HW5 and HW6 Reference

1 HW5

1.1 试述将线性函数$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$用作神经元激活函数的缺陷

（言之有理即可）

解：当单元层和隐藏层激活函数为线性函数$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$时，每一层输出都是上层输入的线性函数，无论神经网络有多少层，输出都是输入的线性组合，神经网络实际仍为原始的感知机；当输出层激活函数也为线性函数时，相当于整体的线性回归。此时的网络无法处理非线性问题。使用非线性激活函数增加了神经网络模型的非线性因素，使得神经网络可以任意逼近任何非线性函数，这样神经网络就可以应用到众多的非线性模型中。

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

解：实数在计算机中以二进制表示，计算时非精确值。当数值过小会取0（下溢出）或数值过大导致上溢出。对于softmax函数，当$x_i \to -\infty, i = 1, 2, \cdots, C$时，$\sum_{i=1}^{C} e^{x_i} \to 0$，分母计算可能四舍五入为0，发生下溢出。当$x_i \to +\infty$时，$e^{x_i} \to +\infty$，分子计算可能出现上溢出。

解决方法：

对于softmax函数，令$M = max(x_i), i = 1, 2, \cdots, C$，将计算$f(x_i)$改为计算$f(x_i - M)$即可。

说明如下：

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

$e^{x_i - M} \leq e^{M-M} = 1$，分子不会发生上溢；

$\sum_{j=1}^{C} e^{x_j - M} \geq e^{M-M} = 1$，分母不会发生下溢。

对于log softmax函数，同上，

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

$1 = e^{M-M} \leq \sum_{j=1}^{C} e^{x_j - M} \leq \sum_{j=1}^{C} e^{M-M} = C$，所以分子不会发生上溢,分母不会发生下溢。

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

（计算对向量梯度需分别计算对每个分量的梯度，计算结果为向量形式）

解：令$f(x_i) = \dfrac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}, g(x_i) = logf(x_i), i, k = 1, 2, \cdots, C$

$k \neq i$时，

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

$k = i$时，

$$\frac{\partial f(x_i)}{\partial x_k} = \frac{e^{x_i}\sum_{m=1,m\neq i}^{C} e^{x_m}}{(\sum_{j=1}^{C} e^{x_j})^2}$$

所以

$$\begin{aligned}\frac{\partial f(x_i)}{\partial \boldsymbol{x}} &= \frac{e^{x_i}}{(\sum_{j=1}^{C} e^{x_j})^2}[-e^{x_1},\cdots,-e^{x_{i-1}},\sum_{m=1,m\neq i}^{C} e^{x_m},\cdots,-e^{x_C}] \\ &= \frac{f(x_i)}{\sum_{j=1}^{C} e^{x_j}}[-e^{x_1},\cdots,-e^{x_{i-1}},\sum_{m=1,m\neq i}^{C} e^{x_m},\cdots,-e^{x_C}]\end{aligned}$$

同理可得，

$$\begin{aligned}\frac{\partial g(x_i)}{\partial \boldsymbol{x}} &= \frac{1}{\sum_{j=1}^{C} e^{x_j}}[-e^{x_1},\cdots,-e^{x_{i-1}},\sum_{m=1,m\neq i}^{C} e^{x_m},\cdots,-e^{x_C}] \\ &= \frac{f(x_i)}{e^{x_i}}[-e^{x_1},\cdots,-e^{x_{i-1}},\sum_{m=1,m\neq i}^{C} e^{x_m},\cdots,-e^{x_C}]\end{aligned}$$

1.4 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$\boldsymbol{x}=(0.2,0.3)$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.

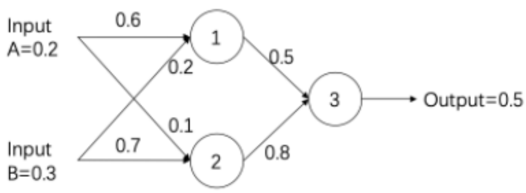

Figure 1: 简单网络

（注意题目中BP算法使用的激活函数为ReLU函数以及链式求导计算梯度）

解：

$$ReLU'(x) = \begin{cases} 1 & x > 0 \\ 0 & x < 0 \end{cases} \tag{1}$$

$v_{11}=0.6, v_{12}=0.1, v_{21}=0.2, v_{22}=0.7, w_1=0.5, w_2=0.8, \alpha_1,\alpha_2,\gamma$为节点1,2,3的输入值，$\beta_1,\beta_2,\hat{\gamma}$为节点1,2,3的输出值

第一次正向传播：

$\alpha_1 = v_{11}x_1 + v_{21}x_2 = 0.6\times 0.2 + 0.2\times 0.3 = 0.18$

$\beta_1 = max(0,\alpha_1) = 0.18$

$\alpha_2 = v_{12}x_1 + v_{22}x_2 = 0.1\times 0.2 + 0.7\times 0.3 = 0.23$

$\beta_2 = max(0,\alpha_2) = 0.23$

$\gamma = w_1\beta_1 + w_2\beta_2 = 0.5 \times 0.18 + 0.8 \times 0.23 = 0.274$
$\hat{\gamma} = max(0, \gamma) = 0.274$
误差$E = \frac{1}{2}(\gamma - \hat{\gamma})^2 = 0.5 \times (0.5 - 0.274)^2 = 0.025538$
误差逆传播计算梯度：
$\frac{\partial E}{\partial v_{11}} = \frac{\partial E}{\partial \hat{\gamma}}\frac{\partial \hat{\gamma}}{\partial \gamma}\frac{\partial \gamma}{\partial \beta_1}\frac{\partial \beta_1}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial v_{11}} = (\hat{\gamma} - \gamma)ReLU'(\gamma)w_1ReLU'(\alpha_1)x_1 = -0.0226$
$\frac{\partial E}{\partial v_{12}} = \frac{\partial E}{\partial \hat{\gamma}}\frac{\partial \hat{\gamma}}{\partial \gamma}\frac{\partial \gamma}{\partial \beta_2}\frac{\partial \beta_2}{\partial \alpha_2}\frac{\partial \alpha_2}{\partial v_{12}} = (\hat{\gamma} - \gamma)ReLU'(\gamma)w_2ReLU'(\alpha_1)x_1 = -0.03616$
$\frac{\partial E}{\partial v_{21}} = \frac{\partial E}{\partial \hat{\gamma}}\frac{\partial \hat{\gamma}}{\partial \gamma}\frac{\partial \gamma}{\partial \beta_1}\frac{\partial \beta_1}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial v_{21}} = (\hat{\gamma} - \gamma)ReLU'(\gamma)w_1ReLU'(\alpha_1)x_2 = -0.0339$
$\frac{\partial E}{\partial v_{22}} = \frac{\partial E}{\partial \hat{\gamma}}\frac{\partial \hat{\gamma}}{\partial \gamma}\frac{\partial \gamma}{\partial \beta_2}\frac{\partial \beta_2}{\partial \alpha_2}\frac{\partial \alpha_2}{\partial v_{22}} = (\hat{\gamma} - \gamma)ReLU'(\gamma)w_2ReLU'(\alpha_1)x_2 = -0.05424$
$\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial \hat{\gamma}}\frac{\partial \hat{\gamma}}{\partial \gamma}\frac{\partial \gamma}{\partial \beta_1} = (\hat{\gamma} - \gamma)ReLU'(\gamma)\beta_1 = -0.04068$
$\frac{\partial E}{\partial w_2} = \frac{\partial E}{\partial \hat{\gamma}}\frac{\partial \hat{\gamma}}{\partial \gamma}\frac{\partial \gamma}{\partial \beta_2} = (\hat{\gamma} - \gamma)ReLU'(\gamma)\beta_1 = -0.05198$
学习率$\eta = 1$，更新权重
$v'_{11} = v_{11} - \eta\frac{\partial E}{\partial v_{11}} = 0.6226$
$v'_{12} = v_{12} - \eta\frac{\partial E}{\partial v_{12}} = 0.13616$
$v'_{21} = v_{21} - \eta\frac{\partial E}{\partial v_{21}} = 0.2339$
$v'_{22} = v_{22} - \eta\frac{\partial E}{\partial v_{22}} = 0.75424$
$w'_1 = w_1 - \eta\frac{\partial E}{\partial w_1} = 0.54068$
$w'_2 = w_2 - \eta\frac{\partial E}{\partial w_2} = 0.85198$
第二次正向传播更新输出使用新权重，计算过程与第一次一致，计算结果如下：
$\alpha'_1 = 0.19469, \alpha'_2 = 0.253504, \beta'_1 = 0.19469, \beta'_2 = 0.253504, \gamma' = 0.3212, \hat{\gamma}' = 0.32125, E' = 0.015976$
$0.015976 \leq 0.025538$，可知参数更新降低了平方损失。

2 HW6

2.1 试讨论线性判别分析与线性核支持向量机在何种条件下等价

（言之有理即可）

解：线性判别分析能够解决 n 分类问题，而线性核 SVM 只能解决二分类问题。当线性判别分析的投影向量和线性核SVM的超平面向量垂直的时候，SVM 的最大间隔就是线性判别分析所要求的异类投影点间距，同时在这种情况下，线性判别分析的同类样例的投影点也会被这个超平面所划分在一起，使其间隔较小。所以（1）线性判别分析求解出来的投影向量和线性核 SVM 求解出来的超平面向量垂直，（2）数据集只有两类，（3）数据集线性可分时，SVM 和 LDA等价。

2.2 试析SVM对噪声敏感的原因

（言之有理即可）

解：（1）SVM 的基本形态是一个硬间隔分类器，它要求所有样本都满足硬间隔约束，因此噪

声很容易影响 SVM 的学习。（2）存在噪声时，SVM 容易受噪声信息的影响，将训练得到的超平面向两个类间靠拢，导致训练的泛化能力降低，尤其是当噪声成为支持向量时，会直接影响整个超平面。（3）当 SVM 推广到到使用核函数时，会得到一个更复杂的模型，此时噪声也会一并被映射到更高维的特征，可能会对训练造成更意想不到的结果。综上， SVM 对噪声敏感。

2.3 试使用核技巧推广对率回归产生“核对率回归”

（使用对率回归函数推导或使用对率损失函数代替0/1损失函数推导也可）

解：对率回归的L2正则化目标函数

$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m}(-y_i\boldsymbol{\beta}^T\hat{x}_i + ln(1+e^{\boldsymbol{\beta}^T\hat{x}_i}))$

$F = \ell(\beta) + \frac{1}{2} \parallel \beta \parallel$

设ϕ为$x->F$的映射，在F中作对率回归，即$h(\hat{x}) = \beta^T\phi(\hat{x})$

由表示定理可得，$h(\hat{x}) = \sum_{i=1}^{m}\alpha_i\kappa(\hat{x},\hat{x_i})$，所以 $\boldsymbol{\beta} = \sum_{i=1}^{m}\alpha_i\phi(\boldsymbol{x}_i)$，带入上式可得：

$$
\begin{aligned}
F &= \sum_{i=1}^{m}(-y_i\boldsymbol{\beta}^T\phi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}) + ln(1+e^{\boldsymbol{\beta}^T\phi(\hat{\boldsymbol{x}}_{\boldsymbol{i}})})) + \frac{\boldsymbol{\lambda}}{2} \parallel \boldsymbol{\beta} \parallel^2 \\
&= \sum_{i=1}^{m}(-y_i\sum_{j=1}^{m}\alpha_j\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j) + ln(1+e^{\sum_{j=1}^{m}\alpha_j\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j)})) + \frac{\boldsymbol{\lambda}}{2} \parallel \sum_{j=1}^{m}\alpha_j\phi(\boldsymbol{x}_j) \parallel^2 \\
&= \sum_{i=1}^{m}(-y_i\sum_{j=1}^{m}\alpha_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) + ln(1+e^{\sum_{j=1}^{m}\alpha_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)})) + \frac{\boldsymbol{\lambda}}{2} \parallel \sum_{j=1}^{m}\alpha_i\alpha_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)
\end{aligned}
$$

目标函数为$\min\limits_{\alpha} F$，得到L2正则化下的核对率回归

2.4 支持向量回归的对偶问题如下，

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

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

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），

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

$$s.t. \quad C \geq \boldsymbol{\alpha} \geq 0 \quad and \quad \boldsymbol{\alpha}^T\boldsymbol{u} = 0$$

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

（注意题目有两个小问）

解：

1.

令$\boldsymbol{\alpha} = [\alpha_1,\alpha_2,\cdots,\alpha_m]^T, \boldsymbol{y} = [y_1,y_2,\cdots,y_m]^T, \boldsymbol{K}_{ij} = \kappa(\boldsymbol{x};\boldsymbol{x}_j), \boldsymbol{\epsilon}^* = [\epsilon,\epsilon,\cdots,\epsilon]^T$

$\boldsymbol{\alpha}^* = [\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}]^T, v = [-\boldsymbol{y} - \boldsymbol{\epsilon}^*, \boldsymbol{y} - \boldsymbol{\epsilon}^*]^T, \boldsymbol{K}^* = \begin{bmatrix} \boldsymbol{K} & -\boldsymbol{K} \\ -\boldsymbol{K} & \boldsymbol{K} \end{bmatrix}$

则

$$\begin{aligned}\sum_{i=1}^{m}(y_i(\hat{\alpha}_i - \alpha_i) - \epsilon(\hat{\alpha}_i + \alpha_i)) &= \sum_{i=1}^{m}(\alpha_i(-y_i - \epsilon)) + \hat{\alpha}_i(y_i - \epsilon)) \\ &= \boldsymbol{\alpha}^T(-\boldsymbol{y} - \boldsymbol{\epsilon}^*) + \hat{\boldsymbol{\alpha}}^T(\boldsymbol{y} - \boldsymbol{\epsilon}^*) \\ &= \boldsymbol{\alpha^*}^T \boldsymbol{v}\end{aligned}$$

$$\begin{aligned}-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) &= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_j - \alpha_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\hat{\alpha}_j \\ &\quad - \hat{\alpha}_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_j + \hat{\alpha}_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\hat{\alpha}_j) \\ &= -\frac{1}{2}(\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\boldsymbol{K}\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^T\boldsymbol{K}\boldsymbol{\alpha} + \hat{\boldsymbol{\alpha}}^T\boldsymbol{K}\hat{\boldsymbol{\alpha}}) \\ &= -\frac{1}{2}\boldsymbol{\alpha^*}^T\boldsymbol{K}^*\boldsymbol{\alpha^*}\end{aligned}$$

其中$\alpha_i^* = \alpha_i$或 $\hat{\alpha}_i$，所以$0 \leq \alpha^* \leq C$.。
综上，SVM的对偶问题可转化为标准型。

2.

设$\boldsymbol{x}$是m维向量，则

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T\boldsymbol{x}_j)^2 \qquad = ($$

$$\begin{aligned}&= (\sum_{u=1}^{m} x_{ui}x_{uj})(\sum_{v=1}^{m} x_{vi}x_{vj}) \\ &= \sum_{1\leq u,v\leq m} x_{iu}x_{iv}x_{ju}x_{jv} \\ &= [x_1x_1, x_1x_2, \cdots, x_1x_m, \cdots, x_2x_m, \cdots, x_mx_m]^T \times [x_1x_1, x_1x_2, \cdots, x_1x_m, \cdots, x_2x_m, \cdots, x_mx_m]^T\end{aligned}$$

则$\phi(x_i)$为一个m^2维向量，每个分量为$x_{iu}x_{iv}, 1 \leq u, v \leq m$，
$\phi(x_i) = [x_1x_1, x_1x_2, \cdots, x_1x_m, \cdots, x_2x_m, \cdots, x_mx_m]^T$，各分量互不相同。