Problem 1

1.1

$E=\frac{1}{2}\sum_{n=1}^{N}(w^Tx_n+w_0-t_n)^2$ 关于 w_0 为二次.

∴最小值在 $\frac{\partial E}{\partial w_0}=0$ 处取到 $\Rightarrow \sum_{n=1}^{N}(w^Tx_n+w_0-t_n)=0$

$\Rightarrow w_0=\frac{1}{N}\sum_{n=1}^{N}(t_n-w^Tx_n)=-\frac{1}{N}\sum_{n=1}^{N}w^Tx_n=-w^Tm$

1.2

$\frac{\partial E}{\partial w}=\sum_{n=1}^{N}(w^Tx_n+w_0-t_n)x_n=0 \Rightarrow \sum_{n=1}^{N}(x_nx_n^T-x_nm^T)w=N(m_1-m_2)$

$S_W=\sum_{C_1}x_nx_n^T-N_1m_1m_1^T+\sum_{C_2}x_nx_n^T-N_2m_2m_2^T$

$S_B=m_2m_2^T+m_1m_1^T-m_1m_2^T-m_2m_1^T$

$S_W+\frac{N_1N_2}{N}S_B=\sum_{n=1}^{N}x_nx_n^T-\frac{N_1^2m_1m_1^T}{N}-\frac{N_2^2m_2m_2^T}{N}-\frac{N_1N_2m_1m_2^T}{N}-\frac{N_1N_2m_2m_1^T}{N}$

$=\sum_{n=1}^{N}x_nx_n^T-N\cdot\left(\frac{N_1m_1+N_2m_2}{N}\right)\left(\frac{N_1m_1+N_2m_2}{N}\right)^T=\left(\sum_{n=1}^{N}x_nx_n^T\right)-Nmm^T$

$=\sum_{n=1}^{N}(x_nx_n^T-x_nm^T)$

$\therefore (S_W+\frac{N_1N_2}{N}S_B)W=N(m_1-m_2)$

1.3

$S_WW=N(m_1-m_2)-\frac{N_1N_2}{N}(m_1-m_2)(m_1-m_2)^Tw$

令 $R=(m_1-m_2)^Tw$

有 $S_WW=\left(N-\frac{N_1N_2R}{N}\right)(m_1-m_2)$

$\therefore w\propto S_W^{-1}(m_1-m_2)$

Problem 2.

2.1 $S_B=S_T-S_W=\sum_{n=1}^{N}(x_nx_n^T-x_nm^T)=\sum_{n=1}^{N}(x_nx_n^T-mm^T)$

$-\sum_{k=1}^{K}\sum_{n\in C_k}(x_nx_n^T-m_km_k^T)$

$=\sum_{k=1}^{K}N_km_km_k^T-Nmm^T$

$=\sum_{k=1}^{K}N_k(m_k-m)(m_k-m)^T$

2.2. $y = W^T x$

$$S_W' = \sum_{k=1}^{K} \sum_{n \in C_k} (W^T x_n - W^T m_k)(W^T x_n - W^T m_k)^T$$

$$= W^T S_W W$$

同理 $S_B' = W^T S_B W$

2.3 $$J(W) = \frac{\prod_{i=1}^{D'} W_i^T S_B W_i}{\prod_{i=1}^{D'} W_i^T S_W W_i}$$

2.4 最大化 $J(W) = \frac{\prod_{i=1}^{D'} W_i^T S_B W_i}{\prod_{i=1}^{D'} W_i^T S_W W_i}$ 由于上下对每个i可单独放缩 ∴ 引入约束 $W_i^T S_W W_i = \tau_i$, $i=1,\cdots D'$

Lagranian 为 $L = \log \prod_{i=1}^{D} W_i^T S_B W_i - \lambda_i (W_i^T S_W W_i - \tau_i)$

$$\frac{\partial L}{\partial W_i} = \frac{S_B W_i}{W_i^T S_B W_i} - \lambda_i S_W W_i = 0$$

令 $W_i^T S_B W_i = R_i$ 有 $S_B W_i = \lambda_i R_i S_W W_i$

$\Rightarrow S_W^{-1} S_B W_i = \lambda_i R_i W_i$

~~对于 对于不同的 引入与前面 W_i 线性 不能设为~~

2.5 $D' = rank(S_B)$ 因为 $S_W^{-1} S_B$ 的0特征值对应的特征向量作为feature并无意义

若 $S_W^{-1} S_B x = 0$ 且由 S_W^{-1} 可逆 $\Rightarrow S_B x = 0 \Rightarrow x^T S_B x = 0$ 类间距离为0

∴ $S_W^{-1} S_B$ 的非零特征向量可作为feature. ∴ $D' = rank(S_W^{-1} S_B) = rank(S_B)$

而 $rank(S_B) = rank(\sum_{k=1}^{K} N_k (m_k - m)(m_k - m)^T) \leq K$

∴ $D' \leq K$

Problem 3

3.1 $P(\text{mistake}) = \sum_{k=1}^{K} P(x \in R_k, C_k^1) = \sum_{k=1}^{K} \int_{x \in R_k} P(x, C_k^1)$

$= \sum_{k=1}^{K} \int_{R_k} (1 - P(x, C_k)) dx$

3.2 用 x_i 为 feature 判决为 ω_1 的概率为: $P(\omega_1 | x_i = 1) = \frac{P(\omega_1) P(x_i = 1 | \omega_1)}{P(x_i = 1)} = \frac{\alpha_i}{\alpha_i + \beta_i}$

$P(\omega_2 | x_i = 1) = \frac{\beta_i}{\alpha_i + \beta_i}$

∴ 当 $x_i = 1$ 时判决为 ω_2

$x_i = 0$ 时判决为 ω_1

$P(\text{error}(x_i)) = P(x_i = 1 | \omega_1) P(\omega_1) + P(x_i = 0 | \omega_2) P(\omega_2)$

$= \frac{1}{2}\alpha_i + \frac{1}{2}(1 - \beta_i) = \frac{1}{2} \cdot \frac{1}{2}(\beta_i - \alpha_i)$

由于 $\beta_1 - \alpha_1 > \beta_2 - \alpha_2 > \beta_3 - \alpha_3 \Rightarrow e(x_1) < e(x_2) < e(x_3)$ □

这是由于第 i 维不同类别 ω_j ($j = 1, 2$) 取值的差别 $|\beta_i - \alpha_i|$ 越大，对判决更有效果

3.3 $e(x_1) = \frac{1}{2} - \frac{1}{2}(\beta_1 - \alpha_1) = 0.1$

$e(x_2) = \frac{1}{2} - \frac{1}{2}(\beta_2 - \alpha_2) = 0.125$

$e(x_3) = \frac{1}{2} - \frac{1}{2}(\beta_3 - \alpha_3) = 0.155$

$P(\omega_1 | \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}) = \frac{P(\omega_1) P(x_1 = 1 | \omega_1)(x_2 = 1 | \omega_1)}{P(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix})} = \frac{\alpha_1 \alpha_2}{\alpha_1 \alpha_2 + \beta_1 \beta_2}$

$P(\omega_2 | \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}) = \frac{\beta_1 \beta_2}{\alpha_1 \alpha_2 + \beta_1 \beta_2} > P(\omega_1 | \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix})$

$P(\omega_1 | \begin{pmatrix} x_i \\ x_j \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}) = \frac{\alpha_i(1 - \alpha_j)}{\alpha_i(1 - \alpha_j) + \beta_i(1 - \beta_j)}$

同此表格对 (x_i, x_j) 的 4 种取值，分子为

(x_i, x_j)	0 0	0 1	1 0	1 1
ω_1	$(1-\alpha_i)(1-\alpha_j)$	$(1-\alpha_i)\alpha_j$	$\alpha_i(1-\alpha_j)$	$\alpha_i\alpha_j$
ω_2	$(1-\beta_i)(1-\beta_j)$	$(1-\beta_i)\beta_j$	$\beta_i(1-\beta_j)$	$\beta_i\beta_j$

在(X_1,X_2) (X_1,X_3), (X_2,X_3) 下feature情况下的判别：

(X_1,X_2):

	ω_1		ω_2
00	0.855	>	0.02
10	0.095	<	0.18
01	0.045	<	0.08
11	0.005	<	0.72

$$e^{(X_1,X_2)} = P\left(\begin{pmatrix}X_1\\X_2\end{pmatrix}\neq\begin{pmatrix}0\\0\end{pmatrix}\middle|\omega_1\right)P(\omega_1)+P\left(\begin{pmatrix}X_1\\X_2\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}\middle|\omega_2\right)$$
$$= 0.0725 + 0.01 = 0.0825$$

(X_1,X_3):

	ω_1		ω_2
00	0.891	>	0.03
10	0.099	<	0.27
01	0.009	<	0.07
11	0.001	<	0.63

$$e(X_1,X_3) = (1-0.891)\frac{1}{2} + 0.03\cdot\frac{1}{2}$$
$$= 0.0685$$

(X_2,X_3):

	ω_1		ω_2
00	0.8405	>	0.06
10	0.0895	<	0.24
01	0.0095	<	0.14
11	0.0005	<	0.56

$$e(X_2,X_3) = (1-0.8405+0.06)\cdot\frac{1}{2}$$
$$= 0.05975$$

发现在选2个 feature 时 选取 X_2,X_3 是 error 最小的。而1个 feature 时 X_1 是 error 最小的

这里因为在选择多个 feature 用于决策时 贝叶斯决策的中 X_1,X_2 并不独立，选择能获得更多信息的 feature 组合不能只是单个 feature 的信息

Problem 4

4.1 $J(w;\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(y_i - w^T\phi(x_i))^2 + \lambda\|w\|_1$

前一项可分 对其中 $\frac{\partial \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(y_i - w^T\phi(x_i))^2}{\partial w_k} = \frac{1}{n}\sum_{i=1}^{n}(y_i + w^T\phi(x_i))\phi_k(x_i)$

$w^T\phi(x_i) = w_k^T\phi_k(x_i) + w_k\phi_k(x_i)$

$= \frac{1}{n}\sum_{i=1}^{n} w_k\phi_k^2(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi_k(x_i)(y_i - w_{-k}^T\phi_{-k}(x_i))$

$= a_k w_k - c_k$

第二项用subgradient: 其 $\frac{\partial \lambda\|w_k\|_1}{\partial w_k} = \begin{cases} -\lambda & w_k<0 \\ [-\lambda,\lambda] & w_k=0 \\ \lambda & w_k>0 \end{cases}$

当 $w_k=0$ 时等于 $a_k w_k = 0$

$\therefore \partial_{w_k} J(w;\lambda) = \begin{cases} a_k w_k - c_k - \lambda, & w_k<0 \\ [-c_k-\lambda, -c_k+\lambda], & w_k=0 \\ a_k w_k - c_k + \lambda, & w_k>0 \end{cases}$

其中 $a_k = \frac{1}{n}\sum_{i=1}^{n}\phi_k^2(x_i)$

$c_k = \frac{1}{n}\sum_{i=1}^{n}\phi_k(x_i)(y_i - w_{-k}^T\phi_{-k}(x_i))$

c_k 为不考虑 $\phi_k(x_i)$ 与当前 w_k 的情况下需要 w_k 来弥补的残差在 $\phi_k(x_i)$ 方向投影之和

4.2 (a) 当 $c_k < -\lambda$ 时, $-c_k - \lambda > 0$, $0 \notin [-c_k-\lambda, -c_k+\lambda]$

$a_k w_k - c_k - \lambda = 0 \Rightarrow w_k = -\frac{(-c_k-\lambda)}{a_k}$ $w_k<0$

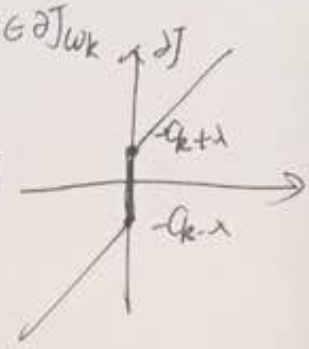
$a_k w_k - c_k + \lambda = 0 \Rightarrow w_k = \frac{-\lambda + c_k}{a_k}$ $w_k>0$ 舍

$\therefore w_k = \frac{c_k+\lambda}{a_k}$ 有 $0 \in \partial J_{w_k}$

(b) $c_k \in [-\lambda, +\lambda]$

$-c_k - \lambda < 0$ $\therefore$ 因为

$-c_k + \lambda > 0$

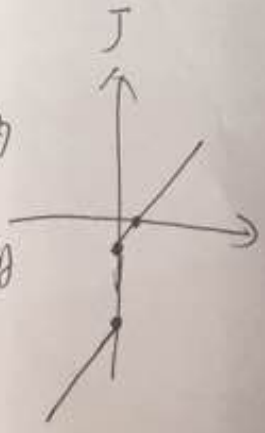


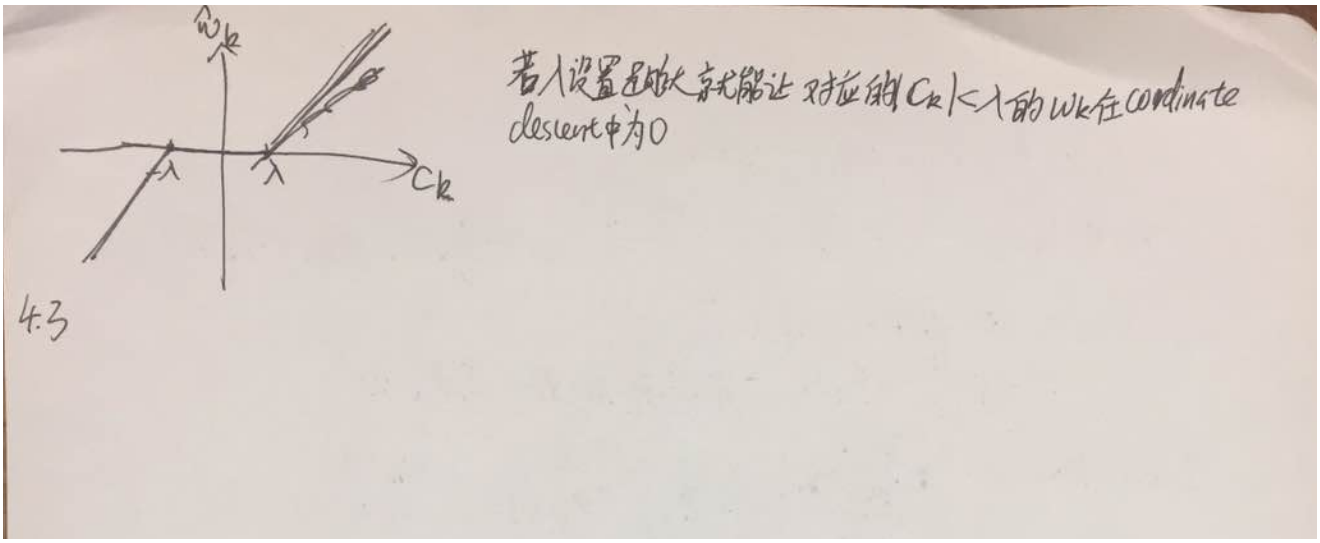
在 $w_k=0$ 时取为 $0 \in \partial J_{w_k}$

(c) $c_k > \lambda$ $-c_k - \lambda < 0$ $-c_k + \lambda < 0$ 因为

$a_k w_k - c_k + \lambda = 0 \Rightarrow w_k = \frac{c_k-\lambda}{a_k}$ >0 取为

$0 \in \partial J_{w_k}$



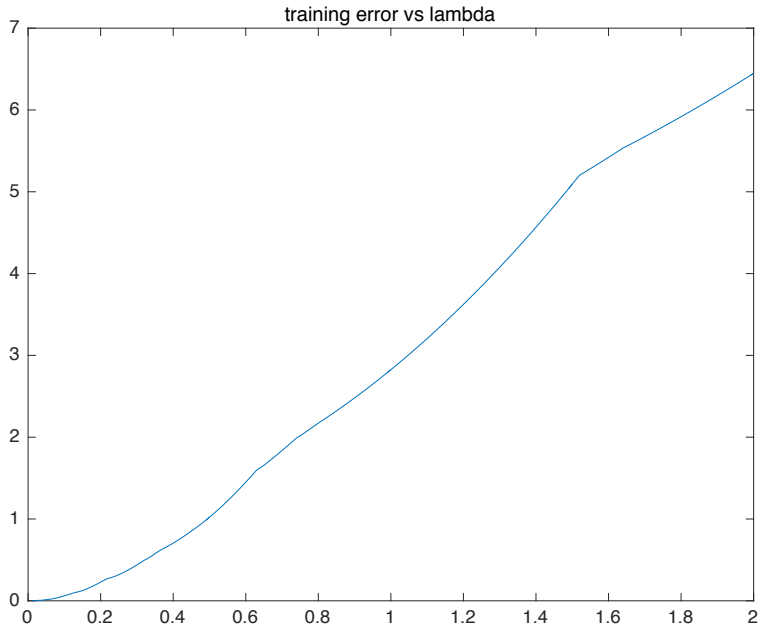


Problem 4 EXP:

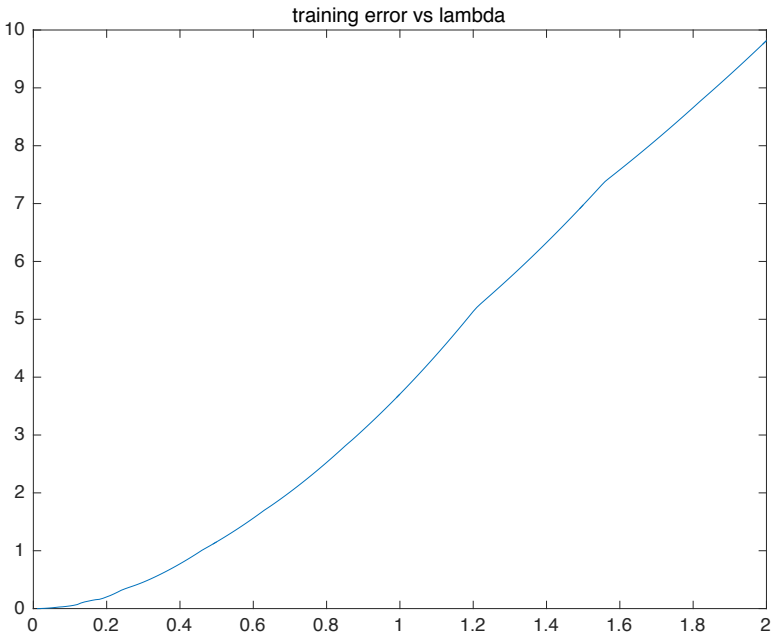
4.3, 4.4:

training error 和 lambda 的关系如下图所示。可以看出随着 regularization 强度的增加，在训练集上的 error 使单调增加的。随着训练数据集的增大，同一个表示能力的模型在训练数据集上的绝对 error 值也在增大。

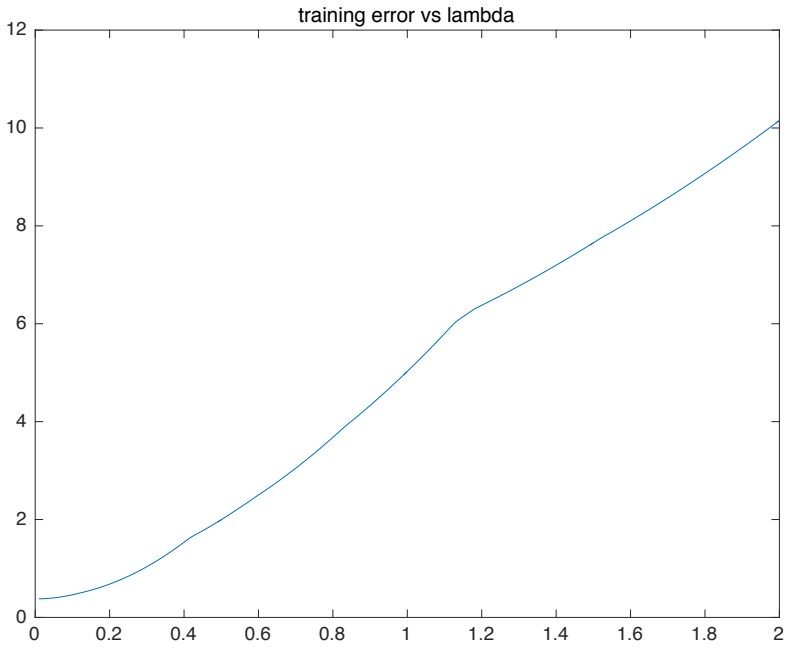
train_small:



train_mid:

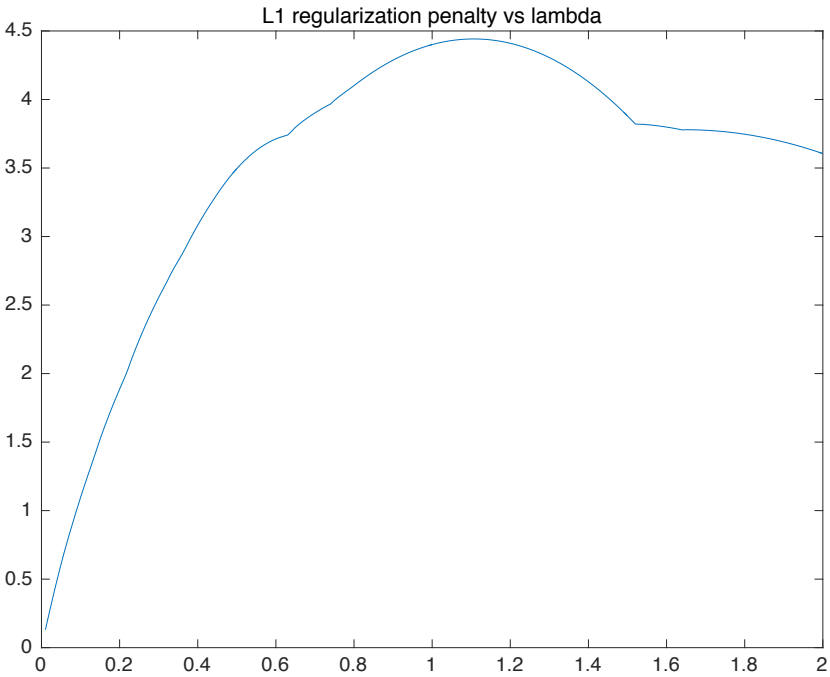


train_large:

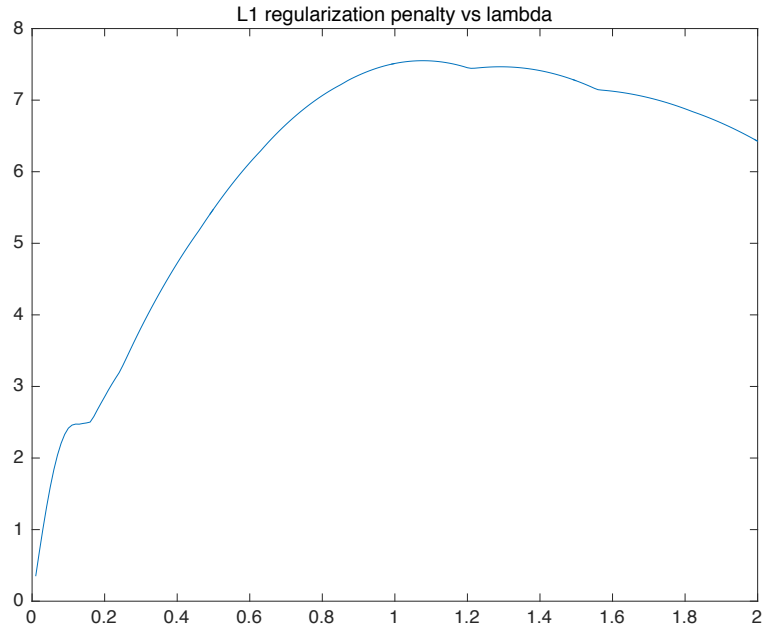


L1 regularization penalty 和 lambda 的关系如下图所示。可以看出 L1 regularization penalty 随着 lambda 的增加先是增加的，但是当 lambda 大到一定程度，因为对应投影残差在该阈值以下的 feature 的 weight 被拉到了 0，regularization penalty 会减少。

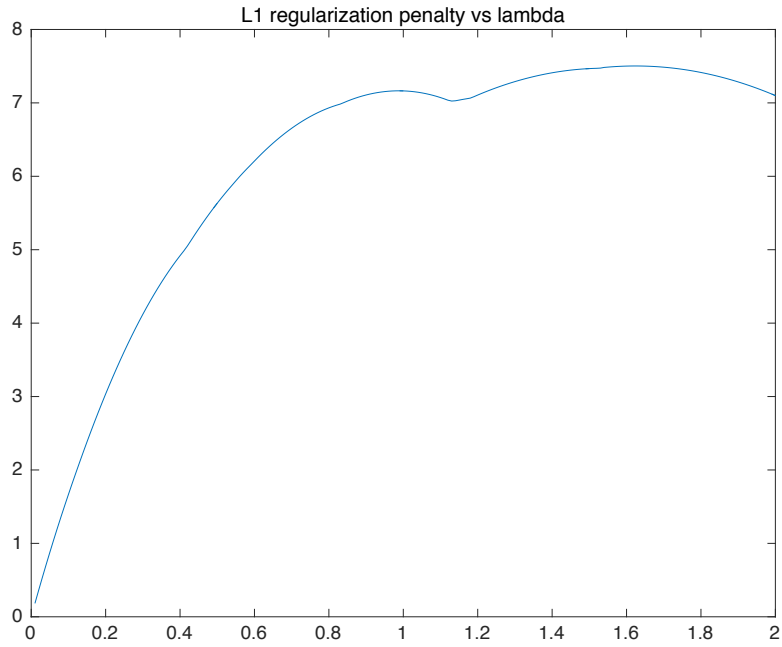
train_small:



train_mid:

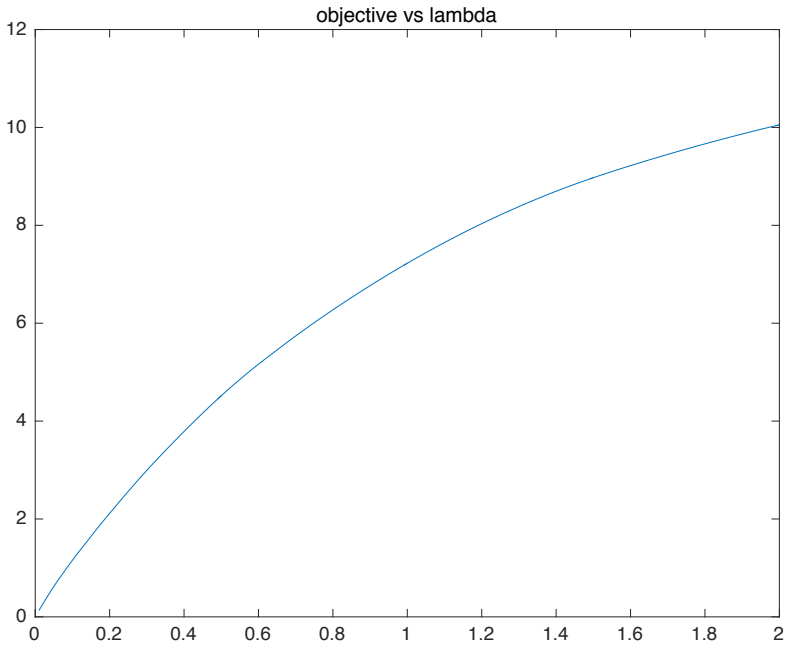


train_large:

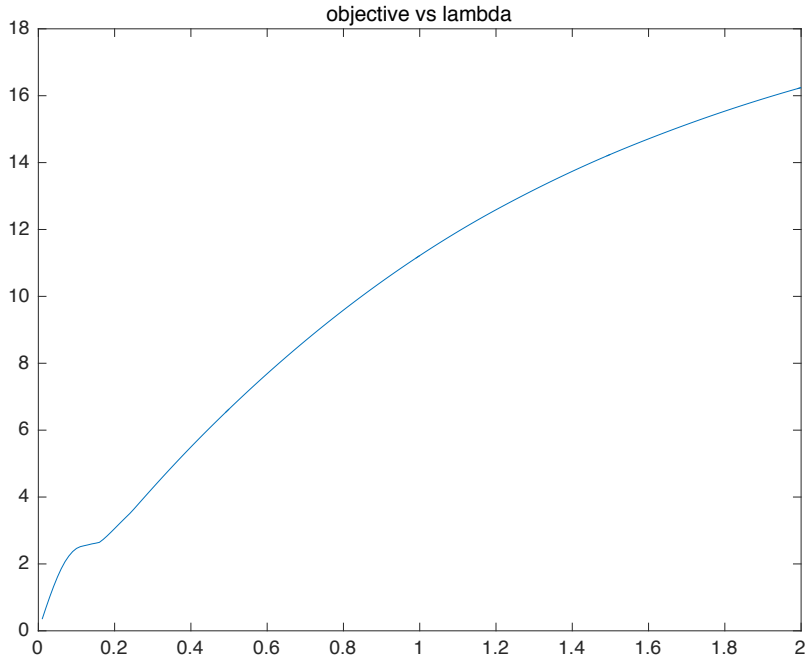


objective function 和 lambda 的关系如下图所示。可以看到，随着 lambda 的增加，也即加入了更多对系数的 penalty，同时对系数的 penalty 会使得 training error 增加，上面两项之和的总的训练的 objective function 一直在增加。

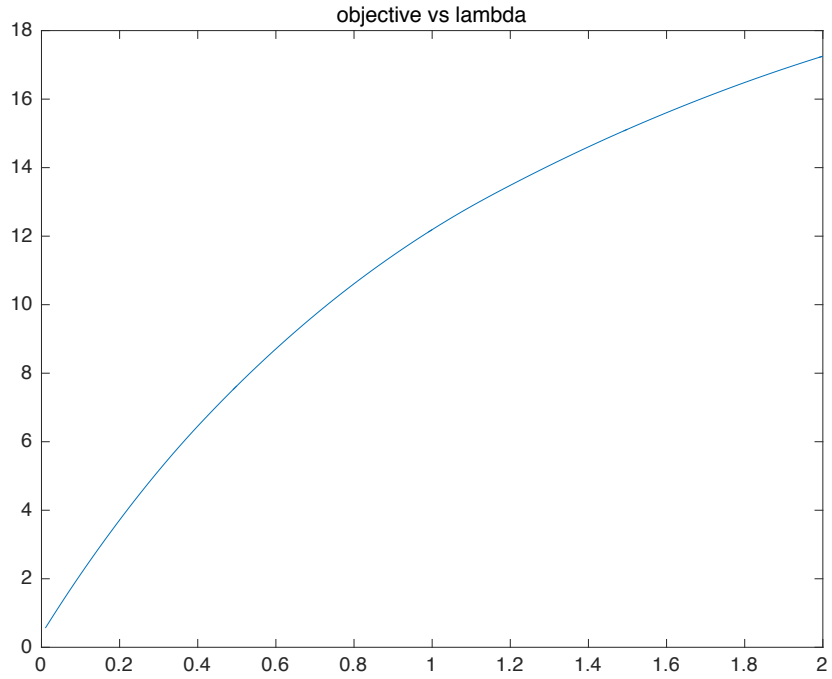
train_small:



train_mid:

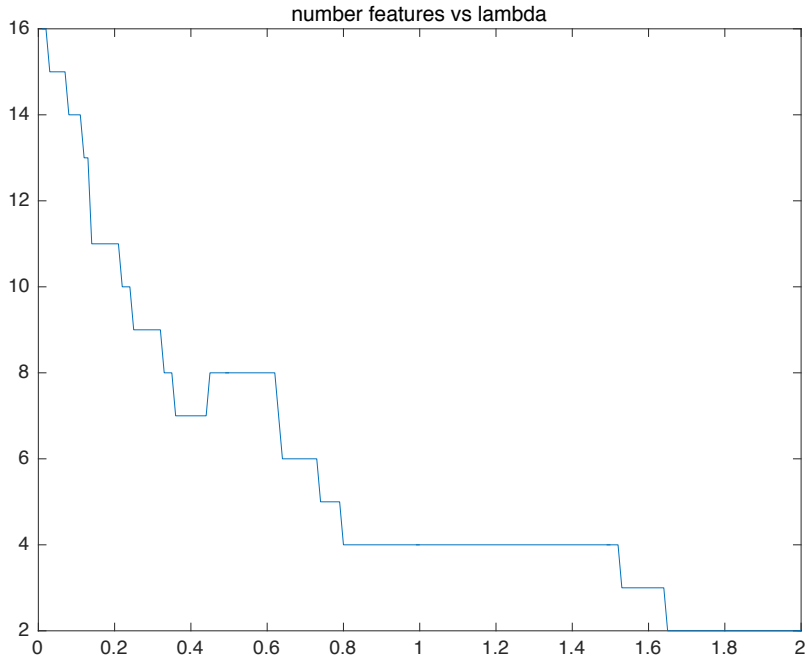


train_large:

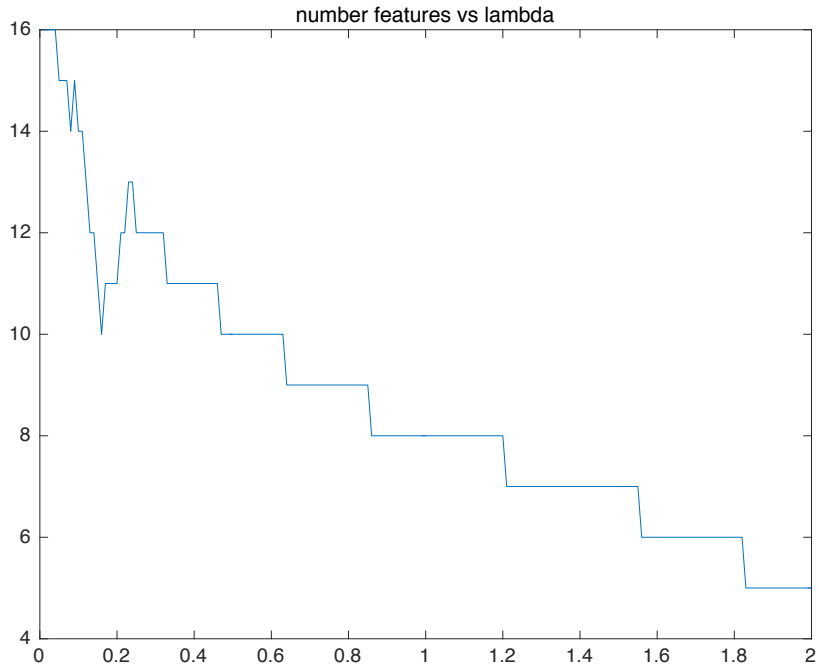


number of features 和 lambda 的关系如下图所示。可以看出基本上随着 lambda 的变大，系数的 0 阶 norm，也就是不为 0 的 lambda 值一直在变少。

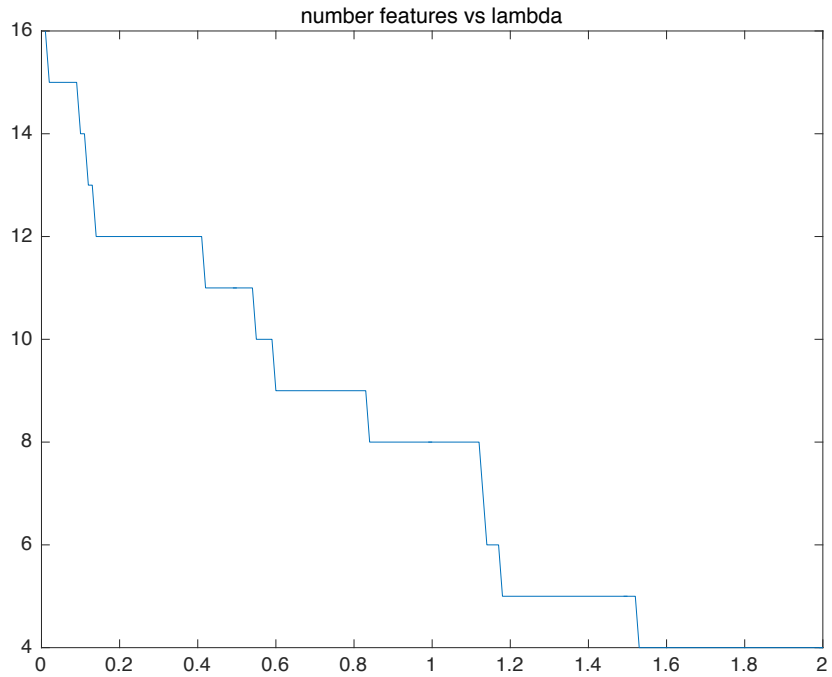
train_small:



train_mid:

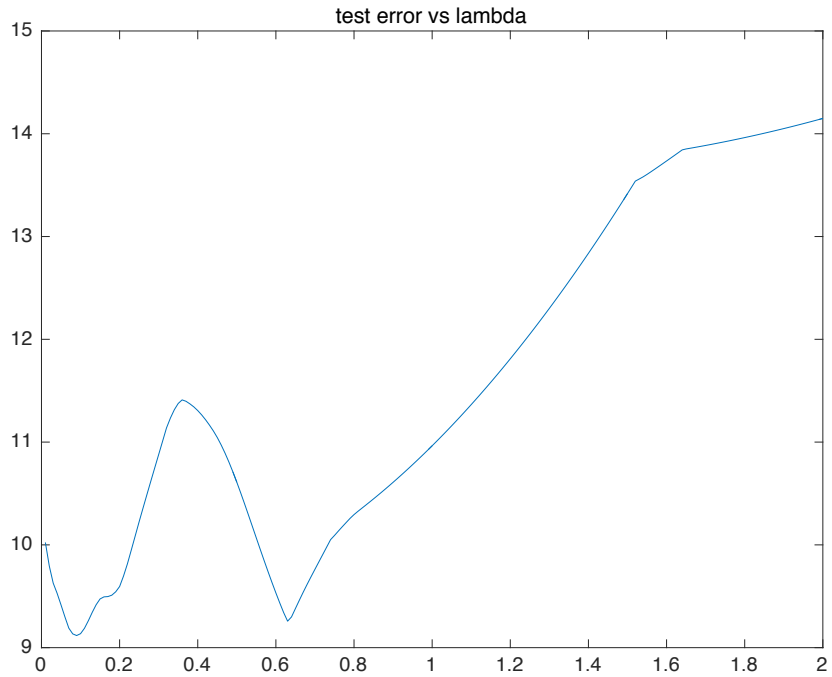


train_large:

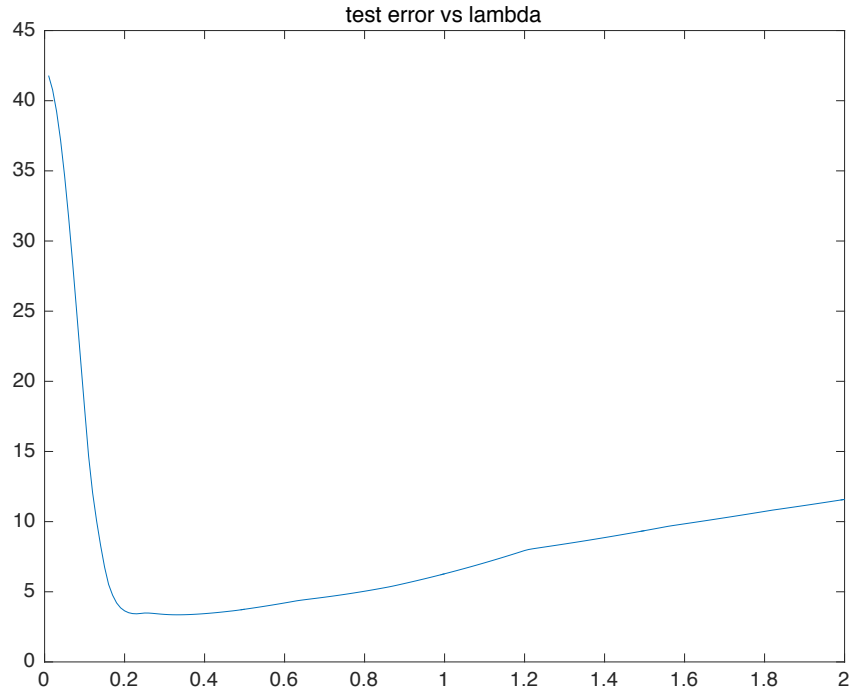


最后是在 test 集上的错误率在不同大小的训练数据集上和 lambda 的关系。可以看出，随着训练数据集的增大，训练数据增多，更加不容易 overfitting，模型在测试集上的效果更好。为了达到在测试集上最好的泛化效果，对模型加入的 L1 penalty 应该要根据训练集大小取一个适当的值，在训练数集据很小时，模型容易 overfitting，应该取较大的 lambda 值。而随着训练数据增多，模型不易 overfitting 此时可以取较小的 lambda 值。

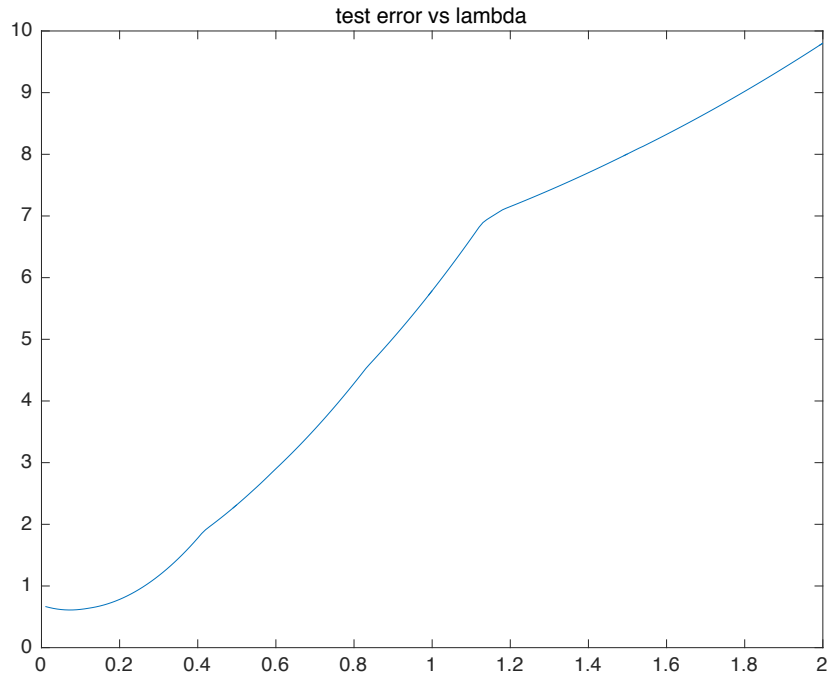
train_small:



train_mid:



train_large:



	small	mid	large
Size	8	16	64
lambda	0.09 (0.63)	0.33	0.07
Feature number	14(7)	11	15
Test error	9.11791 (9.258)	3.358542	0.6108

上表给出了在测试集上效果最好时，对应三个训练集应该使用的 regularization 强度 lambda，和对应的有效 feature 个数，和测试集错误。注意到在训练数据很少时，在 lambda 取一个很小的 0.09 的值时，却意外之中的达到了比较好的测试数据集上的效果，但是 lambda 继续增加一点开始波动，并不是一个平滑稳定的曲线，我认为这是由于训练数据太少，训练和测试结果都不够稳定和准确。可能在第一张图中位于 lambda=0.63，test error 为 9.258,有效 feature 个数为 7 的第二个极小值点的数据更为可靠。

从这些实验结果可以看出，在训练数据集较小时，应该把 regularization 的强度增加，在 L1 情况下就是减少有效 feature 个数，防止 overfitting。从这次试验的结果看来，取 lambda 为训练数据集 size 的一个反比貌似是比较不错的选择。

Problem5

One-versus-one:

对每一对不同的类别 pair 算出投影类间方差和投影类内方差比值最大的 criteria 得到的 w 方向。只取了 10 类，一共 45 个 feature。

Multi-class:

按照 problem2 算出$S_w^{-1}S_b$的前 K 个特征向量为投影方向。其中 K 为S_b的 rank，在这里取为 10(只取了 10 类，在前面 problem2 有分析)。

计算过程中要注意由于每一类也只有 10 个数据，所以 one-versus-one 里 Sw 一般都为奇异，我这里给他加了一个很小的 epsilon*单位阵的正则项。

用 KNN 分类器分类结果如下所示：(aaa 来不及写了)

讨论：基于 one-versus-one feature 的的分类器 inference 的时候需要计算 CLS*(CLS-1)/2 次再做 majority voting，基于 multi-class 协同设计的 feature 可以在上面设计 one-versus-all 或者也可以设计 one-versus-one 的分类器。就灵活程度来说，multi-class 设计的 feature 更灵活，而且需要保存的投影向量 w 也更少。One-versus-one 的分类器本身就可能存在 ambiguos 的区域，没有 majority 的到 voting 的。从 3 类的情况就可以看出，每个类别可能得到一票，那么就没有 majority 的票数。