Homework 3
Decision Tree, KNN, Kernel SVM

CMU 10-701: Machine Learning (Spring 2017)

OUT: Feb 27
DUE: March 10, 11:59 PM

START HERE: Instructions

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 3.4"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Each derivation/proof should be completed on a separate page. Submissions can be handwritten, but should be labeled and clearly legible. Else, submissions can be written in LaTeX. Upon submission, label each question using the template provided by Gradescope.
- **Programming**: All programming portions of the assignments should be submitted to Gradescope as well. We will not be using this for autograding, but rather for plagiarism detection, meaning you may use any language which you like to submit.

Section A : Multiple Choice Questions [10 points] (Dan)

1. **[2 points]** Suppose that you are using ridge regression to estimate the relationship between your data and a value of interest. Your estimate for $\hat{\beta}$ is given by:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

Where λ is a hyper-parameter that governs how much to penalize model complexity. Which of the following would be valid ways to use a 10-fold cross validation scheme? Select all that apply.

A) For different values of λ, train on 9 of the folds and estimate the risk on the 10th fold. Select λ by using the value that has the lowest risk on this 10th fold. Repeat this procedure 10 times for each possible hold out and use the mean of the risks for the estimators selected at each round as an estimate of the true risk of the model.

B) For different values of λ, train on 8 of the folds and estimate the risk on the 9th fold. Select λ by using the value that has the lowest risk on this 9th fold. Then estimate the risk using this λ on the 10th fold. Repeat this procedure 45 (10 choose 2) times and use the mean of the risks for the estimators selected each round as an estimate of the true risk of the model.

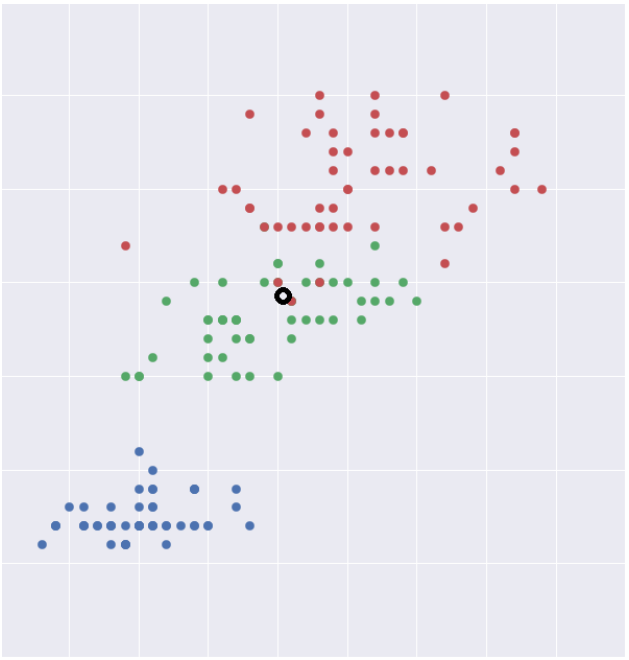
C) For different values of λ, train on 1 of the folds, and estimate the risk on this same fold. Select λ by using the value that has the lowest risk on this fold. Repeat this procedure 10 times (once for each fold), and use the mean of the risks for the estimators as an estimate of the true risk of the model.

D) Select a value of λ before doing your experiment. Train on 9 of the folds and estimate the risk on the 10th fold. Repeat this procedure 10 times (once for each fold), and use the mean of the risks for the estimators as an estimate of the true risk of the model.

Answer:A

- For B: The selected λ is from the 9th fold, but is not the optimized one from 10th fold; therefore, the estimated risk from 10th fold cannot be used to estimate the true risk.
- For C: The selected λ is not from a validation dataset.
- For D: There should be different λ values for model selection from validate dataset.

2. **[2 points]** For a k-Nearest Neighbor classifier and the data set below, which class will the test point (marked by the black circle) be classified as for each of the following values of k?



(i) $k = 1$ **Answer:**A
(ii) $k = 2$ **Answer:**A
(iii) $k = 7$ **Answer:**B

A) Red
B) Green
C) Blue

3. **[2 points]** Which of the following statements is true about the k-Nearest Neighbor classifier? Select all that apply.

   A) As the value of k increases, the variance of the model increases

   B) As the value of k increases, the bias of the model increases

   C) As the value of k increases, the model complexity increases

   D) As the value of k increases, the number of parameters in the model increases

   E) As the number of training data points increases, the memory requirements of the model increase

   Answer:B, E

   - For A: the variance of the model will decrease
   - For C: the model complexity will decrease
   - For D: the number of parameters in the model will not change

4. **[2 points]** What is the maximum training error for a decision tree on an arbitrary data set with k discrete output classes?

   A) $\frac{1}{k}$

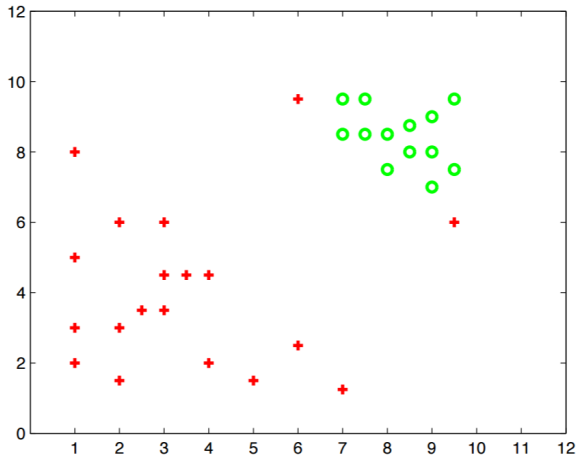
   B) 1

   C) $\frac{1}{\log_2 k}$

   D) $\frac{k-1}{k}$

   Answer:D
   This may happen when $P(Y) = P(Y|X_i)$, $\forall i$, and all data will be in a leaf node with a class label whose prior $P(Y)$ is the highest. To maximize the training error, the minimum training accuracy will be $1/k$; therefore, the maximum training error is $(k-1)/k$

5. **[2 points]** Consider the following data set. Which methods will classify all data points in this set correctly? Select all that apply.



A) Soft-Margin SVM (with no kernel)
B) SVM with a quadratic kernel, when the coefficient on the penalty for slack variables $C = 0$
C) SVM with a quadratic kernel, when the coefficient on the penalty for slack variables $C \to \infty$
D) Logistic regression (no kernel)
E) 3-NN

Answer:B, C

- For A: The data is not linear separable. Two outlier red points will be misclassified.
- For B: Because $C = 0$, the training data will be correctly classified with arbitrarily selected ξ_i
- For C: $C \to \infty$ means hard-margin SVM is used. The quadratic kernel is used, and the data is non-linear separable.
- For D: Logistic is linear classifier.
- For E: The two outlier red points will be misclassified.

Part B, Problem 1: Decision Tree (30 pts) (Hao and Yiting)

1.1 Build Your Own Decision Tree (14 pts)

The following is a small synthetic data set where we try to predict the usage of individual mobile phones based on their income, age, education, and marital status. In this section, you can assume that the decision tree is built using the ID3 algorithm, where each attribute is used only as an internal node.

Income	Age	Education	Martial Status	Usage
Low	Old	University	Married	Low
Medium	Young	College	Single	Medium
Low	Old	University	Married	Low
High	Young	University	Single	High
Low	Old	University	Married	Low
High	Young	College	Single	Medium
Medium	Young	College	Married	Medium
Medium	Old	High School	Single	Low
High	Old	University	Single	High
Low	Old	High School	Married	Low
Medium	Young	College	Married	Medium
Medium	Old	High School	Single	Low
High	Old	University	Single	High
Low	Old	High School	Married	Low
Medium	Young	College	Married	Medium

1. **[2 points]** What is the initial entropy of Usage?
   Answer:
   Usage:
   - Low: 7
   - Medium: 5
   - High: 3

   Therefore,
   $$H(Y) = -\frac{7}{15}\log_2\frac{7}{15} - \frac{5}{15}\log_2\frac{5}{15} - \frac{3}{15}\log_2\frac{3}{15} = 1.506$$

2. **[5 points]** Which attribute should be chosen at the root of the tree? Show your calculation for the information gains (IG) and explain your choice in a sentence.
   Answer:

   - Income:
     - Low: 5 (L: 5, M: 0, H: 0)
     $$P(Low)H(Y|Low) = -\frac{5}{15}(\frac{5}{5}\log_2\frac{5}{5} + \frac{0}{5}\log_2\frac{0}{5} + \frac{0}{5}\log_2\frac{0}{5}) = 0.000$$
     - Medium: 6 (L: 2, M: 4, H: 0)
     $$P(Medium)H(Y|Low) = -\frac{6}{15}(\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6} + \frac{0}{6}\log_2\frac{0}{5}) = 0.367$$
     - High: 4 (L: 0, M: 1, H: 3)
     $$P(Medium)H(Y|High) = -\frac{4}{15}(\frac{0}{4}\log_2\frac{0}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}) = 0.216$$

$$IG(Y; Income) = H(Y) - H(Y|Income) = 1.506 - 0.000 - 0.367 - 0.216 = 0.923$$

- Age:
  - Young: 6 (L: 0, M: 5, H: 1)

$$P(Young)H(Y|Young) = -\frac{6}{15}(\frac{0}{6}\log_2\frac{0}{6} + \frac{5}{6}\log_2\frac{5}{6} + \frac{1}{6}\log_2\frac{1}{6}) = 0.260$$

  - Old: 9 (L: 7, M: 0, H: 2)

$$P(Old)H(Y|Old) = -\frac{9}{15}(\frac{7}{9}\log_2\frac{7}{9} + \frac{0}{9}\log_2\frac{0}{9} + \frac{2}{9}\log_2\frac{2}{9}) = 0.459$$

$$IG(Y; Age) = H(Y) - H(Y|Age) = 1.506 - 0.260 - 0.459 = 0.787$$

- Education:
  - High School: 4 (L: 4, M: 0, H: 0)

$$P(HighSchool)H(Y|HighSchool) = -\frac{4}{15}(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4} + \frac{0}{4}\log_2\frac{0}{4}) = 0.000$$

  - College: 5 (L: 0, M: 5, H: 0)

$$P(College)H(Y|College) = -\frac{5}{15}(\frac{0}{5}\log_2\frac{0}{5} + \frac{5}{5}\log_2\frac{5}{5} + \frac{0}{5}\log_2\frac{0}{5}) = 0.000$$

  - University: 6 (L: 3, M: 0, H: 3)

$$P(University)H(Y|University) = -\frac{6}{15}(\frac{3}{6}\log_2\frac{3}{6} + \frac{0}{6}\log_2\frac{0}{6} + \frac{3}{6}\log_2\frac{3}{6}) = 0.400$$

$$IG(Y; Education) = H(Y) - H(Y|Education) = 1.506 - 0.000 - 0.000 - 0.400 = 1.106$$

- Martial Status:
  - Single: 7 (L: 2, M: 2, H: 3)

$$P(Single)H(Y|Single) = -\frac{7}{15}(\frac{2}{7}\log_2\frac{2}{7} + \frac{2}{7}\log_2\frac{2}{7} + \frac{3}{7}\log_2\frac{3}{7}) = 0.726$$

  - Married: 8 (L: 5, M: 3, H: 0)

$$P(Married)H(Y|Married) = -\frac{8}{15}(\frac{5}{8}\log_2\frac{5}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{0}{8}\log_2\frac{0}{8}) = 0.509$$

$$IG(Y; Martialtion) = H(Y) - H(Y|Martial) = 1.506 - 0.726 - 0.509 = 0.271$$

Because $IG(Y; Education)$ is the largest information gain, the "Education" attribute should be chosen at the root of the tree.

3. **[7 points]** Draw the full decision tree for the data.
Answer:

- Education = High School ⇒ Usage = Low
- Education = College ⇒: Usage = Medium
- Education = University
  - Martial Status = Married ⇒ Usage = Low
  - Martial Status = Single ⇒ Usage = High

1.2 Decision Tress Analysis (16 pts)

1. **[8 points]** Suppose X and Y are discrete variables. Let IG be the information gain and H be the entropy. Show that $IG(X;Y) = H(X,Y) - H(X|Y) - H(Y|X)$.
Answer:

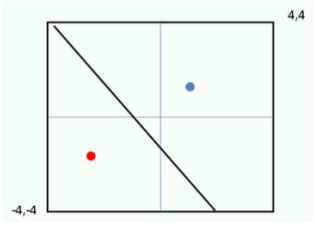
$$\begin{aligned} IG(X;Y) &= H(Y) - H(Y|X) \\ &= -\sum_y P(y)\log_2 P(y) + \sum_x P(x)\sum_y P(y|x)\log_2 P(y|x) \\ &= -\sum_{x,y} P(x,y)\log_2 P(y) - \sum_{x,y} P(x,y)\log_2 P(x|y) \\ &\quad + \sum_{x,y} P(x,y)\log_2 P(x|y) + \sum_{x,y} P(x,y)\log_2 P(y|x) \\ &= -\sum_{x,y} P(x,y)\log_2 P(x,y) \\ &\quad + \sum_y P(y)\sum_x P(x|y)\log_2 P(x|y) + \sum_x P(x)\sum_y P(y|x)\log_2 P(y|x) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

2. **[4 points]** Occam's Razor can be interpreted as simpler hypotheses are generally better than the complex ones. Does ID3 follow Occam's Razor? How about C4.5? Explain briefly (no more than 3 sentences).
Answer:
The ID3 and C4.5 typically don't follow Occam's Razor. The splitting of a decision tree is based on one feature space, but the decision boundary is normally not perpendicular to one feature space; therefore, the decision tree requires a complex combination of boundaries perpendicular to multiple feature spaces to form a simple decision boundary.

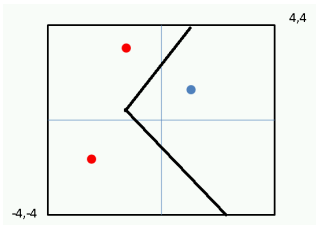
3. **[4 points]** Consider a decision tree T learned on a training set of n instances. Assume that there are two identical instances X and $X\prime$ (i.e. they have exactly the same attributes and labels) in the training set. Can removing $X\prime$ out of the training set produce a different T? Explain Briefly (no more than 4 sentences).
Answer:
Yes, the tree may change, because decision tree classfier requires the availability of all training data to build the tree. E.g. the mid-points of features in C4.5 may change even we just remove a duplicated point.

Part B, Problem 2 K-NN (15 pts) (Weixiang)

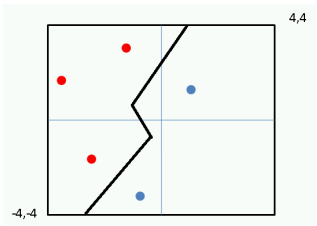
1. **(9 points)** For each of the following figures, we are given a few data points in 2-d space,each of which is labeled as either positive (blue) or negative (red). Assuming that we are using L2 distance as a distance metric, draw the decision boundary for 1-NN for each case. In other words, with your decision boundary, the new test data can be classified into corresponding categories. As an example, we draw the decision boundary for you with figure (a).



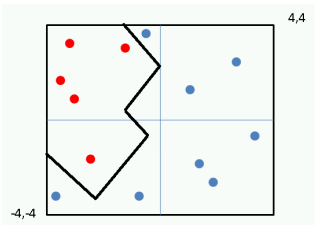
(a)



(b)



(c)



(d)

2. **(3 points)** In class we have mentioned that K-NN is a *lazy* classifier. Do we *always* need to store all training data to build our 1-NN classifier when we have $n(n \geq 3)$ points? Why? Is there a case that you must store all training data points (i.e., use all training data points when doing the classification)? Why?
   Answer:
   No, we don't need to store all data points for 1-NN classifier, because some points are far from the decision boundary, and the classification of the region around these points can be determined by the points near decision boundary. If all points are near the decision boundary, we must store all training data points.

3. **(3 point)** Decision tree classfier requires the availability of all training data to build the tree. Thus when a new training data point comes in, it might influence the structure of the decision tree. Does K-NN also suffer from this problem and why (explian in 1-2 sentences)?
   Answer:
   With the increment of K, the decision boundary relies on more and more training data and even the whole dataset. If this happens, K-NN also suffers from this problem.

Part B, Problem 3: Kernel SVM (15 pts) (Adams)

After Homework 2, you must be very familiar with SVM. In this question, we are considering its Kernel version:

$$\begin{aligned} \min_{w\in\mathbb{R}^d, b, \xi_i\in\mathbb{R}, i=1,2,\ldots,n} \quad & \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^{\mathsf{T}}\phi(x^{(i)}) + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n \\ & \xi_i \geq 0, \quad \forall i = 1, \ldots, n \end{aligned}$$

where $x^{(i)} \in \mathbb{R}^p, i = 1, 2, ..., n$ is the original training data coming along with the label $y^{(i)} \in \{-1, 1\}$, $C > 0$ is a constant and $\phi : \mathbb{R}^p \to \mathbb{R}^d$ is a mapping function that maps the original data to a new space. Generally speaking, $d > p$ (In fact, d can be $+\infty$). Please answer the following questions. Note that the questions with complexity could be answered with big-O notation.

1. **[6 points]** First write down the Lagrangian of the above problem and then derive the dual problem step by step.
   Answer:

$$J(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\gamma}) = \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - \xi_i - y^{(i)}(\vec{w}^T\phi(\vec{x}^{(i)}) + b)) - \sum_{i=1}^{n}\gamma_i\xi_i \quad (\alpha_i \geq 0,\ \gamma_i \geq 0)$$

   Because,

$$\begin{aligned} \tfrac{\partial J}{\partial \vec{w}}\big|_{\vec{w}^*} &= \vec{w}^* - \textstyle\sum_{i=1}^{n}\alpha_i y^{(i)}\phi(\vec{x}^{(i)}) = 0 \\ &\Rightarrow \vec{w}^* = \textstyle\sum_{i=1}^{n}\alpha_i y^{(i)}\phi(\vec{x}^{(i)}) \\ \tfrac{\partial J}{\partial b}\big|_{b^*} &= -\textstyle\sum_{i=1}^{n}\alpha_i y^{(i)} = 0 \\ &\Rightarrow \textstyle\sum_{i=1}^{n}\alpha_i y^{(i)} = 0 \\ \tfrac{\partial J}{\partial \xi_i}\Big|_{\xi_i^*} &= C - \alpha_i - \gamma_i = 0 \\ &\Rightarrow C = \alpha_i + \gamma_i \end{aligned}$$

   Therefore,

$$J(\vec{w}^*, b^*, \vec{\xi}^*, \vec{\alpha}, \vec{\gamma}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\phi^T(\vec{x}^{(i)})\phi(\vec{x}^{(j)}) + \sum_{i=1}^{n}\alpha_i$$

   Therefore, the dual form is:

$$\max_{\vec{\alpha},\vec{\gamma}} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\phi^T(\vec{x}^{(i)})\phi(\vec{x}^{(j)}) + \sum_{i=1}^{n}\alpha_i$$

   s.t.

$$\begin{aligned} \sum_{i=1}^{n}\alpha_i y^{(i)} &= 0 \\ 0 \leq \alpha_i &\leq C \end{aligned}$$

2. **[9 points, (3 for each subproblem)]** Suppose we have obtained the solution of the dual problem, which is denoted as α_i^* for $i = 1, 2, ..., n$. Please find out the corresponding primal solution w^* and b^* in the course slides. Given a test data point $z \in \mathbb{R}^p$,
   Answer:

$$\vec{w}^* = \sum_{i=1}^{n}\alpha_i^* y^{(i)}\phi(\vec{x}^{(i)})$$

$$b^* = y^{(k)} - \sum_{i=1}^{n} \alpha_i^* y^{(i)} \phi^T(\vec{x}^{(i)}) \phi(\vec{x}^{(k)}) \quad (for\ any\ k\ where\ C > \alpha_k > 0)$$

- What is the time complexity of making the classification decision if the mapping is given by

$$\phi(x) = (\underbrace{x_p^2, x_{p-1}^2, ..., x_1^2}_{p}, \underbrace{\sqrt{2}x_p x_{p-1}, ..., \sqrt{2}x_p x_1}_{p-1}, \underbrace{\sqrt{2}x_{p-1}x_{p-2}, ..., \sqrt{2}x_{p-1}x_1}_{p-2}, ..., \sqrt{2}x_2 x_1, \sqrt{2c}x_p, ..., \sqrt{2c}x_1, c)^\mathsf{T}$$

  with a constant $c > 0$ and we need to compute it from scratch?
  Answer:

  - For $\phi(\vec{x}^{(i)})$: the number of multiply operations is $n\frac{(p+2)(p+1)}{2}$
  - For $\vec{w}^*$: the number of multiply operations is $n\frac{(p+2)(p+1)}{2}$
  - For b^*: the number of multiply operations is $2n\frac{(p+2)(p+1)}{2}$
  - For $\phi(\vec{z})$: the number of multiply operations is $\frac{(p+2)(p+1)}{2}$
  - For $sgn(\vec{w}^* \cdot \phi(\vec{z}) + b)$: the number of multiply operations is $\frac{(p+2)(p+1)}{2}$
  - Therefore, the time complexity is $O(np^2)$

- Let $K(u, v) = \phi(u)^\mathsf{T}\phi(v)$ where $\phi(\cdot)$ has the same definition as above. Please give a compact form of $K(u, v)$. What is the time complexity of making the classification decision if we directly compute $K(\cdot, \cdot)$?
  Answer:

$$K(\vec{u}, \vec{v}) = \phi(\vec{u})^T \phi(\vec{v}) = (\vec{u}^T \vec{v} + C)^2$$

  - For $\vec{w}^* \cdot \phi(\vec{z}) = \sum_{i=1}^{n} \alpha_i y_i K(\vec{z}, \vec{x}_i)$: the number of multiply operations is $n(p + 2)$
  - For $b^* = y_k - \sum_{i=1}^{n} \alpha_i y_i K(\vec{x}_k, \vec{z})$: the number of multiply operations is $n(p + 2)$
  - For $sgn(\vec{w}^* \cdot \phi(\vec{z}) + b)$: the number of multiply operations is 0
  - Therefore, the time complexity is $O(np)$

- Now please go back to the dual formulation you derived previously. To avoid repetitive computation, one can precompute all the inner products $K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^\mathsf{T}\phi(x^{(j)})$ before solving the dual problem. What is the space complexity of this approach and what might be a problem if n is huge?
  Answer:
  The space complexity of $K(x^{(i)}, x^{(j)})$ is $O(n^2)$, and this may lead memory overflow if n is huge.

Part C: Implementing SVM Variation [20 points] (Yichong and Prakhar)

Please attach your code as appendix to this problem.
In the last homework, we derived a variant of SVM that explicitly maximizes the margin. You can use any library on quadratic optimization (e.g., CVXOPT for python or quadprog for MATLAB) for this problem. Here is an instruction on CVXOPT.

Installation instructions of CVXOPT can be found here. CVXOPT provides an easy interface for quadratic programming: The function qp(P, q[, G, h[, A, b]]) solves the optimization problem

$$\begin{aligned} \text{minimize}_{x\in\mathbb{R}^n} \ & (1/2)x^T Px + q^T x \\ \text{subject to } & Gx \preceq h \\ & Ax = b \end{aligned}$$

for $P \in \mathbb{R}^{n\times n}, q \in \mathbb{R}^n, G \in \mathbb{R}^{m_1\times n}, h \in \mathbb{R}^{m_1}, A \in \mathbb{R}^{m_2\times n}, b \in \mathbb{R}^{m_2}$. Here $\preceq$ means pointwise less than or equal; i.e., if $a \preceq b$ for $a, b \in \mathbb{R}^n$, then $a_i \leq b_i \forall i = 1, 2, ..., n$, where the subscript indicates coordinates. Look into here for a concrete example including all details.

In last homework we want to solve the following primal problem:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\xi},\rho} \quad & \frac{1}{2}||\boldsymbol{w}||_2^2 + \frac{1}{2}b^2 - \rho + \frac{\lambda}{2}\sum_{i=1}^{n}\xi_i^2 \\ \text{subject to} \quad & y_i(w^T\boldsymbol{x_i} + b) \geq \rho - \xi_i, \quad i = 1, \ldots, n \end{aligned} \tag{1}$$

We derived a dual program of (1) in homework 2. (Have a look at the solutions if you did not solve it - we will release it soon.) Implement the dual program using CVXOPT. Compute results for both $\lambda = 1$ and $\lambda = 10$. The data is a toy dataset *train_data.csv* in csv format of size $\mathbb{R}^{100\times 2}$, and the training label is contained in the file *train_label.csv* of size $\mathbb{R}^{100\times 1}$. Be careful that CVXOPT uses a slightly different matrix format than numpy; so either create your matrix in CVXOPT format, or use a numpy array and convert it into CVXOPT format using A = cvxopt.matrix(A).
(a) [**4 points**] Suppose you use $\boldsymbol{\alpha}$ as the Lagrange multipliers in dual program. Given the dual solution $\boldsymbol{\alpha}$, compute the primal solution $\boldsymbol{w}, \boldsymbol{\xi}, \rho$ in terms of training data, λ and $\boldsymbol{\alpha}$. (Hint: This has been computed in last homework, and you do not need to redo them.)
(b) [**8 points**] For each value of λ, draw a scatter plot of the data and plot the decision border (where the predicted class label changes) as well as the boundaries of the margin (the area in which there is a nonzero penalty for predicting any label). Use different colors for margins and border. Also use different colors for positive ($y = 1$) and negative ($y = -1$) samples.
(c) [**4 points**] Report the test error on the test set *test_data.csv* and *test_label.csv* for each value of λ.
(d) [**4 points**] What is the difference between the result obtained in $\lambda = 1$ and $\lambda = 10$? Why is that?