3.7

由逻辑函数的定义可知$y=w^Tx+b$会先经过sigmod函数映射到$(0,1)$区间上

$y=\frac{1}{1+e^{-(w^Tx+b)}}$

我们有

$$ln\frac{y}{1-y}=w^Tx+b$$

 则若讲y视为类后验概率估计$p(y=1|x)$，则

$$ln\frac{p(y=1|x)}{p(y=-1|x)}=w^Tx+b$$

显然有

$$p(y=1|x)=\frac{e^{w^Tx+b}}{1+e^{w^Tx+b}}=\frac{1}{e^{-(w^Tx+b)}+1}$$
$$p(y=-1|x)=\frac{1}{1+e^{w^Tx+b}}$$

所以

$$p(y_i|x)=\frac{1}{1+e^{-y_i(w^Tx+b)}}$$

则对数似然函数可以写为

$$l(w,b)=\sum_{i=1}^{N}logP(y_i|x_i)=\sum_{i=1}^{N}log(\frac{1}{1+e^{-y_i(w^Tx_i+b)}})=-\sum_{i=1}^{N}(1+exp(-y_i(w^Tx+b)))$$

3.8

由{-1,+1}逻辑回归模型对数似然函数的梯度为

$$\nabla E(w)=-\frac{1}{N}\sum_{n=1}^{N}\frac{y_nx_n}{1+e^{y_nw^Tx_n}}$$

所以，当某个样本对梯度的改变值越大，则说明它对训练的贡献更大

由改公式可知y_nx_n都是常量，所以只需要分母更小，及$e^{y_nw^Tx_n}$更小，因为e^x是在R上的增函数，所以当误分类时$y_iw^Tx_i<0$，正确分类时$y_jw^Tx_j>0$，所以肯定$y_iw^Tx_i<y_jw^Tx_j$，所以误分类对梯度的改变值更大，对训练的贡献更高。

3.9

由感知机的算法的收敛性可知

令$R=\max\limits_{1\le i\le N}|\hat{x}_i|$，则感知机算法在训练集上的误分类次数k满足不等式

$$k\le(\frac{R}{\gamma})^2,\gamma=\min_i\{y_i(w_{opt}\cdot x_i+b_{opt})\}$$

所以$R^2=3^2+(-9)^2=90$

若根据算法最后得出的模型为 $w_{opt}\cdot x_i+b_{opt}=0$

则其在T数据集上的误分类次数k最多为

$$k=\frac{90}{\min\limits_i\{y_i(w_{opt}\cdot x_i+b_{opt})\}}$$

若初始$w=(0,0)^T,b=0,\eta=1$

对$x_1=(1,3)^T,y_1(w_0+x_1+b_0)=0$，未能被正确分类，更新$w,b$

$$\begin{gathered}w_1=w_0+\eta y_1x_1=(1,3)^T,b_1=b_0+\eta y_1=1\\ \Rightarrow w_1\cdot x+b_1=1x^{(1)}+3x^{(2)}+1\end{gathered}$$

x_1,x_2,x_3,x_4,x_5五个点都被正确分类，则没必要再进行下去了

所以当取第4个点的时候，γ取得最小值

$$\gamma=\min_i\{y_i(w_{opt}\cdot x_i+b_{opt})\}=1$$

所以$k\le\frac{90}{1}=90$，所以k的最大值为90