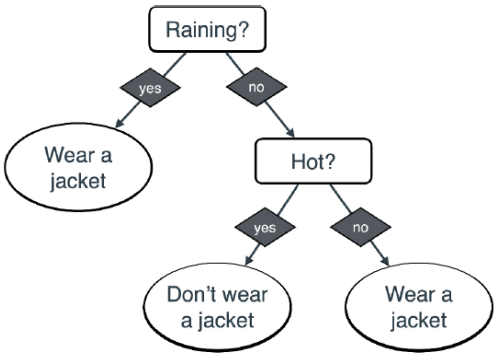
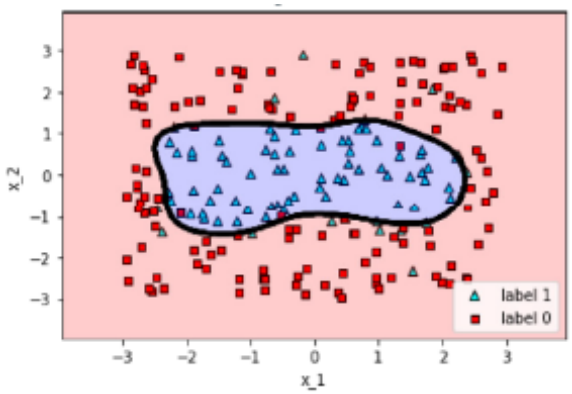
COMP-2704:
Supervised Machine Learning





Final Project: Decision Trees vs. Support Vector Machines

Overview:

In this project you will develop machine learning algorithms for classification. You will create two notebooks, one for a decision tree and another for a support vector machine. Both algorithms are to use the same data and attempt to solve the same problem.

You are to explore various options and do your best to optimize each notebook, using comments to add/remove lines of code as you work. Keep a record of everything you try, along with the results, by using comments and markdown text within the notebook. The code lines you keep should be the options that work best.

In the end, you are to recommend either the decision tree or support vector machines as the optimal solution, based on a quantitative assessment of the performance of the two notebooks.

Details about what is to be done are given below, and a rubric is provided on the course website which lists the marks for each step. After submission, you will have a meeting with your instructor to explain what is done in your notebooks and answer questions.

Steps to Complete

Use Case and Data Selection

This step involves choosing your project topic and data; it must be completed and approved by the instructor before continuing. A due data for Use Case and Data Selection will be posted by the instructor.

Think about what kind of data you are interested in working with – whether it be something of personal interest to you or something that might benefit you in a future job application. Imagine a realistic use case for this related to software or business intelligence. Create a notebook satisfying each requirement and upload it to the Final Project dropbox.

You can work with data in CSV format, or with text data, but image data will not be suitable for this project. A few good places to look for data:

- An excellent online resource for machine learning: www.kaggle.com/datasets/
- The University of California, Irvine collection: Home - UCI Machine Learning Repository
- The Canadian Federated Research Data Repository: Home | FRDR-DFDR
- The NLTK package contains text data from publications. See: NLTK :: Installing NLTK Data.
- The Twitter API: Twitter API Documentation | Docs | Twitter Developer Platform, for example.
- You may even obtain data from a database or web page using methods learned in COMP-2040.

Requirements

1) *[3 marks]* Realistic use case for data with at least three classes, written in markdown. Provide a description of each feature in the data set.
2) *[1 mark]* Discussion of prediction errors. Are false positives or negatives for one of the classes to be avoided? Or are all errors equally bad?
3) *[2 marks]* Data cleaning, with check that it is clean using Pandas. Save the data as a CSV file for later use.
4) *[1 mark]* Statistical or categorical description using the Pandas describe function. Provide written observations.
5) *[1 mark]* Bar graph showing the count of each class. Provide written observations.

Model Development

After getting approval from the instructor for your use case and data, proceed with model development.

Decision tree

Develop decision trees using the sci-kit learn module within a single notebook.

1) *[1 mark]* Split the data into training, validation, and testing sets. Try different proportions and justify the final choices.
2) *[3 marks]* Try both Gini impurity index and Entropy as a condition for splitting branches. Experiment with different values of *max_depth, max_features, min_impurity_decrease, min_samples_leaf, min_samples_split*, and any other hyperparameters you wish. Keep your best two (or more) models.
3) *[2 marks]* On your best decision tree models, use the *classification_report* and *confusion_*matrix functions in sklearn to display metrics, using training and validation (but not testing) data. Also use the *display_tree* method in *utils.py* to display each tree.
4) *[2 marks]* Select and justify your final choice of hyperparameters based on the training and validation metrics. Provide a written analysis in markdown.

Support Vector Machine

Develop support vector machines (SVMs) using the sci-kit learn module within a single notebook.

1) *[1 mark]* Split the data into training, validation, and testing sets. Try different proportions and justify the final choices.
2) *[3 marks]* Experiment with different values of the *C* parameter; try the linear, rbf (with different choices of gamma) and polynomial kernels (with different degrees); try both options for *decision_function_shape.* Keep your best two models.
3) *[2 marks]* Use the *classification_report* and *confusion_*matrix functions in sklearn to display metrics for your best models, using training and validation (but not testing) data.
4) *[2 marks]* Select and justify your final choice of hyperparameters based on the training and validation metrics. Provide a written analysis in markdown.

Comparison

Choose between the decision tree and SVM model. Write this section at the bottom of the chosen notebook.

1) *[2 marks]* Provide a written comparison of the training and validation metrics for the best SVM and decision tree models using markdown. Select the best model and justify your choice.
2) *[2 marks]* Use your selected model to make predictions on the test set. Use the *classification_report* and *confusion_*matrix functions in sklearn to display metrics for the test data.
3) *[2 marks]* Use markdown to review the best model; restate important metrics nd describe how well it will work for your use case.

Submission

Submit the "Use Case and Data Selection" part of your project at the earlier due date posted by the instructor. Save the notebook with output displayed and upload it ot the Final Project dropbox. You do not need to upload your data.

Submit the "Model Development" part of your project by the final due date. Save all notebooks with output displayed and upload them to the Final Project dropbox. You do not need to upload your data. Late submissions will lose 10%.

Total marks = 30