Using Machine Learning Tools 2020
Assignment 2: Breast Cancer Classification

Overview

In this assignment, you will apply two classifier types, Decision Trees and Support Vector Machines, to the problem of classifying breast cancer from a set of characteristics of the cell nuclei in an image of a fine needle aspirate of a breast mass. The Wisconsin Breast Cancer data set is available from the collection of example data sets in scikit learn.

The main aims of the assignment are:

- To use and compare two different classifier approaches on the same data set;
- To evaluate the classifiers and their structure in a white box fashion;
- To practice using pipelines and hyper parameter optimisation;
- To explore a multi-dimensional feature space and handle multi-dimensional data.

This assignment relates to the following ACS CBOK areas: abstraction, design, hardware and software, data and information, HCI and programming.

Instructions

While you are free to use whatever IDE you like to develop your code, your submission should be formatted as a Jupyter notebook that interleaves Python code with output, commentary and analysis.

- Your code must use the current stable versions of python libraries, not outdated versions!
- All data processing must be done within the notebook after calling the load function.
- Comment your code, so that its purpose is clear to the reader!
- **Before submitting your notebook, make sure to run all cells in your final notebook so that it works correctly!**

This assignment is divided into several tasks. Use the template notebook, which uses **the same numbering as in this document** (e.g. "3.2 Display decision tree"), and enter your code, results and answer text analysis under the exact number that it belongs to!

Make sure to answer every question with **separate answer text ("Answer: …") in a Markdown cell** and check that you answered all sub-questions/aspects within the question. The text answers are worth points!

Make the **figures self-explanatory and unambiguous**. Always include axis labels, if available with units, unique colours and markers for each curve/type of data, a legend and a title. Give every figure a number (e.g. at start of title), so that it can be referred to from different parts of the text/notebook. This is also worth points!

1 Understand the dataset (15%)

Load the Wisconsin breast cancer data set from the scikit-learn sample data collection. Read the full documentation: https://scikit-learn.org/stable/datasets/index.html#breast-cancer-dataset. Figure 1 below shows an example image of a tissue sample obtained by fine needle aspiration. A more detailed description of the parameters can be found here:
https://minds.wisconsin.edu/bitstream/1793/59692/1/TR1131.pdf

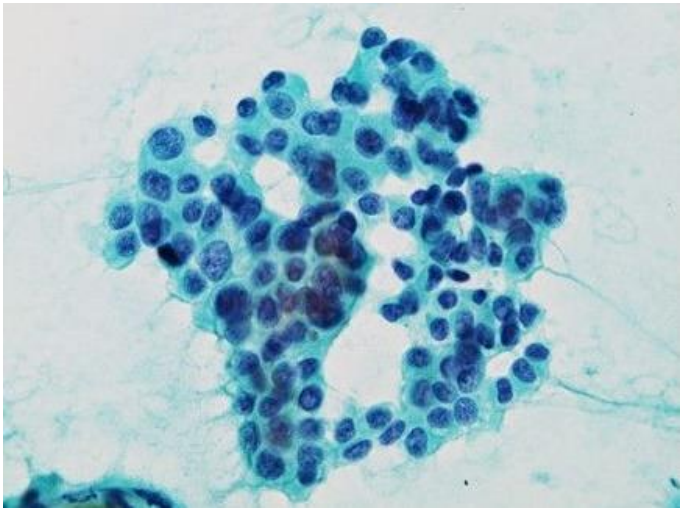

Figure 1. Example of a fine needle aspiration tissue image showing cells and cell nuclei (dark) (Shigematsu et al. 2011, Creative Commons 2.0 License)

1.1 Question: Briefly describe what each of the 10 parameters of the cell nuclei mean, using the documentation of the dataset and the example image in Figure 1. What could be the reasons for using the mean, standard error and maximum of each of the 10 parameters?

1.2 Plot histograms of each of the 30 features, using two distributions, one for each class, in each diagram. Use 3 figures with 10 subplots each.

1.3 Plot receiver-operating-characteristic (ROC) curves of the individual features into 3 figures, one figure for each of the groups of 10.

1.4 Question: Which of the parameters seems promising based on the histograms and ROC curves? Justify your choice while referring to the particular features in the figures that indicate a good separation. Choose your top five candidate features.

1.5 Analysis Point: Calculate the mean of all instances of the malignant class (centre of mass in high dimensional feature space) and the mean of all instances of the benign class. Save the mean between those two as the "Analysis Point". It is a point in the feature space that is approximately between both classes.

2 Train a decision tree classifier (15%)

2.1 Construct a decision tree classifier using the gini criterion and random_state=0. Below, you will perform a hyper parameter search of max_depth and min_samples_leaf. Check the following remaining parameters of the classifier and either keep the default value or select a different value: min_samples_split, min_weight_fraction_leaf, max_features, max_leaf_nodes, min_impurity_decrease, min_impurity_split and class_weight. Question: Describe each choice briefly in one sentence.

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

2.2 Build a pipeline including any pre-processing steps that you think are necessary. Question: Do the data need to be scaled for decision tree classification? Are the different class sizes a problem, and if so what are you doing about it?

2.3 Perform a grid search using five-fold cross validation over values of the maximum depth (max_depth) and the minimum number of samples per leaf (min_samples_leaf). Choose the value range yourself. Question: What is the rationale for your choice?

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV

3 Evaluate the decision tree classifier (20%)

3.1 Calculate the confusion matrix, precision and recall of the final classifier. Question: Based on these metrics, what is the chance of failing to detect a sample with cancer? What are the strengths and weaknesses of the classifier?

3.2 Display decision tree using plot_tree(). Question: Describe the structure. What do each of the entries in the first node mean? Are the features in the decision tree matching the initial candidate features from Section 1?

https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot_tree.html

3.3 Display the decision boundaries (use function predict()) together with a scatter plot of the data using two features at a time.

- Select the five most important features from the decision tree attribute "clf.feature_importances_". Make a figure with 4x4 subplots, where each row and column is one of the features and the subplots are the 2D displays using the corresponding features.
- For the decision boundary, use code that samples the prediction value in the plane of the respective two features (see example below) at the Analysis Point. This means that all other feature dimensions stay constant at the values of the Analysis Point. Use a margin of 30% of each feature value range (max-min) around the data cloud. Use the contourf() function like in the example and its parameters "levels" and "colors". It is highly recommended to program this display as a function for later reuse. Example: https://scikit-learn.org/stable/auto_examples/svm/plot_iris_svc.html#sphx-glr-auto-examples-svm-plot-iris-svc-py
- Plot the Analysis Point with a blue "+" marker in each diagram.

3.4 Question: Is the class differentiation well characterised by the node thresholds or is it modelling the boundary using a rigid or stair case pattern? Why are there few 2D scatterplots with only one class shown as prediction contour?

4 Train a support vector classifier with RBF kernel (15%)

4.1 Construct a support vector classifier with a radial basis function kernel. Below, you will perform a hyper parameter search of C and gamma. Check the following remaining parameters of the classifier and either keep the default value or select a different value: tol, class_weight and max_iter. Question: Describe each choice briefly in one sentence.

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

4.2 Build a pipeline including any pre-processing steps that you think are necessary. Question: Do the data need to be scaled for support vector classification? Are the different class sizes a problem, and if so what are you doing about it?

4.3 Perform a grid search using five-fold cross validation over values of the regularisation parameter C and the kernel coefficient gamma. Choose the value ranges yourself. Question: What is the rationale for your choice?

5 Evaluate the support vector classifier (20%)

5.1 Calculate the confusion matrix, precision and recall of the final classifier. Question: Based on these metrics, what is the chance of failing to detect a sample with cancer? What are the strengths and weaknesses of this classifier?

5.2 Display the decision boundary (use function decision_function()) together with a scatter plot of the data using the same features and figure layout as in the decision tree display for direct comparability. This time, use a suitable colormap (parameter “cmap”) in the contourf() function. Mark the support vectors.

5.3 Question: What is the meaning of the support vectors? Where can we see their purpose in the diagrams?

6 Compare the classifiers and interpret (15%)

6.1 Question: Compare the classifier structures and decision boundaries of both classifiers. Point out similarities and differences. How do the classifiers compare outside the areas of dense sampling in the parameter space, e.g. towards the edges of the scatterplot (extrapolation)?

6.2 Question: Generalisability: Do you see sources of bias in the two classifiers? Are the models showing any signs of overfitting (variance error)?

6.3 Question: Table 1 from Street et al. (1993) below shows the accuracies of their classifiers for different numbers of features and different numbers of hyperplanes used. Compare the number of features (decision tree), selection of features and accuracy of your classifiers with this table. Is there only one good set of features, many different sets or is there a pattern of similar feature combinations?

		Number of Planes		
		1	2	3
Number of Features	1	SE Perimeter 71.8		
	2	SE Perimeter SE Smoothness 74.2	W Radius W Fractal Dim 79.8	
	3	SE Area SE Compactness SE Fractal Dim 75.0	M Area W Concave Pts W Fractal Dim 81.5	M Smoothness M Compactness M Fractal Dim 83.9
	4	M Radius M Area SE Concave Pts SE Fractal Dim 76.6	M Texture W Area W Concavity W Fractal Dim 86.3	M Texture M Compactness W Area W Fractal Dim 81.4

Table 1: Features and Testing Correctness for Prognosis Data
All subsets of k features were tested for training set separation. The subset which demonstrated the best separation was then tested using the leave-one-out approach, and the percent correctness is shown. (M = mean, SE = standard error, W = worst)

Assessment

Hand in your notebook as an .ipynb file via the MyUni page. Make sure your notebook includes your code and formatted (Markdown) text blocks explaining what you have done. Your mark will be based on both code correctness and the quality of your comments and analysis.

The assignment is worth 35% of your overall mark for the course. It is due on Monday May 11 at 11.59pm.

Stephan Lau
April 2020

References:

Shigematsu, H., Kadoya, T., Kobayashi, Y. *et al.* A case of HER-2-positive recurrent breast cancer showing a clinically complete response to trastuzumab-containing chemotherapy after primary treatment of triple-negative breast cancer. *World J Surg Onc* **9,** 146 (2011). https://doi.org/10.1186/1477-7819-9-146

Street, W.N., Wolberg, W.H., & Mangasarian, O.L. (1993). Nuclear feature extraction for breast tumor diagnosis. Electronic Imaging. https://doi.org/10.1117/12.148698, Available online: https://minds.wisconsin.edu/bitstream/1793/59692/1/TR1131.pdf