Elements of Machine Learning, WS 2024/2025
Prof. Dr. Isabel Valera and Dr. Kavya Gupta
Assignment Sheet #5: *Trees and SVMs*

UNIVERSITÄT DES SAARLANDES

Deadline: Wednesday, January 22nd, 2025 23:59 hrs

This problem set is worth a total of 55 points, consisting of 3 theory questions and 1 programming question. Please carefully follow the instructions below to ensure a valid submission:

- You are encouraged to work in groups of two students. Register your team (of 1 or 2 members) on the CMS at least ONE week before the submission deadline. You have to register your team for each assignment.
- All solutions, including coding answers, must be uploaded individually to the CMS under the corresponding assignment and problem number. On CMS you will find FOUR problems under each assignment. Make sure you upload correctly each of your solution against $Assignment\#X - Problem\ Y$(where X- Assignment number and Y is the problem number) on CMS. In total you have to upload THREE PDFs (theoretical problems) and ONE ZIP file (programming problem).
- For each **theoretical question**, we encourage using LaTeX or Word to write your solutions for clarity and readability. Scanned handwritten solutions will be accepted as long as they are clean and easily legible. Final submission format must always be in a single PDF file per theoretical problem. Ensure your name, team member's name (if applicable), and matriculation numbers are clearly listed at the top of each PDF.
- For **programming question**, you need to upload a ZIP file to CMS under $Assignment\#X - Problem$ 4. Each ZIP file must contain a PDF or HTML exported from Jupyter Notebook and the .ipynb file with solutions. Make sure all cells in your Jupyter notebook contain your final answers. For creating PDF/HTML, use the export of the Jupyter notebook. Before exporting, ensure that all cells have been computed. To do this:
  - Go to the "Cell" menu at the top of the Jupyter interface.
  - Select "Run All" to execute every cell in your notebook.
  - Once all cells are executed, export the notebook: Click on "File" in the top menu.
  - Choose "Export As" and select either PDF or HTML.

  The submission should include your name, team member's name, and matriculation numbers at the top of both PDF/HTML and .ipynb file document.
- Finally, ensure academic integrity is maintained. Cite any external resources you use for your assignment.
- If you have any questions follow the instructions here.

Problem 1 (Decision Trees). (15 Points)

1. Given the dataset D of 11 points, shown in Table 1: (5 Points)

Index	X_1	X_2	Y
1	4	2	0
2	3	8	1
3	3	1	0
4	6	3	1
5	5	1	0
6	1	6	1
7	2	7	1
8	2	3	0
9	4	7	1
10	5	8	1
11	1	2	0

Table 1: Dataset with 11 data points showing X_1 and X_2 values.

Using the split $X_2 = 4.5$:

   (a) Calculate accuracy of this model (Decision Stump).

   (b) Determine the probabilities of **both classes in both subspaces**.

   (c) If we change the split to the following values: (i) $X_2 = 3.5$, (ii) $X_2 = 5.5$, analyze and describe what changes occur in accuracy and class probabilities.

2. Figure 1 below shows the decision boundary of a fitted model. (3 points)

   (a) How can we determine that this decision boundary corresponds to a decision tree model?

   (b) What issue do you observe in this image with respect to the model's behavior/decision boundary?

   (c) Suggest one modification or technique to address the identified problem.

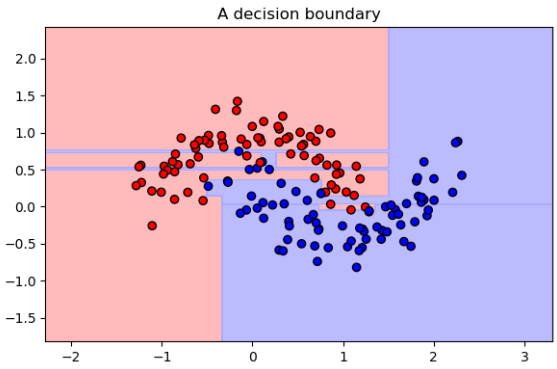

Figure 1: A decision boundary in a binary classification problem.

3. Given a data point $Z = (X_1, X_2) = (1, -2.75)$, (7 points)

   (a) Determine which class it belongs to according to the decision tree shown in Figure 2.

   (b) Write down the sequence of nodes visited during the traversal of the tree.

   (c) Plot the decision boundary between the two classes, shown by the tree. There is no need to plot each individual subspace.

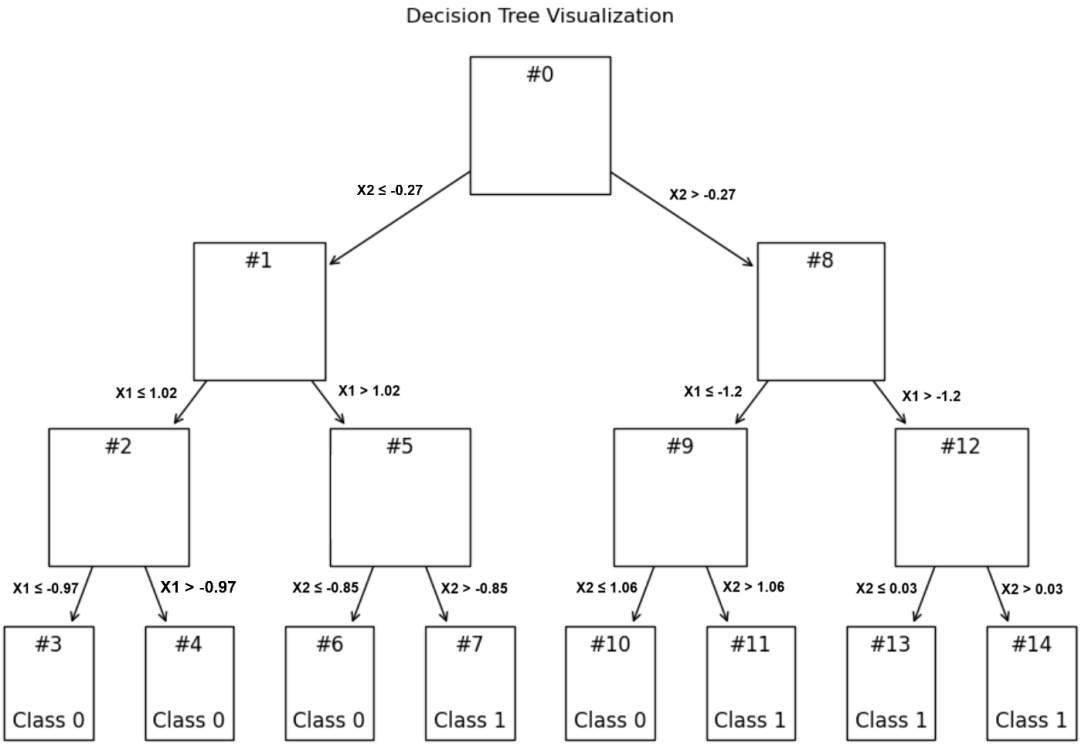

Figure 2: A partially hidden plot of a decision tree.

(d) After drawing the boundaries, clearly label which class each region belongs to.

Problem 2 (Ensemble). (13 Points)

1. If you have trained five different models, such as SVMs, decision trees, nearest neighbor classifiers, etc., on the same training data, and all achieve 95% precision, is it possible to combine these models to improve performance? If so, how can this be achieved? If not, what is the reason? (4 Points)
2. Is it possible to speed up the training of a bagging ensemble by distributing it across multiple servers? How about the boosting ensemble methods we discussed in class, and random forests? Explain how for each of the method. (3 Points)
3. In Figure 3, the left column displays individual models trained, while the right column shows the cumulative predictions of the ensemble. Based on this, which ensemble approach and base learner do you think are used, and why? (6 Points)

Problem 3 (Support Vector Machines). (12 Points)

1. In an online learning scenario, a model is updated incrementally as new data points arrive. In this case, we have a deployed Support Vector Machine (SVM) model, and our system continuously logs new data points. When would we choose **not** to update the model after receiving a new data point in a:

   (a) SVM Regression scenario,

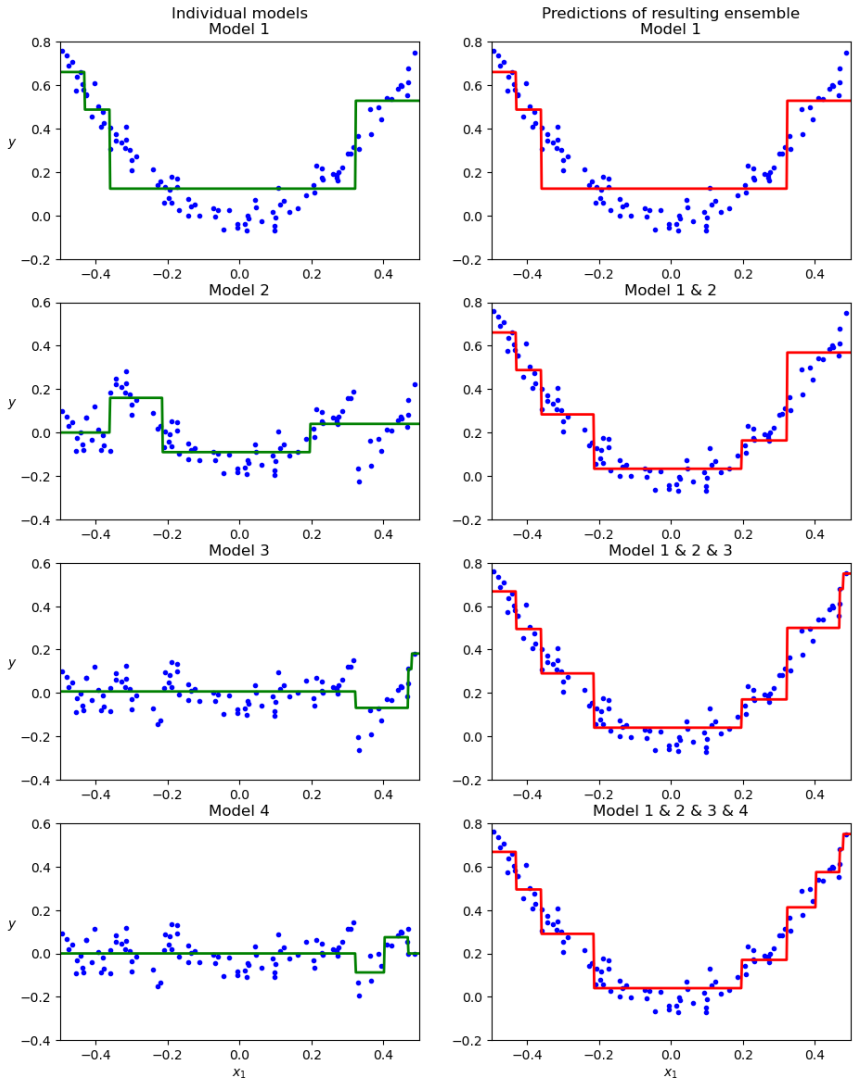

Figure 3: The green lines represent the individual predictors, while the red line represents the ensemble.

(b) SVM Classification scenario,

Answer in terms of the margins of the SVM model. (2 Points)

2. In Figure 4 which of these models is likely to correspond to a (hard margin) SVM classifier, and why? (2 Points)

3. Given a dataset and a hard margin SVM model, identify the support vectors in the plot seen in Figure 5. How can you visually distinguish the support vectors from other data points? (2 Points)

4. Which observations could we (slightly) move without affecting the maximal margin hyperplane shown in Figure 5. Explain why. (2 Points)

5. You are provided with a dataset (Figure 6) that you have plotted. Based on the characteristics of the data distribution in the plot, should you use a hard margin or soft margin Support Vector Machine (SVM) to classify the data? Explain your reasoning. (4 Points)

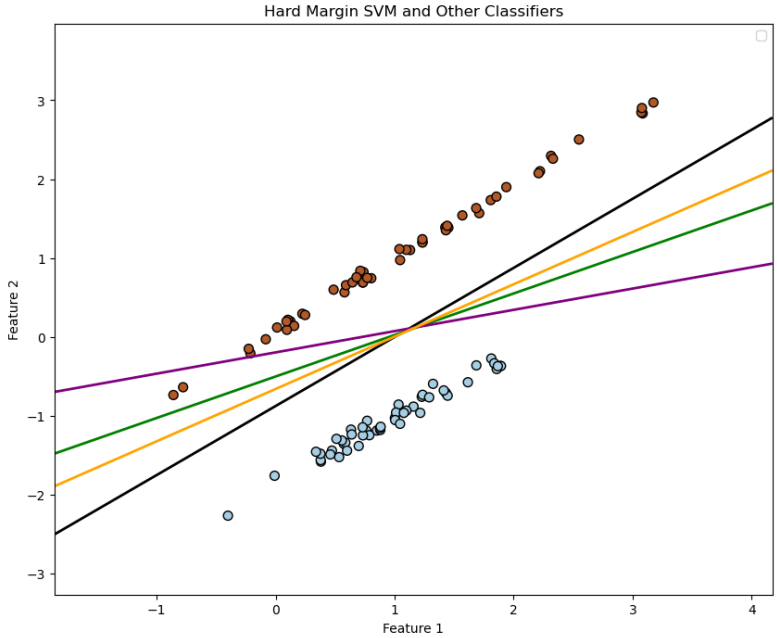

Figure 4: Decision Boundaries of Multiple Fitted Classifiers

Problem 4 (Trees and SVMs). (10 (+ 5 bonus) Points)

In this assignment, you will learn how to train Decision Trees, Support Vector Machines as well as revise previous assignments.

Please refer to the file `assignment_5_handout.ipynb` and **only** complete the sections marked in red and missing codes denoted with `#TODO`. Once you have filled in the required parts, revisit the submission instructions to check how to submit it.

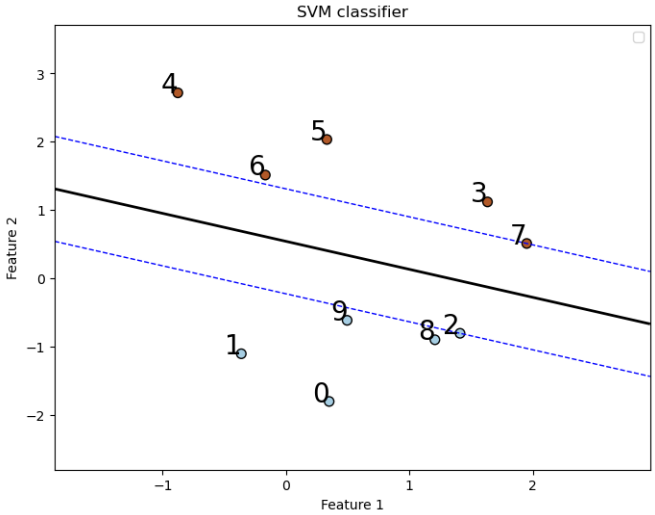

Figure 5: Decision Boundaries of Multiple Fitted Classifiers

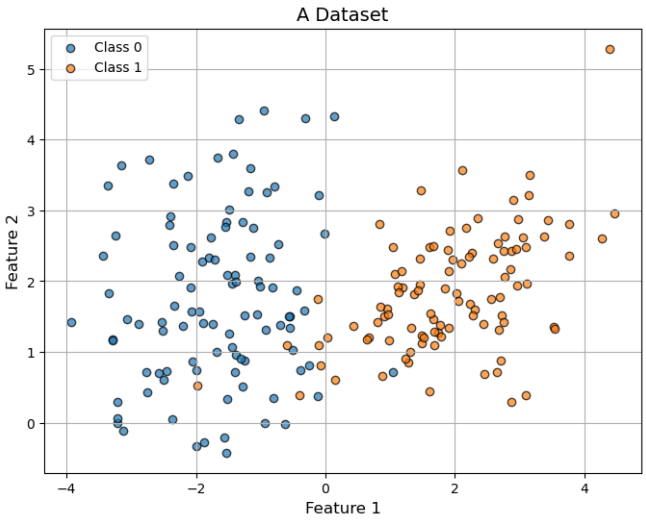

Figure 6: A random dataset.