1}} \end{cases} \end{cases}$$

W 即是网络参数 θ

G 一般是 ReLU 激活函数

⑤ 如何求 W_i？

$$\begin{cases} Loss = \frac{1}{m} \sum_{i=1}^{m} Loss(h_\theta(x^i), y^i) \\ \frac{\partial Loss}{\partial W}, \text{又成为求偏导的问题} \end{cases}$$

~~基于 G_i 是 softmax 等~~ 可以推导出 $\frac{\partial Loss}{\partial W_i} = \frac{\partial Loss}{\partial z_{L+1}} \cdot \frac{\partial z_{L+1}}{\partial z_L} \cdots \frac{\partial z_{i+1}}{\partial W_i}$

此时我们看到 **反向传播**，

记 $G_{i+1} = \frac{\partial Loss}{\partial z_{L+1}} \cdot \frac{\partial z_{L+1}}{\partial z_L} \cdots \frac{\partial z_{i+2}}{\partial z_{i+1}}$，那么 $\frac{\partial Loss}{\partial W_i} = G_{i+1} \cdot \frac{\partial z_{i+1}}{\partial W_i}$

$G_{L+1} = \frac{\partial Loss}{\partial z_{L+1}}$，而 $z_{L+1} \to softmax \to Loss$，

$\therefore G_{L+1} = S - I_y$

$G_i = G_{i+1} \cdot \frac{\partial z_{i+1}}{\partial z_i} = G_{i+1} \cdot \frac{\partial G(z_i W_i)}{\partial z_i} = G_{i+1} \cdot G'(z_i W_i) \cdot W_i$

$\frac{\partial Loss}{\partial W_i} = G_{i+1} \cdot \frac{\partial z_{i+1}}{\partial W_i}$，而 G_{i+1} 由 G_{i+2} 得出，即是反向传播！

⑥ 虽然有了反向传播的理念，我们还是要基于一个easy例子来理解它

$$\begin{cases} h_\theta(x) = \sigma(X W_1) W_2 \\ loss = \frac{1}{m}\sum_{i=1}^{m} (Loss\ c h_\theta(x_i, y_i)) \end{cases}$$

$$X_{m\times n} \xrightarrow{W_1} \xrightarrow{\sigma} H_1 \xrightarrow[L\times k]{W_2} H_2 \xrightarrow{softmax} Z_{m\times k}$$

$$Loss = \frac{1}{m}\sum_{i=1} \log(Z_{iy_i}),\quad Z_{iy_i} = \frac{e^{h_{iy_i}}}{e^{h_{i1}} + e^{h_{i2}} + \cdots + e^{h_{ik}}}$$

⑦ $\frac{dloss}{dW_2} = \frac{dLoss}{dH_2} \cdot \frac{dH_2}{dW_2}$

1) $\frac{dLoss}{dH_{ij}} = \frac{\frac{1}{m} d\left(\sum_{a=1}^{m} \log \frac{e^{ha y_a}}{e^{ha1} + e^{ha2} + \cdots + e^{hak}}\right)}{dh_{ij}} = \frac{1}{m} \cdot \frac{d \log \frac{e^{hiy_i}}{e^{hi1} + e^{hi2} + \cdots + e^{hik}}}{dh_{ij}}$

$$= \frac{1}{m} \frac{d\left(h_{iy_i} - \log\left(e^{hi1} + e^{hi2} + \cdots + e^{hik}\right)\right)}{dh_{ij}}$$

$$\begin{cases} \frac{1}{m}\left(1 - \frac{e^{hij}}{e^{hi1} + e^{hi2} + \cdots + e^{hik}}\right), & j = y_i \\ \frac{1}{m}\left(0 - \frac{e^{hij}}{e^{hi1} + e^{hi2} + \cdots + e^{hik}}\right), & j \neq y_i \end{cases}$$

即是 $\frac{dLoss}{dH_{ij}} = \frac{1}{m}(I - Z_{ij})$

推广到矩阵 $\frac{dLoss}{dH_2} = \frac{1}{m}(I - Z)$

2) $\frac{dH_{ij}}{dW_2} =$ $H_{ij} = H_{i1}W_{1j} + H_{i2}W_{2j} + \cdots + H_{iL} \cdot W_{Lj}$

H_{ij} 只与 $W_{1\sim L,j}$ 有关，$\frac{dH_{ij}}{dW_{1j}} = H_{i1}$，$\frac{dH_{ij}}{dW_{2j}} = H_{i2}$ ⋯ $\frac{dH_{ij}}{dW_{Lj}} = H_{iL}$

$\frac{dH_{ij}}{dW_{aj}} = H_{ia}$，$a \in \{1, 2, \cdots, L\}$，即是 $\frac{dH_2}{dW_2} = H_1$

那么 $\frac{dLoss}{dW_2} = \frac{1}{m}(I-Z)\cdot H_1$，整理为 $\frac{dLoss}{dW_2} = \frac{1}{m}\cdot H_1^T(I-Z)$

Lxk　　mxk　　mxL

⑧ 继续求 $\frac{dLoss}{dW_1}$　　$\frac{dLoss}{dW_1} = \frac{dLoss}{dH_2}\cdot\frac{dH_2}{dH_1}\cdot\frac{dH_1}{dW_1}$

1) $\frac{dLoss}{dH_2} = \frac{1}{m}(I-Z)$，⑦已经有证明

2) $\frac{dH_2}{dH_1} = W_2$　　$(H_2 = H_1W_2)$

mxk　mxL Lxk

3) $\frac{dH_1}{dW_1} = \frac{d\ G(XW_1)}{dW_1} = \frac{d\ G(XW_1)}{d(XW_1)}\cdot\frac{d(XW_1)}{dW_1} = G'(XW_1)\cdot X$

那么 $\frac{dLoss}{dW_1} = \frac{1}{m}(I-Z)\cdot W_2\cdot G'(XW_1)\cdot X$

nxL　　mxk　　Lxk　　mxL　　mxn

整理得 $\frac{dLoss}{dW_1} = \frac{1}{m}X^T\times\left((I-Z)\times W_2^T\circ G'(XW_1)\right)$

nxL　　nxm　　mxk　　kxL　　mxL

矩阵乘积　　标量积

⑨ 那么 W_1/W_2 都有了 gradient descent

$$\begin{cases}\frac{dLoss}{dW_1} = \frac{1}{m}X^T\left((I-Z)W_2^T\circ G'(XW_1)\right)\\ \frac{dLoss}{dW_2} = \frac{1}{m}H_1^T(I-Z)\end{cases}$$