5.1

如果使用$y=w^Tx$作为激活函数，单元值仍然是输入值的线性组合，类似于线性回归，达不到神经网络激活与筛选的目的

▲ 讨论$\frac{e^{x_i}}{\sum_{j=1}^{C}e^{x_j}}$和$log\sum_{j=1}^{C}e^{x_j}$的数值溢出问题

对于

$$\frac{e^{x_i}}{\sum_{j=1}^{C}e^{x_j}}$$

如果x_i均取相同的数，则上式的值应该为$\frac{1}{C}$

但是如果x_i的值比较极端，则可能出现计算时溢出的现象，这时对上式取对数

$$log(\frac{e^{x_i}}{\sum_{j=1}^{C}e^{x_j}})=x_i-log(\sum_{j=1}^{C}e^{x_j})$$

给x_j减去一个常数，处理溢出的问题，则上式等于

$$x_i-log(\sum_{j=1}^{C}e^{x_j-Const})-Const$$

一般的，我们可以取

$$Const=\max_i\{x_i\}$$

可以发现，$log\sum_{j=1}^{C}e^{x_j}$其实是上述讨论式子中的一个部分，所以溢出情况和上述类似

▲ 计算$\frac{e^{x_i}}{\sum_{j=1}^{C}e^{x_j}}$和$log\sum_{j=1}^{C}e^{x_j}$关于向量$x=[x_1,\cdots,x_c]$的梯度

先计算

$$\nabla\frac{e^{x_i}}{\sum_{j=1}^{C}e^{x_j}}$$

先求

$$\frac{\partial}{\partial x_m}\frac{e^{x_i}}{\sum_{j=1}^{C}e^{x_j}}=-\frac{e^{x_i+x_m}}{(\sum_{j=1}^{C}e^{x_j})^2}$$

再求当$i=m$时

$$\begin{aligned}\frac{\partial}{\partial x_i}\frac{e^{x_i}}{\sum_{j=1}^{C}e^{x_j}}&=\frac{e^{x_i}(\sum_{j=1}^{c}e^{x_j})-e^{2x_i}}{(\sum_{j=1}^{C}e^{x_j})^2}\\&=\frac{e^{x_i}(\sum_{j=1}^{C}e^{x_j}-e^{x_i})}{(\sum_{j=1}^{C}e^{x_j})^2}\end{aligned}$$

则

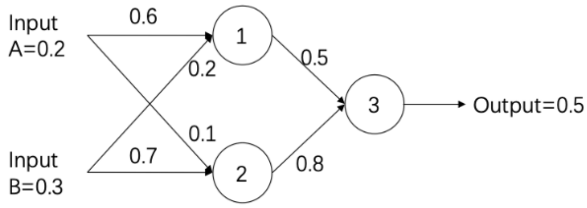
$$\nabla\frac{e^{x_i}}{\sum_{j=1}^{C}e^{x_j}}=[-\frac{e^{x_i+x_1}}{(\sum_{j=1}^{C}e^{x_j})^2},\cdots,\frac{e^{x_i}(\sum_{j=1}^{C}e^{x_j}-e^{x_i})}{(\sum_{j=1}^{C}e^{x_j})^2},\cdots,-\frac{e^{x_i+x_C}}{(\sum_{j=1}^{C}e^{x_j})^2}]$$

类似可以得到

$$\nabla log\sum_{j=1}^{C}e^{x_j}=[\frac{e^{x_1}}{\sum_{j=1}^{C}e^{x_j}},\cdots,\frac{e^{x_C}}{\sum_{j=1}^{C}e^{x_j}}]$$

▲

考虑如下简单网络，假设激活函数为ReLU，用平方损失$\frac{1}{2}(y-\hat{y})^2$计算误差，请用BP算法更新一次所有参数（学习率为1），给出更新后的参数值（给出详细计算过程），并计算给定输入值x=(0.2,0.3)时初始时和更新后的输出值，检查参数更新是否降低了平方损失值.



$$b_0=\begin{bmatrix}0.18\\10.23\end{bmatrix}$$
$$\hat{y}=0.274$$
$$\frac{1}{2}(y-\hat{y})^2=0.025538$$
$$V_0=\begin{bmatrix}0.6 & 0.1\\0.2 & 0.7\end{bmatrix}\qquad w_0=\begin{bmatrix}0.5\\0.8\end{bmatrix}$$

参数更新:

$$V_1=V_0-(\hat{y}-y)w_0^Tx=\begin{bmatrix}0.6226 & 0.1339\\0.2362 & 0.7542\end{bmatrix}$$
$$w_1=w_0-(\hat{y}-y)b_0=V_0=\begin{bmatrix}0.5407\\0.8520\end{bmatrix}$$

更新后输出和平方损失

$$b_1=\begin{bmatrix}0.1954\\0.2531\end{bmatrix}$$
$$y_1=0.321227$$
$$\frac{1}{2}(y-\hat{y})^2=0.0160$$

显然，平方损失值下降了

6.4

二分类问题且数据是线性可分的

6.6

线性超平面是由少数支持向量决定的，噪声很有可能成为某个支持向量，则会对整个分类产生很大影响

6.9

对于 w^Tx+b 形式的模型，即线性模型，使用核技巧的关键点在于最优的 w^* 可以由训练集的线性组合表示， 即 $w^*=\sum_i\beta_ix_i$ ，使得模型可表示为 $\sum_i\beta_i\langle x_i,x\rangle+b$ ， 进而使用核函数直接计算数据点在高维空间内积，而不显式的计算数据点从低维到高维的映射。

原命题：事实上对于任何 L2 正则化的线性模型： $\min_w \frac{\lambda}{N}w^Tw+\frac{1}{N}\sum_{n=1}^{N}err(y_n,w^Tx_n)$ ， 这里，其最优值都可以表示为 $w^*=\sum_i\beta_ix_i$

▲

支持向量回归的对偶问题如下，

$$\max_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}} g(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}})=-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)+\sum_{i=1}^{m}(y_i(\hat{\alpha}_i-\alpha_i)-\epsilon(\hat{\alpha}_i+\alpha_i))$$

$$s.t.\ C\geqslant\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}\geqslant 0 \text{ and } \sum_{i=1}^{m}(\alpha_i-\hat{\alpha}_i)=0$$

请将该问题转化为类似于如下标准型的形式（$\boldsymbol{u},\boldsymbol{v},\boldsymbol{K}$均已知），

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha})=\boldsymbol{\alpha}^\top\boldsymbol{v}-\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha}$$

$$s.t.\ C\geqslant\boldsymbol{\alpha}\geqslant 0 \text{ and } \boldsymbol{\alpha}^\top\boldsymbol{u}=0$$

例如在软间隔SVM中$\boldsymbol{v}=\boldsymbol{1},\boldsymbol{u}=\boldsymbol{y},\boldsymbol{K}[i,j]=y_iy_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$.

若$\kappa(x_i,x_j)=\phi_(x_i)^T\phi_(x_j)=(x_i^Tx_j)^2$，求$\phi(x_i)$表达式。

《机器学习概论》 2022/10/17

取

$$\alpha'=\begin{bmatrix}\alpha\\ \hat{\alpha}\end{bmatrix}$$

$$v=\begin{bmatrix}-y-\epsilon 1_m\\ y-\epsilon 1_m\end{bmatrix}$$

$$u=\begin{bmatrix}1_m\\ -1_m\end{bmatrix}$$

$$K'=\begin{bmatrix}I_m\\ -I\end{bmatrix}K[I \quad -I]$$

$$K[i,j]=\kappa(x_i,x_j)$$