HW5.

5.1 $f(x)=w^Tx$ 线性激活函数在定义域内变换情况相同. 而理想中的激活函数是能够把可能在较大范围内变化的输入值挤入到 (0,1) 输出值范围内的函数。

2. 讨论 $\frac{\exp(x_i)}{\sum_{j=1}^{c}\exp(x_j)}$ 和 $\log\frac{e^{x_i}}{\sum_{j=1}^{c}\exp(x_j)}$ 的数值溢出问题

实数在计算机中以二进制表示，数值过小会取0（下溢出）、数值过大会上溢出。令 $f(x)=\frac{\exp(x_i)}{\sum_{j=1}^{c}\exp(x_j)}$, $g(x)=\log\frac{e^{x_i}}{\sum_{j=1}^{c}e^{x_j}}$

当 $x_i\to-\infty$, $i=1,2,\cdots c$ 时 $\sum_{j=1}^{c}e^{x_j}\to0$，易发生下溢出

当 $x_i\to+\infty$, $i=1,2,\cdots,c$ 时，$\exp(x_i)$ 可能发生上溢出.

对 $f(x)$，令 $X=\max(x_i)$, $i=1,2,\cdots,c$，把 $f(x)$ 改为计算 $f(x_i-X)$ 即可避免溢出

$$\frac{e^{x_i}}{\sum_{j=1}^{c}e^{x_j}}=\frac{\frac{e^{x_i}}{X}}{\sum_{j=1}^{c}\frac{e^{x_j}}{X}}=\frac{e^{x_i-X}}{\sum_{j=1}^{c}e^{x_j-X}}\quad\begin{cases}e^{x_i-X}\le e^{M-M}=1，分子不会上溢出\\ \sum_{j=1}^{c}e^{x_j-M}\geqslant e^{M-M}=1，分母不会下溢出\end{cases}$$

$$\log\frac{e^{x_i}}{\sum_{j=1}^{c}e^{x_j}}=\log\frac{\frac{e^{x_i}}{e^X}}{\sum_{j=1}^{c}\frac{e^{x_j}}{e^X}}=\log\frac{e^{x_i-X}}{\sum_{j=1}^{c}e^{x_j-X}}=x_i-X-\log\sum_{j=1}^{c}e^{x_j-X}$$

$$\because 1=e^{X-X}\leq\sum_{j=1}^{c}e^{x_j-X}\leq\sum_{j=1}^{c}e^{X-X}=c$$

分子不会上溢出，分母不会下溢出

3. 计算 $\frac{\exp(x_i)}{\sum\limits_{j=1}^{c}\exp(x_j)}$ 和 $\log\frac{\exp(x_i)}{\sum\limits_{j=1}^{c}\exp(x_j)}$ 关于向量 $x=[x_1,\cdots,x_c]$ 的梯度 4.

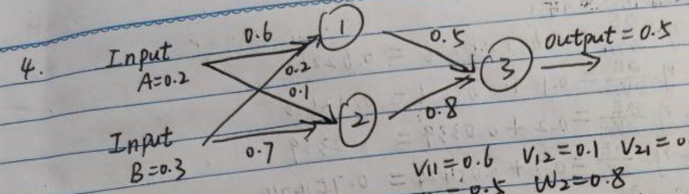
$$\frac{\partial \frac{\exp(x_i)}{\sum_{j=1}^{c}\exp(x_j)}}{\partial x_k}=\begin{cases}-\dfrac{\exp(x_i+x_k)}{\sum\limits_{j=1}^{c}\exp(x_j)^2}, & k\neq i\\[2ex] \dfrac{\exp(x_k)\sum\limits_{j=1,j\neq k}^{c}\exp(x_j)}{\sum\limits_{j=1}^{c}\exp(x_j)^2}, & k=i\end{cases}$$

$$\frac{\partial\left[\log\frac{\exp(x_i)}{\sum_{j=1}^{c}\exp(x_j)}\right]}{\partial x_k}=\begin{cases}-\dfrac{\exp(x_k)}{\sum\limits_{j=1}^{c}\exp(x_j)}, & k\neq i.\\[2ex] 1-\dfrac{\exp(x_k)}{\sum\limits_{j=1}^{c}\exp(x_j)}, & k=i\end{cases}$$

即有 $$\frac{\partial\frac{\exp(x_i)}{\sum_{j=1}^{c}\exp(x_j)}}{\partial x}=\frac{e^{x_i}}{(\sum\limits_{j=1}^{c}e^{x_j})^2}\left[-e^{x_1},\cdots,-e^{x_{i-1}},\sum_{m=1,m\neq i}^{c}e^{x_m},\cdots,-e^{x_c}\right],$$

$$\frac{\partial\left[\log\frac{\exp(x_i)}{\sum_{j=1}^{c}\exp(x_j)}\right]}{\partial x}=\frac{1}{\sum\limits_{j=1}^{c}e^{x_j}}\left[-e^{x_1},\cdots,-e^{x_{i-1}},\sum_{m=1,m\neq i}^{c}e^{x_m},\cdots,-e^{x_c}\right]$$

4.



$V_{11}=0.6$ $V_{12}=0.1$ $V_{21}=0.2$ $V_{22}=0.7$

$W_1=0.5$ $W_2=0.8$

ReLU: $\max(0,x)$ 当 $x>0$ 时 $ReLu'(x)=1$，当 $x\leq 0$ 时，$ReLu'(x)=0$

正向传播: 对于 ① 的输入 $\alpha_1 = 0.2\times0.6+0.3\times0.2=0.18$

$\beta_1=\max(0,x_1)=0.18$

对于 ② 的输入 $\alpha_2=0.2\times0.1+0.7\times0.3=0.23$

$\beta_2=\max(0,x_2)=0.23$

$\hat{Y}=\max(0,W_1\beta_1+W_2\beta_2)=\max(0,0.274)=0.274$

$E=\frac{1}{2}(Y-\hat{Y})^2=0.5\times(0.5-0.274)^2=0.025538$

计算梯度:

$\frac{\partial E}{\partial V_{11}}=\frac{\partial E}{\partial \hat{Y}}\frac{\partial \hat{Y}}{\partial Y}\frac{\partial Y}{\partial \beta_1}\frac{\partial \beta_1}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial V_{11}}=(\hat{Y}-Y)ReLu'(Y)\cdot W_1\cdot ReLu'(\alpha_1)\cdot A$

$=(-0.226)\times1\times0.5\times1\times0.2=-0.0226$

$\frac{\partial E}{\partial V_{12}}=\frac{\partial E}{\partial \hat{Y}}\cdot\frac{\partial \hat{Y}}{\partial Y}\cdot\frac{\partial Y}{\partial \beta_2}\cdot\frac{\partial \beta_2}{\partial \alpha_2}\cdot\frac{\partial \alpha_2}{\partial V_{12}}=(\hat{Y}-Y)ReLu'(Y)\cdot W_2\cdot ReLu'(\alpha_2)\cdot A$

$=(-0.226)\times1\times0.8\times1\times0.2=-0.03616$

$\frac{\partial E}{\partial V_{21}}=\frac{\partial E}{\partial \hat{Y}}\cdot\frac{\partial \hat{Y}}{\partial Y}\cdot\frac{\partial Y}{\partial \beta_1}\cdot\frac{\partial \beta_1}{\partial \alpha_1}\cdot\frac{\partial \alpha_1}{\partial V_{21}}=(\hat{Y}-Y)ReLu'(Y)\cdot W_1\cdot ReLu'(\alpha_1)\cdot B$

$=(-0.226)\times1\times0.5\times1\times0.3=-0.0339$

$\frac{\partial E}{\partial V_{22}}=\frac{\partial E}{\partial \hat{Y}}\cdot\frac{\partial \hat{Y}}{\partial Y}\cdot\frac{\partial Y}{\partial \beta_2}\cdot\frac{\partial \beta_2}{\partial \alpha_2}\cdot\frac{\partial \alpha_2}{\partial V_{22}}=(\hat{Y}-Y)ReLu'(Y)\cdot W_2\cdot ReLu'(\alpha_2)\cdot B$

$=(-0.226)\times1\times0.8\times1\times0.3=-0.05424$

$\frac{\partial E}{\partial W_1}=\frac{\partial E}{\partial \hat{Y}}\frac{\partial \hat{Y}}{\partial Y}\frac{\partial Y}{\partial W_1}=(\hat{Y}-Y)ReLu'(Y)\cdot\beta_1=(-0.226)\times1\times0.18=-0.04068$

$\frac{\partial E}{\partial W_2}=\frac{\partial E}{\partial \hat{Y}}\cdot\frac{\partial \hat{Y}}{\partial Y}\cdot\frac{\partial Y}{\partial W_2}=(\hat{Y}-Y)ReLu'(Y)\beta_2=(-0.226)\times1\times0.23=-0.05198$

取学习率 $\eta=1$，更新权重得

$$v_{11}' = v_{11} - \eta \cdot \frac{\partial E}{\partial v_{11}} = 0.6 + 0.0226 = 0.6226$$

$$v_{12}' = v_{12} - \eta \cdot \frac{\partial E}{\partial v_{12}} = 0.1 + 0.03616 = 0.13616$$

$$v_{21}' = v_{21} - \eta \cdot \frac{\partial E}{\partial v_{21}} = 0.2 + 0.0339 = 0.2339$$

$$v_{22}' = v_{22} - \eta \cdot \frac{\partial E}{\partial v_{22}} = 0.7 + 0.05424 = 0.75424$$

$$w_1' = w_1 - \eta \cdot \frac{\partial E}{\partial w_1} = 0.5 + 0.04068 = 0.54068$$

$$w_2' = w_2 - \eta \cdot \frac{\partial E}{\partial w_2} = 0.8 + 0.05198 = 0.85198$$

第二次正向传播：

$$\alpha_1' = 0.2 \times 0.6226 + 0.3 \times 0.2339 = 0.19469$$

$$\alpha_2' = 0.2 \times 0.13616 + 0.3 \times 0.75424 = 0.253504$$

$$\beta_1' = 0.19469$$

$$\beta_2' = 0.253504$$

$$\hat{y}' = 0.19469 \times 0.54068 + 0.253504 \times 0.85198 = 0.321245$$

$$E' = \frac{1}{2}(y - \hat{y}')^2 = \frac{1}{2} \times (0.5 - 0.321245)^2 = 0.015977 < E$$

综上，可见参数更新降低了均方误差

HW6.

6.4 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一个，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) 数量集线性可分时，SVM 和 LDA 等价。

6.6 (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，可能会对训练造成意想不到的结果。

综上，SVM 对噪声敏感。

6.9 对率回归的 L2 正则化目标函数.

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

$$F=l(\beta)+\frac{1}{2}\|\beta\|$$

由表示定理可知. $\beta=\sum_{i=1}^{m}\alpha_i\phi(x_i)$，带入上式可知.

$$F=\sum_{i=1}^{m}\left(-y_i\beta^T\phi(\hat{x}_i)+\ln\left(1+e^{\beta^T\phi(\hat{x}_i)}\right)\right)+\frac{\lambda}{2}\|\beta\|^2.$$

$$=\sum_{i=1}^{m}\left(-y_i\sum_{j=1}^{m}\alpha_j\phi(x_i)\phi(x_j)+\ln\left(1+e^{\sum_{j=1}^{m}\alpha_j\phi(x_i)\phi(x_j)}\right)\right)+\frac{\lambda}{2}\left\|\sum_{j=1}^{m}\alpha_j\phi(x_j)\right\|^2$$

$$=\sum_{i=1}^{m}\left(-y_i\sum_{j=1}^{m}\alpha_j k(x_i,x_j)+\ln\left(1+e^{\sum_{j=1}^{m}\alpha_j k(x_i,x_j)}\right)\right)+\frac{\lambda}{2}\left\|\sum_{j=1}^{m}\alpha_i\alpha_j k(x_i,x_j)\right\|^2$$

目标函数为 $\min_{\alpha} F$，得到 L_2 正则化下的核对率回归.

4. 把 $\max_{\alpha,\hat{\alpha}} g(\alpha,\hat{\alpha})=-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)k(x_i,x_j)+\sum_{i=1}^{m}\left(y_i(\hat{\alpha}_i-\alpha_i)-\varepsilon(\hat{\alpha}_i+\alpha_i)\right)$

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

转化为类似标准型 $\max_{\alpha} g(\alpha)=\alpha^T v-\frac{1}{2}\alpha^T K\alpha$

s.t. $C\geq\alpha\geq 0$ and $\alpha^T u=0$ 的形式.

① 令 $\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_m]^T$，$y=[y_1,y_2,\cdots,y_m]^T$.

$k_{ij}=k(x_i,x_j)$ $\quad \varepsilon^*=[\varepsilon,\varepsilon,\cdots,\varepsilon]^T$.

$\alpha^*=[\alpha,\hat{\alpha}]^T$，$V=[-y-\varepsilon^*,\ y-\varepsilon^*]$ $\quad K^*=\begin{bmatrix}K & -K\\ -K & K\end{bmatrix}$.

$u=[\underbrace{1,\cdots,1}_{m个},\underbrace{-1,\cdots,-1}_{m个}]^T$

则 $\sum_{i=1}^{m}(y_i(\hat{\alpha}_i-\alpha_i)-\varepsilon(\hat{\alpha}_i+\alpha_i)) = \sum_{i=1}^{m}(\alpha_i(-y_i-\varepsilon))+\hat{\alpha}_i(y_i-\varepsilon))$

$= \alpha^T(-y-\varepsilon^*)+\hat{\alpha}^T(y-\varepsilon^*)$

$= \alpha^{*T}V$

$-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)k(x_i,x_j))$

$= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i k(x_i,x_j)\alpha_j-\alpha_i k(x_i,x_j)\hat{\alpha}_j-\hat{\alpha}_i k(x_i,x_j)\alpha_j+\hat{\alpha}_i k(x_i,x_j)\hat{\alpha}_j)$

$= -\frac{1}{2}(\alpha^T K\alpha-\alpha^T K\hat{\alpha}-\hat{\alpha}^T K\alpha+\hat{\alpha}^T K\hat{\alpha})$

$= -\frac{1}{2}\alpha^{*T}K\alpha^*$

即有原 $\max_{\alpha,\hat{\alpha}} g(\alpha,\hat{\alpha}) = -\frac{1}{2}\alpha^{*T}K\alpha^*+\alpha^{*T}V$

再看约束条件：由 $C\geq\alpha,\hat{\alpha}\geq 0$ 且 $\alpha^*=[\alpha,\hat{\alpha}]^T$ 可知.

$\alpha_i^*=\alpha_i$ 或 $\hat{\alpha}_i$，$0\leq\alpha^*\leq C$.

由 $\sum_{i=1}^{m}(\alpha_i-\hat{\alpha}_i)=0$ 可知 $\alpha^{*T}u=0$.

综上，SVM的对偶问题可转化为题意所示的标准形.

② 在软间隔SVM中，$v=1$，$u=y$，$K[i,j]=y_iy_jk(x_i,x_j)$

$k(x_i,x_j)=\phi(x_i)^T\phi(x_j)=(x_i^Tx_j)^2$，求 $\phi(x_i)$ 表达式

设 x 是 m 维向量，则

$k(x_i,x_j)=(x_i^Tx_j)^2=(\sum_{u=1}^{m}x_{ui}x_{uj})(\sum_{v=1}^{m}x_{vi}x_{vj})$

$= \sum_{1\leq u,v\leq m} x_{iu}x_{iv}x_{ju}x_{jv}$

$= [x_1x_1, x_1x_2, \cdots x_1x_m, \cdots, x_2x_m\cdots, x_mx_m]^T$

$\times [x_1x_1, x_1x_2, \cdots, x_1x_m, \cdots, x_2x_m, \cdots, x_mx_m]^T$

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.