HW4

1.

如果是 one-versus-one 的話，一次訓練的資料量是 2N/K。所以一次的訓練時間是$\frac{8\mathrm{aN}^3}{K^3}$。總共需要訓練$C_2^K = \frac{\mathrm{K(K-1)}}{2}$個 classifier，所以總共需要$\frac{8\mathrm{aN}^3}{K^3} * \frac{\mathrm{K(K-1)}}{2} = \frac{4\mathrm{aN}^3(K-1)}{K^2}$的訓練時間。

2.

squared error 的話就是$\sum_{n=1}^{N}(y_n w^T z_n - 1)^2$。我們可以像之前 linear regression 那樣把 X,w,y 的計算寫成矩陣形式。令最高次為 Q 的 Vandermonde matrix 是$\begin{pmatrix} 1 & \cdots & x_1^Q \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_N^Q \end{pmatrix}$，則$\begin{pmatrix} 1 & \cdots & x_1^Q \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_N^Q \end{pmatrix} w = \begin{pmatrix} w^T z_1 \\ w^T z_2 \\ \cdots \\ w^T z_N \end{pmatrix}$。題目要求說明存在某些 Q 使$\begin{pmatrix} w^T z_1 \\ w^T z_2 \\ \cdots \\ w^T z_N \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix}$。因為 V 是 invertible, 所以當 Q=N-1 時，w 的解就是$\mathrm{w} = \mathrm{V}^{-1}\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix}$。只要不存在誤差，那這個 w 的解就可以完美使$\begin{pmatrix} w^T z_1 \\ w^T z_2 \\ \cdots \\ w^T z_N \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix}$，也就是 squared error=0。

3.

把 x1 到 xN mapping 到 z1 到 zN，把 z1 到 zN 排成一個矩陣 Z，因為對於每個 zk 都是第 k 項是 1 其他都是 0，所以可以排成一個 N*N 的單位矩陣。也就是說，在這種情況下，最佳解 w 就是$\mathrm{w} = \mathrm{Z}^{-1}\mathrm{y} = \mathrm{y}$。這時，在沒有誤差的狀況下，因為$\mathrm{Zw} = \mathrm{y}$，$\mathrm{y} - \mathrm{y}$的向量每一項都是 0，所以 Ein 會是 0。

因為是在 uniform distribution 裡面取 testing data，因為選到某個特定數字的機率是 0，所以取到跟某個 training data 完全一樣的數字的機率是無限小的。按照題目的 transform，所有 testing data 作完 transform 都會對應到一個 N 維的 0 向量。這樣的話，w 乘上一個 0 向量得到結果是 0，err 就是 y^2。而 Eout 就是這個 y^2 的期望值。所以 Eout=E((x + ε)^2)=Var(ε)+E(ε)^2+E(x^2)=1+E(x)^2+Var(x)=4/3。其中因為 x 跟ε互相獨立，所以$\mathrm{E}(\mathrm{x}\varepsilon) = \mathrm{E}(\mathrm{x})\mathrm{E}(\varepsilon) = 0$

4.

$\mathrm{X}_{\mathrm{h}}^{T}X_h$是個 N*N 矩陣。原本$\mathrm{X}^{\mathrm{T}}\mathrm{X}$的第 ij 項是$\sum_{n=1}^{N} x_{ni}x_{nj}$，現在$\mathrm{X}_{\mathrm{h}}^{T}X_h$的第 ij 項是$\sum_{n=1}^{N} x_{ni}x_{nj}$+$\sum_{n=1}^{N}(x_{ni} + \varepsilon_{ni})(x_{nj} + \varepsilon_{nj})$。這裡$\varepsilon$是每個ϵ裡面的一個值。展開可得$\sum_{n=1}^{N} x_{ni}x_{nj} + \sum_{n=1}^{N}(x_{ni} + \varepsilon_{ni})(x_{nj} + \varepsilon_{nj}) = 2\sum_{n=1}^{N} x_{ni}x_{nj} + \sum_{n=1}^{N} \varepsilon_{ni}\varepsilon_{nj} + \sum_{n=1}^{N} \varepsilon_{ni}x_{nj} + \sum_{n=1}^{N} x_{ni}\varepsilon_{nj}$。因$\varepsilon$的 mean 是 0，variance 是$\frac{\delta^2}{3}$，所以期望值 E 是$2\sum_{n=1}^{N} x_{ni}x_{nj} +$

$E(\varepsilon^2)+\sum_{n=1}^{N}\varepsilon_{ni}x_{nj}+\sum_{n=1}^{N}x_{ni}\varepsilon_{nj}=2\sum_{n=1}^{N}x_{ni}x_{nj}+\frac{\delta^2}{3}=2\mathrm{X}^{\mathrm{T}}\mathrm{X}+\frac{\delta^2}{3}$

5.

$\mathrm{w_{t+1}}=w_t-\eta\nabla E_{aug}(w_t)=w_t-\eta\left(\nabla E_{in}(w_t)+\frac{2\lambda}{N}w_t\right)=$ $w_t\left(1-\frac{2\lambda\eta}{N}\right)-\eta\nabla E_{in}(w_t)=\left(1-\frac{2\lambda\eta}{N}\right)(w_t-\frac{\eta}{1-\frac{2\lambda\eta}{N}}\nabla E_{in}(w_t))$。所以 $\alpha=1-\frac{2\lambda\eta}{N}$，$\beta=\frac{\eta}{1-\frac{2\lambda\eta}{N}}$。

6.

令所有x_n的集合所形成的向量是 X，y_n對應的向量是 Y。直接代公式就是$\mathrm{w}^*=\frac{X^TY}{X^TX+\lambda}$。根據題目說明，C 是($\frac{X^TY}{X^TX+\lambda}$)^2，所以$\frac{X^TY}{X^TX+\lambda}=\sqrt{C}$，$X^TX+\lambda=\frac{X^TY}{\sqrt{C}}$，$\lambda=\frac{X^TY}{\sqrt{C}}-X^TX$。所以$\alpha=X^TY=\sum_{n=1}^{N}x_ny_n$ ，$\beta=-X^TX=-\sum_{n=1}^{N}x_n^2$。

7.

假設說 L1-regularized linear regression 得到的最佳解為w^*，則我們可以生出一個對應的$\mathrm{w}^{**}=\left(\mathrm{w}^{*T}V\right)^{\mathrm{T}}$。這時，$\mathrm{w}^{*T}\Phi(x_n)=\left(\mathrm{w}^{*T}\mathrm{V}\right)x_n=\mathrm{w}^{**}x_n$，所以$\frac{1}{\mathrm{N}}\sum_{n=1}^{N}(\mathrm{w}^{*T}\Phi(x_n)-y_n)+\frac{\lambda}{N}\|\mathrm{w}^*\|_1=\frac{1}{\mathrm{N}}\sum_{n=1}^{N}(\mathrm{w}^{**}x_n-y_n)+\frac{\lambda}{N}\|V^{-1}\mathrm{w}^{**}\|_1$。所以$\Omega(\mathrm{w})=\|V^{-1}\mathrm{w}^{**}\|_1$，其中$\mathrm{w}^{**}$就是題目要求的 optimal w，$\mathrm{w}^*$就是題目要求的 optimal $\widetilde{w}$，之

間的關係是$\mathrm{w} = \mathrm{V}\tilde{w}$。

8.

當操作 loocv 時，挑到的點如果是負的，則這時 training data 裡面有 N 個正的跟 N-1 個負的。這時 A 會回傳負的，所以做 validation 時會回傳負的，0/1 error 是 0。反過來也一樣，0/1 error 是 0。因此全部加起來，$\mathrm{E}_{\mathrm{loocv}}(A_{minority}) = 0$。

9.

回到 Training vs Testing 的課程。裡面有個 fun time 是說當有 5 個點在一個圓上的時候，能完全分割的 case 只有 22 個狀況。這題的點都在 x^2+y^2=2，所以有 10 個狀況無法線性分割。直接窮舉

當$\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3 = 1$，$\mathrm{x}_4, \mathrm{x}_5 = -1$時，$\min \mathrm{E}_{\mathrm{in}}(\mathrm{w}) = \frac{1}{5}$

當$\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3 = -1$，$\mathrm{x}_4, \mathrm{x}_5 = 1$時，$\min \mathrm{E}_{\mathrm{in}}(\mathrm{w}) = \frac{1}{5}$

當$\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_4 = 1$，$\mathrm{x}_3, \mathrm{x}_5 = -1$時，$\min \mathrm{E}_{\mathrm{in}}(\mathrm{w}) = \frac{1}{5}$

當$\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_4 = -1$，$\mathrm{x}_3, \mathrm{x}_5 = 1$時，$\min \mathrm{E}_{\mathrm{in}}(\mathrm{w}) = \frac{1}{5}$

當$\mathrm{x}_4, \mathrm{x}_5, \mathrm{x}_3 = 1$，$\mathrm{x}_1, \mathrm{x}_2 = -1$時，$\min \mathrm{E}_{\mathrm{in}}(\mathrm{w}) = \frac{1}{5}$

當$\mathrm{x}_4, \mathrm{x}_5, \mathrm{x}_3 = -1$，$\mathrm{x}_1, \mathrm{x}_2 = 1$時，$\min \mathrm{E}_{\mathrm{in}}(\mathrm{w}) = \frac{1}{5}$

當$\mathrm{x}_1, \mathrm{x}_4, \mathrm{x}_5 = 1$，$\mathrm{x}_2, \mathrm{x}_3 = -1$時，$\min \mathrm{E}_{\mathrm{in}}(\mathrm{w}) = \frac{1}{5}$

當$\mathrm{x}_1, \mathrm{x}_4, \mathrm{x}_5 = -1$，$\mathrm{x}_2, \mathrm{x}_3 = 1$時，$\min \mathrm{E}_{\mathrm{in}}(\mathrm{w}) = \frac{1}{5}$

當$x_5, x_2, x_3 = 1$，$x_4, x_1 = -1$時，$\min E_{in}(w) = \frac{1}{5}$

當$x_5, x_2, x_3 = -1$，$x_4, x_1 = 1$時，$\min E_{in}(w) = \frac{1}{5}$

所以答案是$\frac{1}{5} * 10/32 = \frac{1}{16}$。

10.

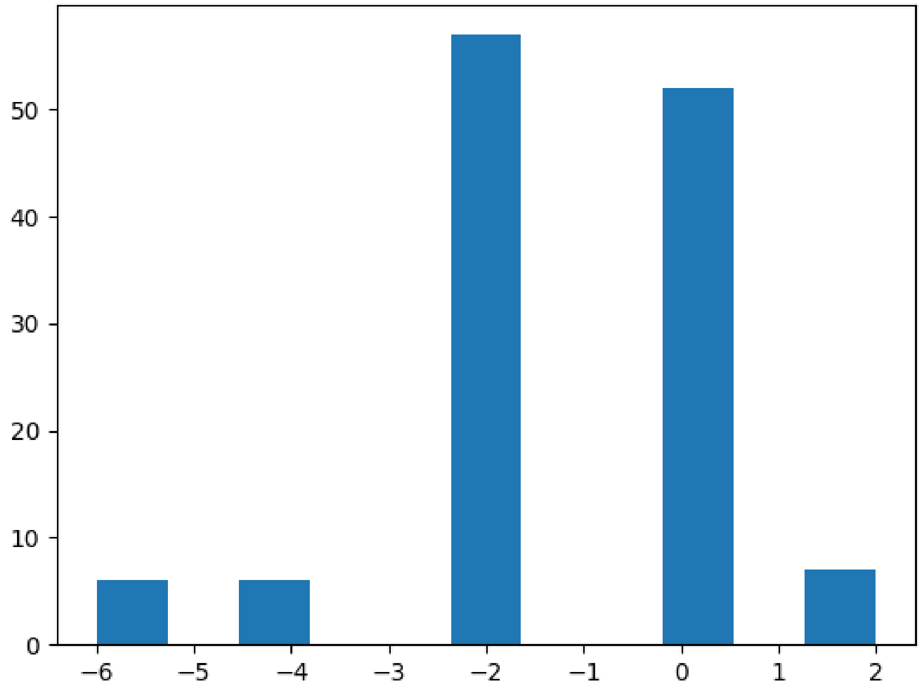
從-6 到 2，5 次結果是

Accuracy = 98% (196/200) (classification)
Accuracy = 93% (186/200) (classification)
Accuracy = 91% (182/200) (classification)
Accuracy = 88% (176/200) (classification)
Accuracy = 80.5% (161/200) (classification)

所以最好的λ是-6

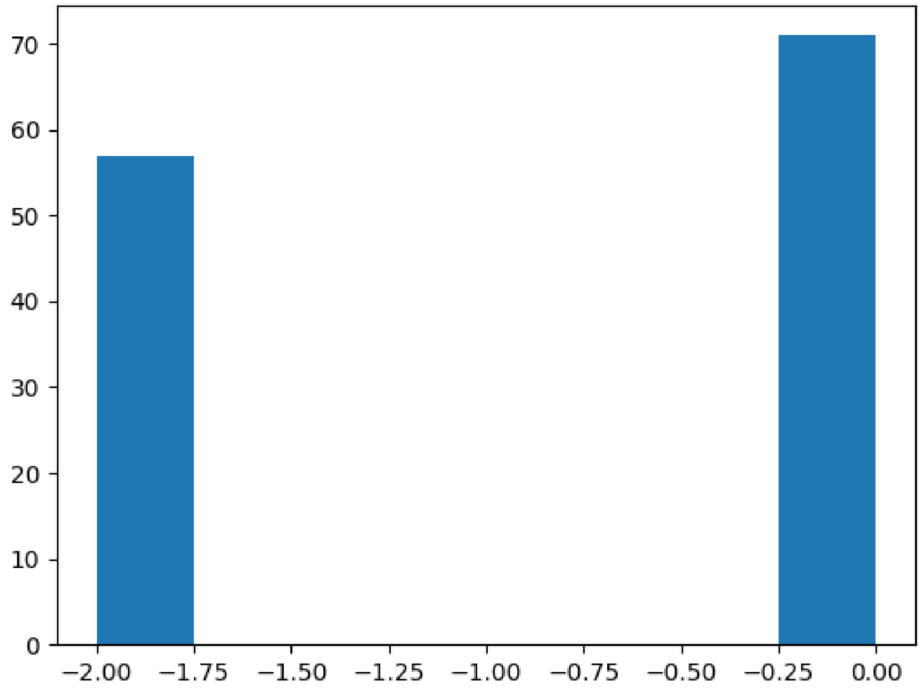
11.

這是我畫的圖



12.

這是我畫的圖



跟前一張圖比較，明顯在$\lambda = 0$的地方更集中。第 10 題的部分，當λ越小，則 regularize 的能力就越差，Ein 就會越小。第 11 題的部分，可以看出來說λ集中在 0 跟-2，代表 validation 有在發揮作用。Ein 比較接近 Eout。第 12 題的部分，因為每筆 datas 都做了 5 次 validation，而且用的 training data 更多，所以誤差更小，Ein 更接近 Eout，所以得到的λ更加集中。

13.

Linear regression 在 augment error 下的公式解是$w_{\text{reg}} =$

$(\mathrm{X^TX} + \lambda\mathrm{I})^{-1}X^Ty$。而在正常情況下的公式解是$\mathrm{w_{lin}} = (\mathrm{X^TX})^{-1}X^Ty$。當$\mathrm{X^TX} = \alpha\mathrm{I}$時，$\mathrm{w_{reg}} = (\mathrm{X^TX} + \lambda\mathrm{I})^{-1}X^Ty = (\alpha\mathrm{I} + \lambda\mathrm{I})^{-1}X^Ty = \frac{X^Ty}{\lambda+\alpha}$，$\mathrm{w_{lin}} = (\mathrm{X^TX})^{-1}X^Ty = \frac{X^Ty}{\alpha}$，兩者間只差常數倍，所以$\mathrm{w_C}$的 scaling 操作就可以把$\mathrm{w_{lin}}$轉換成$\mathrm{w_{reg}}$，可以解決 C-constrained 的問題。

如果$\mathrm{w_{reg}} = \mathrm{w_C}$的話，則$(\mathrm{X^TX} + \lambda\mathrm{I})^{-1}X^Ty = (\mathrm{X^TX})^{-1}X^Ty * \frac{\sqrt{C}}{\|\mathrm{w_{lin}}\|}$，所以$X^Ty * \frac{\sqrt{C}}{\|\mathrm{w_{lin}}\|} = (\mathrm{X^TX})(\mathrm{X^TX} + \lambda\mathrm{I})^{-1}X^Ty$，這時$X^Ty$是$(\mathrm{X^TX})(\mathrm{X^TX} + \lambda\mathrm{I})^{-1}$的特徵向量，$\frac{\sqrt{C}}{\|\mathrm{w_{lin}}\|}$是 eigenvalue。令$(\mathrm{X^TX})(\mathrm{X^TX} + \lambda\mathrm{I})^{-1} = A$，則$\mathrm{X^TX} = (\mathrm{X^TX} + \lambda\mathrm{I})A = \mathrm{X^TXA} + \lambda\mathrm{A}$，$\mathrm{X^TX(I - A)} = \lambda\mathrm{A}$