Honor code. This assignment is individual work. The goal of this assignment is for you to put in practice the concepts we learned in the video recordings, as well as explore complementary concepts. As mentioned in the synchronous lecture, academic integrity will be strictly enforced. If for any reason you are tempted to cheat (i.e., because you are facing personal hardship), contact the instructor immediately by email.

Instructions. To facilitate grading, please follow the following guidelines:

- You should submit a single PDF with all of your answers in the same order than this handout.
- You can use any tool you want to generate the PDF (e.g., word, LaTeX, scan your handwriting), but each page of the PDF should be easy to read and oriented properly.
- If you decide to handwrite your exam rather than typeset it, ensure your handwriting is readable otherwise TAs will have the discretion to not grade your answer.
- Clearly state any questions that you skip by writing down the question number along with "I skip this question"
- Graphs produced should be clearly interpretable. Include labels on axes and a legend.
- Attach a python script (.py) or an iPython notebook for the questions that require handing in code.
- Only your last Quercus submission will be graded. Make sure it contains a single PDF for your solutions and one .py/.ipynb script for your code,

Assignment structure. The assignment contains 21 questions worth a total of 32 points.

Problem 1 The goal of this problem is to derive the parameters of a support vector machine with a hard margin. We will consider the following toy dataset of 4 points: $x^{(1)} = (1,1)^\top$, $x^{(2)} = (-1,-1)^\top$, $x^{(3)} = (1,0)^\top$, and $x^{(4)} = (0,1)^\top$. $x^{(1)}$ is the only point from the positive class, whereas $x^{(2)}$, $x^{(3)}$, and $x^{(4)}$ are from the negative class. Put another way, $y^{(1)} = 1$ and $y^{(2)} = y^{(3)} = y^{(4)} = -1$.

1. (2 points) Plot the data and draw the maximum-margin hyperplane. By inspecting the plot, give the equation of this hyperplane. You will use it to verify your answers at the end of this problem.
2. (1 point) Which of the points are support vectors?
3. (1 point) Recall the optimization problem we defined for the SVM with a hard margin:

$$\min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|^2 \texttt{ s.t. } \quad \forall j(\vec{w} \cdot x^{(j)} + b)y^{(j)} \geq 1 \tag{1}$$

Write down the Lagrangian $L(\vec{w}, b, \vec{\alpha})$ corresponding to this problem, where $\vec{\alpha}$ is the vector of dual variables.

4. (2 points) We can therefore rewrite Equation 1 as follows:

$$\min_{\vec{w},b} \max_{\vec{\alpha}} L(\vec{w}, b, \vec{\alpha})$$

Because Slater's condition holds, this is equivalent to solving:

$$\max_{\vec{\alpha}} \min_{\vec{w},b} L(\vec{w}, b, \vec{\alpha}) \tag{2}$$

Given a fixed $\vec{\alpha}$, solve the inner (minimization) problem. You should obtain two relationships taking the form of

$$A = \sum_j B_j C_j \vec{D}_j$$

$$\sum_j E_j F_j = G$$

5. (2 points) Simplify Equation 2 with the result of question 4. Show that it takes the form of

$$\max_{\vec{\alpha}} \sum_j A_j - \frac{1}{2} \sum_j \sum_k B_j C_k D_j E_k \vec{F} \cdot \vec{G}$$

6. (1 point) We now come back to our toy dataset. Because it has 4 trainings points, we have $j \in 1..4$, i.e., four dual variables. For points that are not support vectors, recall from class that the corresponding constraints in Equation 1 can be removed. Deduce the value of one of the dual variable α_j corresponding to one value of j (i.e., one constraint) which we will write j^* to not reveal the answer. Follow indexes introduced in the problem statement above when describing the dataset.

7. (1 point) Deduce from the second expression found in question 4 (i.e., $\sum_j E_j F_j = G$), a relationship between all remaining dual variables α_j for $j \neq j^*$

8. (3 points) Solving the optimization problem from question 5, and using question 7 in addition, find values for all dual variables α_j for $j \in 1..4$.

9. (1 point) Deduce the numerical values of the two components of $\vec{w}$.

10. (2 points) Through an analysis of constraints in Equation 1 which are tight, deduce the numerical value of b. Hint: recall the role of support vectors here.

11. (0 points) Check that the hyperplane you found analytically is the same than the one you found geometrically in the first question.

Problem 2 - Binary linear classification vs. SVM For this problem, we will empirically compare the performance of a binary linear classifier to a SVM. Because the purpose of this problem is to compare the performance of the two classifiers, you can use their implementation in sklearn. However, note that (a) you should understand how to implement these classifiers from scratch and (b) you may be evaluated on this later in the course.

Consider the iris dataset included in sklearn. For all questions of this problem (except the last question), we will only consider the **first 100 entries** of the dataset. With the help of train_test_split from sklearn.model_selection, split the dataset into a training set and a test set. For now, you can do so by setting the argument test_size to 0.8 in train_test_split. To ensure results below are comparable and reproducible, set the random state argument of your train_test_split calls to 0: this will control the shuffling applied to the data before applying the split.

1. (2 points) Implement a binary linear classifier on the first two dimensions (sepal length and width) of the iris dataset and plot its decision boundary. (Hint: sklearn refers to the binary linear classifier as a LogisticRegression, we will see why later in the course.)

2. (1 point) Report the accuracy of your binary linear classifier on both the training and test sets.

3. (2 points) Implement a linear SVM classifier on the first two dimensions (sepal length and width). Plot the decision boundary of the classifier and its margins.

4. (1 point) Circle the support vectors. Please justify how to identify them through the duality theorem. (hint: KKT condition)

5. (1 point) Report the accuracy of your linear SVM classifier on both the training and test sets.

6. (1 point) What is the value of the margin? Justify your answer.

7. (1 point) Which vector is orthogonal to the decision boundary?

8. (3 points) Split the iris dataset again in a training and test set, this time setting test_size to 0.4 when calling train_test_split. Train the SVM classifier again. Does the decision boundary change? How about the test accuracy? Please justify why (hint: think about the support vectors), and illustrate your argument with a new plot.

9. (1 point) Do the binary linear classifier and SVM have the same decision boundaries?

10. (3 points) Now consider all 150 entries in the iris dataset, and retrain the SVM. You should find that the data points are not linearly separable. How can you deal with it? Justify your answer and plot the decision boundary of your new proposed classifier.

*
* *