# COMP9417 - Machine Learning
# Homework 2: Perceptrons, Kernels & `scikit-learn`

**Introduction** In this homework, we will explore perceptron learning in more depth. We will then dive further into some existing sklearn modules and compare their perforamance on simulated data.

**Question 1. Diving Deeper into sklearn**

In this question, we will consider the scikitlearn implementations of the following classifiers:

- Decision Trees
- Random Forest
- AdaBoost
- Logistic Regression
- Multilayer Perceptron (Neural Network)
- Support Vector Machine

You are required to compare the performance of the above models on a classifiation task. The following code loads in these classifiers and defines a function to simulate a toy dataset:

```
        import numpy as np
        import matplotlib.pyplot as plt
        from matplotlib.colors import ListedColormap
        import warnings
        warnings.simplefilter(action='ignore', category=FutureWarning)

        import time
        from sklearn.svm import SVC
        from sklearn.linear_model import LogisticRegression
        from sklearn.ensemble import AdaBoostClassifier
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.tree import DecisionTreeClassifier
        from sklearn.neural_network import MLPClassifier
        from sklearn.model_selection import train_test_split
        from sklearn.preprocessing import StandardScaler
        from sklearn.datasets import make_classification

        def create_dataset(n=1250, nf=2, nr=0, ni=2, random_state=125):
            '''
            generate a new dataset with
            n: total number of samples
            nf: number of features
            nr: number of redundant features (these are linear combinatins of informative
    features)
            ni: number of informative features (ni + nr = nf must hold)
            random_state: set for reproducibility
```

```
        '''
        X, y = make_classification( n_samples=n,
                                    n_features=nf,
                                    n_redundant=nr,
                                    n_informative=ni,
                                    random_state=random_state,
                                    n_clusters_per_class=2)
        rng = np.random.RandomState(2)
        X += 3 * rng.uniform(size = X.shape)
        X = StandardScaler().fit_transform(X)
        return X, y


```

(a) Generate a dataset using the default parameters of `create_dataset`. Then, randomly split the dataset into training set Xtrain and test set Xtest (use the sklearn `train_test_split` function with `random_state=45`), with 80 % of examples for training and 20 % for testing. Fit each of the models to the training data and then plot the decision boundaries of each of the classifiers (using default parameter settings) on the test set. If you prefer, you may use the following plotter function which plots the decision boundary and works for any sklearn model.

```
def plotter(classifier, X, X_test, y_test, title, ax=None):
    # plot decision boundary for given classifier
    plot_step = 0.02
    x_min, x_max = X[:, 0].min() - 1, X[:,0].max() + 1
    y_min, y_max = X[:, 1].min() - 1, X[:,1].max() + 1
    xx, yy = np.meshgrid(np.arange(x_min, x_max, plot_step),
                         np.arange(y_min, y_max, plot_step))
    Z = classifier.predict(np.c_[xx.ravel(),yy.ravel()])
    Z = Z.reshape(xx.shape)
    if ax:
        ax.contourf(xx, yy, Z, cmap = plt.cm.Paired)
        ax.scatter(X_test[:, 0], X_test[:, 1], c = y_test)
        ax.set_title(title)
    else:
        plt.contourf(xx, yy, Z, cmap = plt.cm.Paired)
        plt.scatter(X_test[:, 0], X_test[:, 1], c = y_test)
        plt.title(title)

```

Note that you can use the same general approach for plotting a grid as used in Homework 1, and the plotter function supports an 'ax' argument. *What to submit: a single* $3 \times 2$ *plot, a print screen of the python code used, a copy of your python code in solutions.py.*

(b) Next, we will study how the performance of each of the clasifiers varies as you increase the size of the training set. Fix your training and test sets from part (a). Then, starting from a random subset (no need to set a seed) of your training set of size 50 (chosen with replacement), train your classification model, and compute the accuracy on the test set. Repeat this process for training set sizes of $[50, 100, 200, 300, \ldots, 1000]$. Repeat the experiment a total of 10 times for each classifier. Then, for each classifier, plot its average accuracy (achieved on the test set) for each training set size. Compare the accuracy across different algorithms in a **single** figure, and in 5 lines or less, discuss your observations: For the models covered in the course so far, use what you know about the bias-variance decomposition to inform your discussion. Which model you prefer for this task?

Please use the following color scheme for your plots: [Decision Tree, Random Forest, AdaBoost, Logistic Regression, Neural Network, SVM], and please include a legend in your plot. *What to*

*submit: a discussion of your observation, a single plot, a print screen of the python code used, a copy of your python code in solutions.py.*

(c) Using the time module, record the training time for each of the classifiers at each of the training set sizes. Plot the log of the average training time over the 10 trials of each classifier as a function of the training size (use log base $e$). You may add code for this section to your code in part (b). What do you observe? In 5 lines or less, discuss your observations. Use the same color scheme as in (b). *What to submit: a discussion of your observation, a single plot, a print screen of the python code used, a copy of your python code in solutions.py.*

(d) We now focus on the `DecisionTreeClassifier` module. Use `create_dataset` from the previous question to generate a new dataset consisting of 2000 samples, 20 total features, 8 of which are informative, and use a random state of 25. Next, generate equally sized train/test data, using a random state of 15. Fit a deciscion tree classifier (using default parameters except for setting `random_state` to 0 ) and report both its train and test accuracy. *What to submit: train accuracy of your model, test accuracy of your model, a screen shot of your code, a copy of your python code in solutions.py.*

(e) Fit multiple decision trees with `min_samples_leaf` varying over $k = 2, 3, \ldots, 130$, and with `random_state` set to 0 For each iteration, compute the AUC score on both your train and test datasets from the previous part. Provide a plot of both train and test AUC over $k$. *What to submit: a single plot, a screen shot of your code, a copy of your python code in solutions.py.*

(f) In this part, you will perform cross validation over the min_samples_leaf hyperparameter from scratch (**Do not use existing cross validation implementations in this question, doing so will result in a mark of zero.**) For `min_samples_leaf` ranging from $k = 2$ to $k = 95$, run 10-fold cross validation (i.e. split the training data into 10 folds, fit a decision tree with `min_samples_leaf` set to $k$ (and `random_state` set to 0 ) on 9 of those folds, and record the auc score on the 10-th, repeating this process 10 times). For this question, your 10 folds must be created by taking the first fold to be the first 100 rows, the second fold to be the next 100 rows, and so on.

Produce a plot displaying your results: the x-axis should be `min_samples_leaf`, and for each $k$ plot a Box-plot over the 10 CV scores. Further, report the value of $k$ that results in the highest CV score, re-fit the Decision tree (with `random_state` set to 0 ) with this choice and report both train and test accuracy. *What to submit: a single plot, train accuracy of your model, test accuracy of your model, a screen shot of your code, a copy of your python code in solutions.py.*

(g) Finally, we will use scikit-learn's built in `GridSearchCV` to repeat the analysis in the previous part. Report the optimal `min_samples_leaf` chosen by this process. Re-fit the Decision tree (with `random_state` set to 0 ) with this choice and report both train and test accuracy. Explain why the results might differ between your implementation and sklearn's. *What to submit: train accuracy of your model, test accuracy of your model, a screen shot of your code, a copy of your python code in solutions.py, some comments to address the results.*

**Question 2. (Perceptron Learning & Features)**

This question requires you to refer to the following training data for parts (a)-(c). You are only permitted to make use of numpy and matplotlib. You are not permitted to make use of any existing numpy implementations of perceptrons (if they exist).

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| -0.8 | 1 | 1 |
| 3.9 | 0.4 | -1 |
| 1.4 | 1 | 1 |
| 0.1 | -3.3 | -1 |
| 1.2 | 2.7 | -1 |
| -2.45 | 0.1 | -1 |
| -1.5 | -0.5 | 1 |
| 1.2 | -1.5 | 1 |

(a) Plot the data. Recall that the polynomial kernel is defined as $k(x, y) = (m + x^T y)^d$ for $m \in \{0, 1, 2, \dots\}$ and $d \in \{1, 2, \dots\}$. Each such kernel corresponds to a feature representation of the original data. Find the simplest polynomial kernel for which this data becomes linearly separable (**note:** simplest in this case is defined as the polynomial kernel with the smallest values of **both** $m$ and $d$). *What to submit: a single plot, a value for $m$, a value for $d$.*

(b) The optimal kernel found in (a) induces a feature representation in $\mathbb{R}^p$ for some integer $p$ determined by your choice of kernel, write down this feature representation. Then, choose a subset of two coordinates from this $p$-dimensional vector and plot the transformed data. Is your transformed data linearly separable? For example, for vector $(w, x, y, z)^T \in \mathbb{R}^4$, a subset of two coordinates could be the first two coordinates $(w, x)$, or the first and third coordinates $(w, y)$, etc.). *What to submit: a single plot, a description of the feature vector, and an answer to the question of linear separability.*

(c) Train a perceptron on the transformed data (using your feature transformation in (b)) with initial weight vector $w^{(0)} = \mathbf{1}_{p+1}$ (the vector of ones in $\mathbb{R}^{p+1}$). Use a learning rate of $\eta = 0.2$. Note: this should be done in numpy. Provide a table outlining all updates of the weight vector, and the iteration number at which the update occured. State the final learned perceptron and the number of iterations until convergence. Demonstrate that your perceptron correctly classifies each of the points. You may use the following table as a template for presenting your results:

| Iteration No. | $w_0$ | $w_1$ | $\dots$ | $w_p$ |
|---|---|---|---|---|
| 0 | 1 | 1 | $\dots$ | 1 |
| first update iteration | 1+$\delta_0$ | 1+$\delta_1$ | $\dots$ | 1+$\delta_p$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

where $\delta_j$ is the update for $w_j$ computed at the first iteration. To demonstrate that your perceptron classifies each point correctly, use the following table:

| $x_i$ | $\phi(x_i)$ | $y_i \phi^T(x_i) w^*$ |
|---|---|---|
| $[-0.8, 1]^T$ | $\phi([-0.8, 1]^T)$ | $r_1 > 0$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $[1.2, -1.5]^T$ | $\phi([1.2, -1.5]^T)$ | $r_8 > 0$ |

where $r_i = y_i \phi^T(x_i) w^*$ should be positive if your perceptron has converged. *What to submit: the*

*final weight vector, a table of iterations, a table showing that your model has converged, a print screen of the python code used, a copy of the python code used in your solutions.py file.*

(d) Let $x, y \in \mathbb{R}^n$ (i.e. $x$ and $y$ are $n$ dimensional vectors), and consider the following kernel:

$$k(x, y) = \prod_{i=1}^{n}(1 + x_i y_i).$$

What feature representation does this kernel induce? **Hint: try the cases when $n = 2$, $n = 3$, then figure out the pattern for larger $n$.** *What to submit: a solution can either be typed or handwritten (take a photo and upload to the pdf). All working must be shown for full marks.*

**Points Allocation** There are a total of 15 marks, the available marks are:

- Question 1 a): 1 mark
- Question 1 b): 1.5 marks
- Question 1 c): 2 marks
- Question 1 d): 1 mark
- Question 1 e): 1 mark
- Question 1 f): 2 marks
- Question 1 g): 1 marks
- Question 2 a): 1 marks
- Question 2 b): 1 marks
- Question 2 c): 2.5 marks
- Question 2 d): 1 marks

**What to Submit**

- A single PDF file which contains solutions to each question. For each question, provide your solution in the form of text and requested plots. For any question in which you use code, provide a copy of your code at the bottom of the relevant section.
- **.py file(s) containing all code you used for the project, which should be provided in a seperate .zip file.** This code must match the code provided in the report.
- You may be deducted points for not following these instructions.
- You may be deducted points for poorly presented/formatted work. Please be neat and make your solutions clear.
- You cannot submit a python notebook, this will receive a mark of zero. This does not stop you from developing your code in a notebook and then copying it into a .py file though.

**When and Where to Submit**

- Due date: Monday **March 29th**, 2021 by **5:00pm**.

- Late submission incur a penalty of 10% per day for the first 5 days and 100% for more than 5 days.
- Submission has to be done through Moodle.

**Final Reminder: You are required to submit one PDF file, AND also to submit the Python file(s) with all your code as a separate .zip file.**