Workshop 4

Learning From Big Data

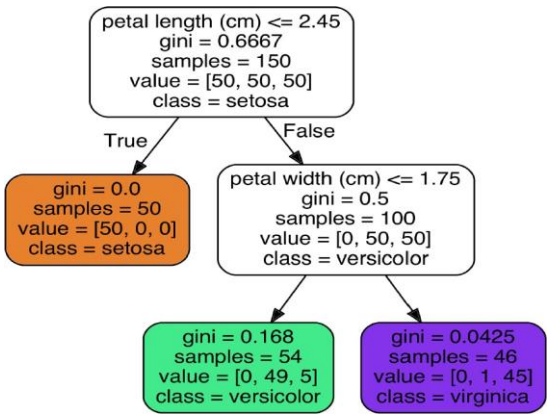
Decision Trees, Ensemble Learning and Random Forests

*Please do the exercises or tasks without "Optional" mark at first. After that, if you still have some time, please try the tasks with "Optional".

Part 1: Decision Trees

* We will use Scikit-Learn APIs to have a practice on Decision Trees at first.

1. Referring to a PDF file, called "Text for the explanation of decision trees, given in /Lectures/Session 4, together with my lecture slides, try to understand how to use a Decision-Tree for making predictions. In other words, you may well explain the diagram as below:



2. Review "**The CART Training Algorithm"** given in Section 4) of the slides and understand how this algorithm training or building a decision tree.

3. Study the notebook document, called decision_trees.ipynb, located in …/Labs/Lab4/Decision_trees, and understand:

   i) How to train and visualize decision trees;
   ii) How to predict classes and class probabilities;
   iii) The sensitivity of training set details.

4. **(Optional):** I did not give a detailed discussion about how to use Decision Tress for regression tasks in my lecture. However, once you have a good understanding of how to use Decision Trees for classification, you may study on "Regression Trees" by yourself.

- Please refer to a PDF file, called "Regression Trees", given in …/Labs/Lab4, together with the last part of decision_trees.ipynb notebook, to learn this topic.

Part 2: Ensemble Learning and Random Forests

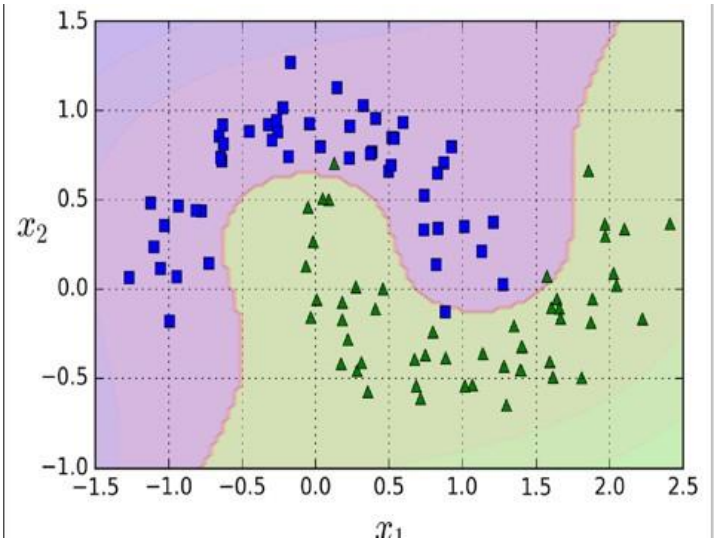
By using Scikit-Learn, let us do the following exercises dealing with ensemble learning and random forests.

Regarding ensemble learning, a number of methods are introduced in my lecture, such as voting classifiers, bagging and pasting, random forests, as well boosting (such as, Adaboosting and Gradient Boosting).

1. Before you study the following examples, please make sure of understanding these concepts by reviewing the lecture slides. Particularly, have a more attention to Random Forests and Gradient Boosting, as they are very popular in use today.

2. Study the coding samples - Ensemble_learning_and_random_forests.ipynb, located in …/Labs/Lab4, which will be helpful for you to understand ensemble learning techniques or methods. [I have added more comments in the notebook].

- For your understanding what **"moon dataset"** looks like, I put the following diagram as below. You will see the data distributed in two interleaving circles.



Part 3: Tree methods with Spark

Similar to Scikit-Learn, Spark also provides APIs with MLlib, for Decision Trees, Random Forests and Gradient Boosting Algorithms. In this part, we will have a practice with Spark around these topics.

Exercises:

1. Please visit the notebook, called Testing_Three_Tree Methods.ipynb, given in …/Labs/Lab4. In this example, we will code along with some data and test out 3 different tree methods:

   - A single decision tree
   - A random forest
   - A gradient boosted tree classifier

   *We will be using a college dataset to try **to classify** colleges as Private or Public based off these features:

   Private A factor with levels No and Yes indicating private or public university

   Apps Number of applications received

   Accept Number of applications accepted

   Enroll Number of new students enrolled

   Top10perc Pct. new students from top 10% of H.S. class

   Top25Perc Pct. new students from top 25% of H.S. class

   F.Undergrad Number of fulltime undergraduates

   P.Undergrad Number of parttime undergraduates

   Outstate Out-of-state tuition

   Room.Board Room and board costs

   Books Estimated book costs

   Personal Estimated personal spending

   PhD Pct. of faculty with Ph.D.'s

   Terminal Pct. of faculty with terminal degree

   S.F.Ratio Student/faculty ratio

   perc.alumni Pct. alumni who donate

   Expend Instructional expenditure per student

   Grad.Rate Graduation rate

   Action:

   Study this example and see how to use the **three tree methods** for this classification and what is the result of each method.