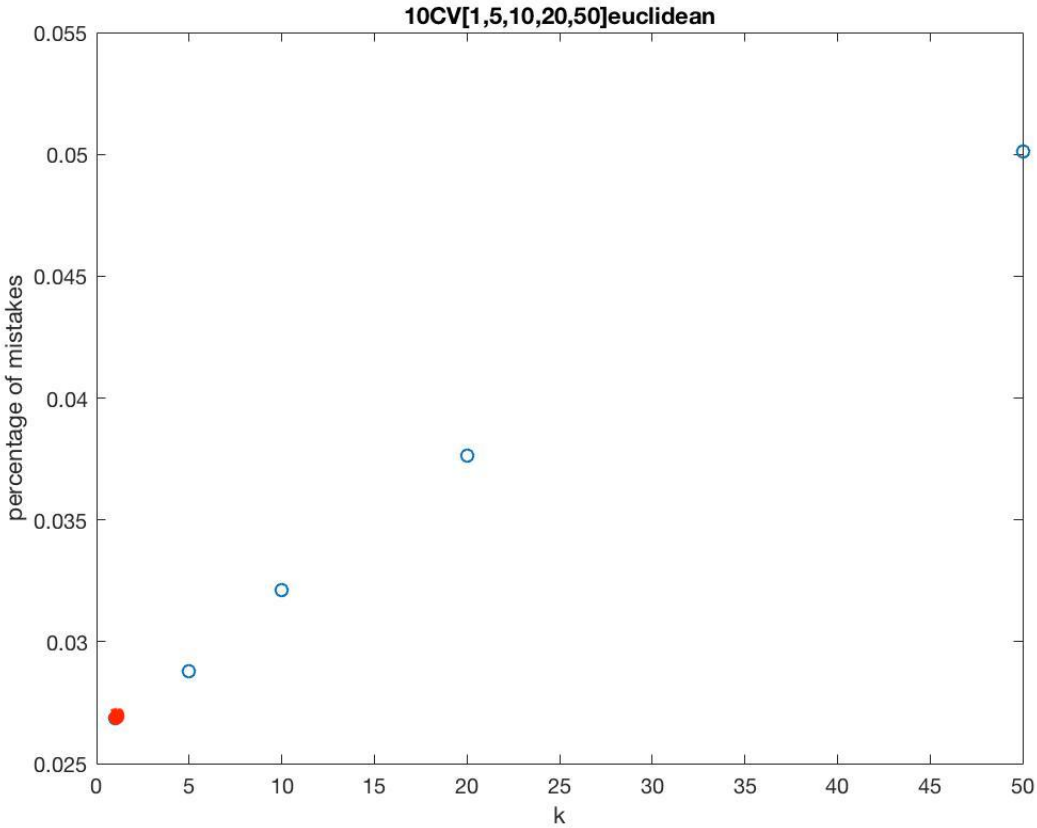
CS 698: Assignment 2

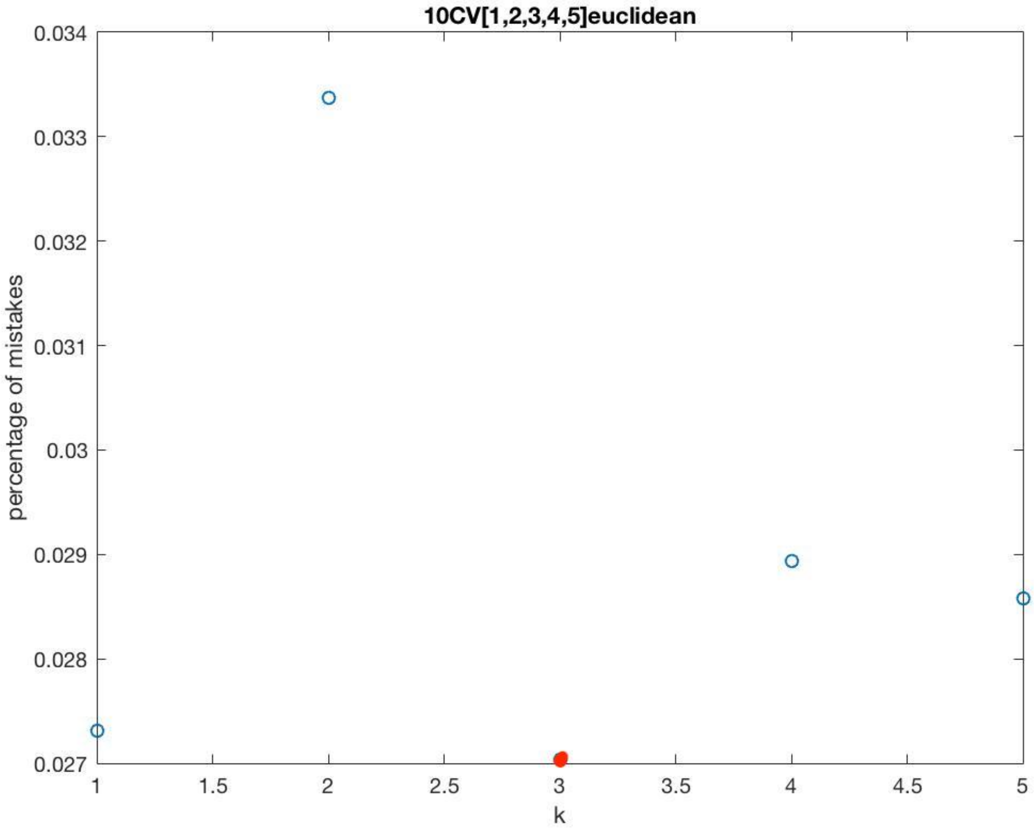
Ronghao Yang

October 12, 2017

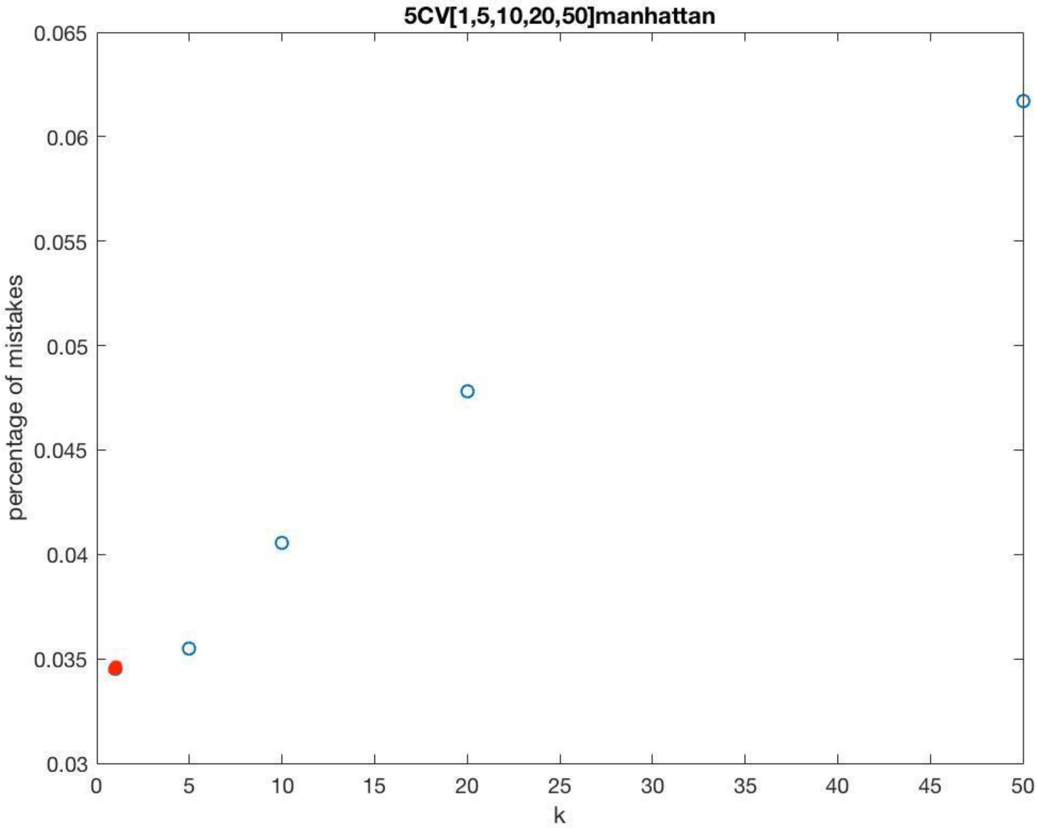
1 Question 1



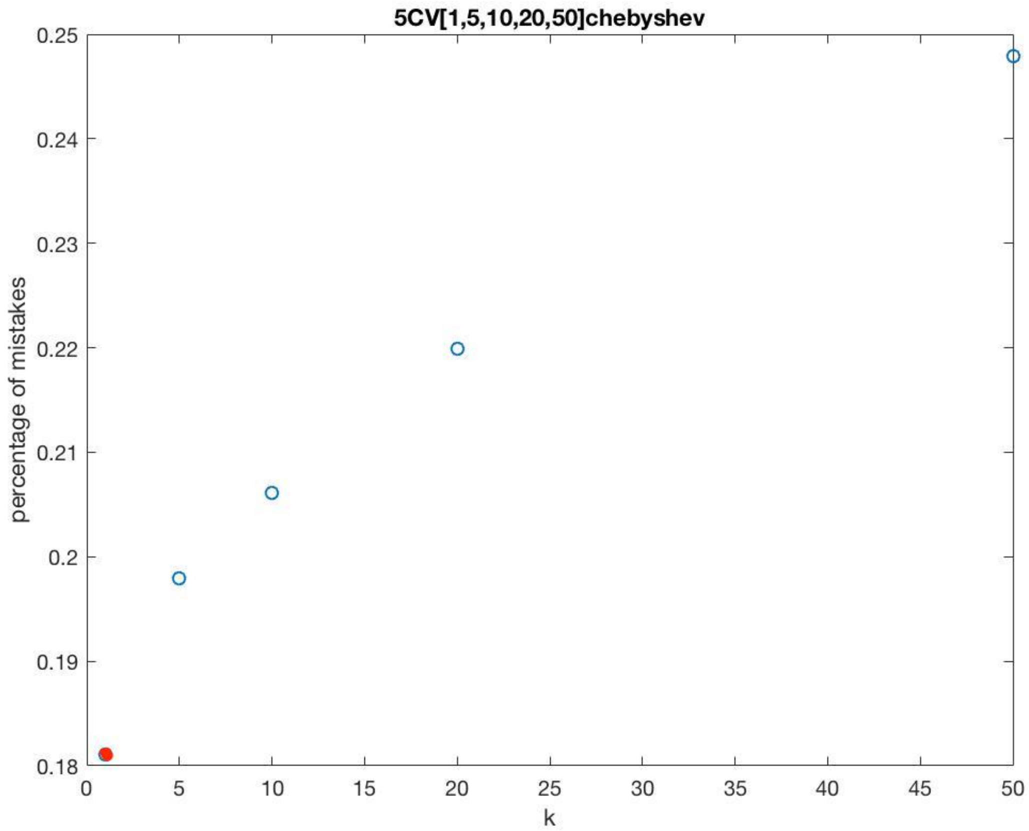
For the first case, we set k to be 1, 5, 10, 20, 50, and the distance metric to be *euclideandistance*. When using 10-fold cross validation on the training set, the optimal k is 1. When using k=1 on the testing set, the error rate is 3.09%



For the second case, we set k to be 1, 2, 3, 4, 5, and the distance metric to be *euclideandistance*. When using 10-fold cross validation on the training set, the optimal k is 3. When using k=3 on the testing set, the error rate is 2.95%



For the first case, we set k to be 1, 5, 10, 20, 50, and the distance metric to be *manhattandistance*. When using 5-fold cross validation on the training set, the optimal k is 1. When using k=1 on the testing set, the error rate is 3.69%



For the first case, we set k to be 1, 5, 10, 20, 50, and the distance metric to be *chebyshevdistance*. When using 10-fold cross validation on the training set, the optimal k is 1. When using k=1 on the testing set, the error rate is 17.41%

2 Question 2

$$\underline{Pr(f(x) \neq Y \mid X = x)} = 1 - Pr(f(x) = Y \mid X = x)$$

$$= 1 - Pr(f(x)=1, Y=1 \mid X=x) - Pr(f(x)=2, Y=2 \mid X=x) - \cdots - Pr(f(x)=m, Y=m \mid X=x)$$

We can show that $f(x) = i$ and $Y = i$ are independent.

$$= 1 - [E[1_{f(x)=1} P(Y=1 \mid X)] + E[1_{f(x)=2} P(Y=2 \mid X) + \cdots + E[1_{f(x)=m} P(Y=m \mid X)]]$$

If we want to minimize $Pr(f(x) \neq Y \mid X = x)$.

We need to maximize $E[1_{f(x)=1} P(Y=1 \mid X)] + E[1_{f(x)=2} P(Y=2 \mid X)] + \cdots E[1_{f(x)=m} P(Y=m \mid X)]$

$$= E[1_{f(x)=1} P(Y=1 \mid X) + 1_{f(x)=2} P(Y=2 \mid X) + \cdots 1_{f(x)=m} P(Y=m \mid X)]$$

Since $1_{f(x)=i} = 1$ only for one i, to ~~maximize~~ minimize, $Pr(f(x) \neq Y \mid X = x)$, we need to maximize $Pr(f(x) = Y \mid X = x)$, therefore if $f^*(x) = \underset{m=1,\dots,C}{\arg\max} Pr(Y=m \mid X)$, then $Pr(f(x) = Y \mid X = x)$ is maximized.

3 Question 3

3.1 Question 3.1

Question 3.1

$$\min_{w,b} \frac{c}{2}\|w\|_2^2 + \sum_{i=1}^{n} (\rho - y_i \hat{y}_i)_+ , \quad \text{where } \hat{y}_i = w^T x_i + b$$

if we let $w' = w/\rho$, $c' = c * \rho$

Then we have

$$\min_{w,b} \frac{c}{2}\| w'\rho\|_2^2 + \sum_{i=1}^{n} (\rho(1 - y_i(w'^T x_i + b)))_+$$

$$\Rightarrow \min_{w,b} \frac{c'}{2\rho} \cdot \|w'\|_2^2 \cdot \rho^2 + \rho \cdot \sum_{i=1}^{n} (1 - y_i \hat{y}_i)_+ , \quad \text{where } \hat{y}_i = w'^T x_i + b$$

$$\Rightarrow \min_{w,b} \frac{c'}{2}\|w'\|_2^2 \cdot \rho + \rho \cdot \sum_{i=1}^{n} (1 - y_i \hat{y}_i)_+$$

$$\Rightarrow \min_{w,b} \frac{c'}{2}\|w'\|_2^2 + \sum_{i=1}^{n} (1 - y_i \hat{y}_i)_+$$

So ρ can be reduced to 1

]

3.2 Question 3.2

Question 3.2

If $w' = wC$, $\rho' = \rho C$.

$$\min_{w,b,\rho} \frac{C}{2}\|w\|_2^2 - \nu\rho + \sum_{i=1}^{n}(\rho - y_i(w^T x_i))_+ \quad \text{[ignore } b \text{ in this case]}$$

$$= \min_{w,b,\rho} \frac{C}{2}\|Cw\|^2 \cdot \frac{1}{C^2} - \nu \cdot \frac{\rho C}{C} + \sum_{i=1}^{n}\left(\frac{\rho C}{C} - \frac{y_i(w^T C x_i)}{C}\right)_+$$

$$= \min_{w,b,\rho} \frac{1}{C}\left[\frac{1}{2}\|Cw\|^2 - \nu(\rho C) + \sum_{i=1}^{n}(\rho C - y_i(w^T C x_i))_+\right]$$

$$= \frac{1}{C}\min_{w,b,\rho} \frac{1}{2}\|Cw\|^2 - \nu(\rho C) + \sum_{i=1}^{n}(\rho C - y_i(w^T C x_i))_+$$

Therefore, letting $w' = wc$, $\rho' = \rho c$ can leave the problem reduced to c being 1.

]

3.3 Question 3.3

uestion 3.3.

Introducing the slack variable ε_i here. $\forall i$, $(1-y_i\hat{y}_i)_+ \le \varepsilon_i$

The lagrangian is

$$\min_{w,b}\ \max_{\alpha\ge 0,\beta\le 0}\ \frac{1}{2}\|w\|_2^2 - \nu\rho + \sum_i \varepsilon_i + \alpha_i(\rho - y_i\hat{y}_i - \varepsilon_i) + \beta\varepsilon_i$$

$$\rightarrow \max_{\alpha\ge 0,\beta\le 0}\ \min_{w,b}\ \frac{1}{2}\|w\|_2^2 - \nu\rho + \sum_i \varepsilon_i + \alpha_i(\rho - y_i\hat{y}_i - \varepsilon_i) + \beta\varepsilon_i$$

$$\rightarrow \max_{\alpha\ge 0,\beta\le 0}\ \min_{w,b}\ \frac{1}{2}\|w\|_2^2 - \nu\rho + \sum_i \varepsilon_i + \alpha_i(\rho - y_i(w^Tx_i+b) - \varepsilon_i) + \beta\varepsilon_i$$

$$\frac{d}{dw} = w - \sum\alpha_i y_i x_i = 0\ ; \quad \frac{d}{d\rho} = -\nu + \sum_i \alpha_i = 0$$

$$\frac{d}{db} = -\sum\alpha_i y_i = 0\ ; \quad \frac{d}{d\varepsilon_i} = 1 - \alpha_i + \beta_i = 0$$

Plug back. $\min \frac{1}{2}\|\sum\alpha_i y_i x_i\|_2^2$ +

$\cancel{\sum \alpha_i = 0}$, $\alpha_i y_i$ $\quad \sum_i \alpha_i y_i = 0, \quad \nu = \sum_i \alpha_i, \quad 0 \le \alpha_i \le 1$]

3.4 Question 3.4

Question 3.4.

We first sort a_i from the smallest to the largest

$$a_1 \le a_2 \cdots \le a_n$$

Let $a_j \le p \le a_{j+1}$

So the initial minimization function becomes.

$$\min_p -kp + \sum_{i=1}^{j} (p - a_i)$$

$$\rightarrow \min_p -kp + jp - \sum_{i=1}^{j} a_i$$

$$\rightarrow \min_p (j-k)p - \sum_{i=1}^{j} a_i.$$

1) when $j > k$:

when j increases by 1, $(j-k)p$ increases by p

$(\sum_{i=1}^{j} a_i)$ increases by less than p

So the entire equation increases. ~~Since~~ So we want j to be as small as possible

2) when $j < k$:

when j decreases by 1, $(j-k)p$ decreases by p

$(\sum_{i=1}^{j} a_i)$ decreases by more than p

So the entire equation increases as j decreases. Since we want to minimize it, so we want j to be as big as possible

Therefore, if we let $j=k$, we can achieve the minimal. $a_k \le p \le a_{k+1}$, $\min_p -kp + \sum_{i=1}^{n} (p-a_i)_+ = -\sum_{i=1}^{k} a_i$]

3.5 Question 3.5

Question 3.5

In hard margin S.V.M problem, the two convex halls are well seperated, therefore, we can formulate the minimal (squared) distance between convex hulls.

However, in soft S.V.M, convex hulls are not well seperated, the previous formulation does not work in this case.

From question 3.3, if we do the following substitutions

$$w = \frac{w}{v}, \quad u = \frac{2}{v}, \quad \rho = \frac{\rho}{v}, \quad \xi_i = \frac{\xi_i}{v}, \quad b = \frac{b}{v}.$$

Then we have the new minimization problem

$$\min \|w\|^2 - 2\rho + u \sum \xi_i$$ whose dual is

$$\min \frac{1}{4} \sum_{i,j}\sum \alpha_i \alpha_j y_i y_j x_i x_j$$ subject to

$$\sum \alpha_i y_i = 0, \quad \sum \alpha_i = 2, \quad 0 \le \alpha_i \le u$$

Since $0 \le \alpha_i \le u$, we can control u to shrink the convex hull to be a sub-set of the original convex hull.

Therefore, a better v can help ~~with the problem~~ make the formulation work.

[Some insights from "A Geometric Interpretation of v-SVM Classifiers]

]