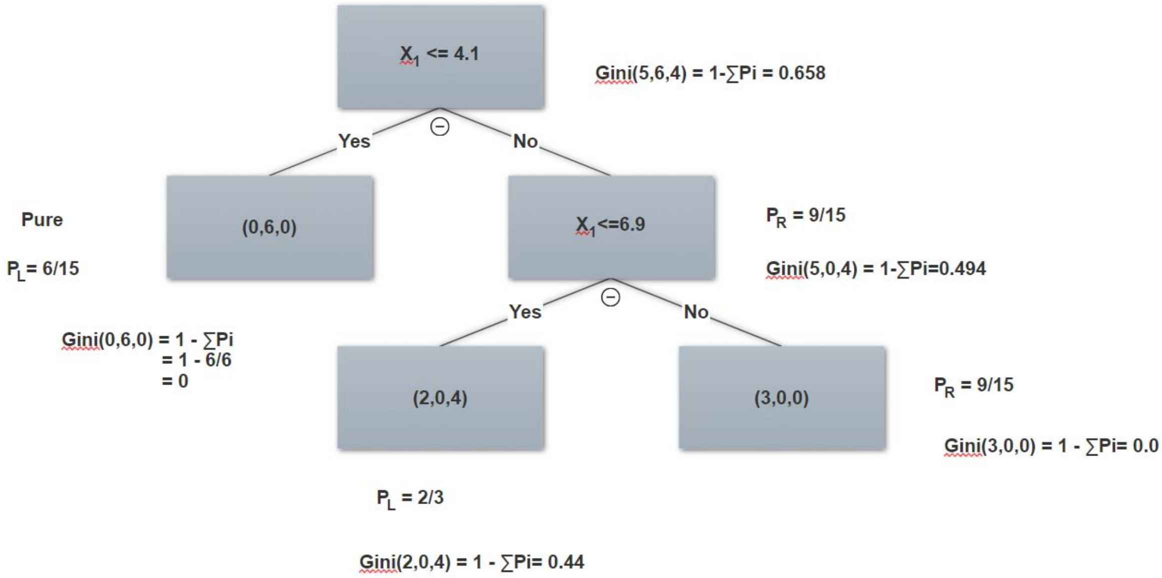
Machine Learning – Assignment 1

Akshay Tumkur Renukaprasad

MTK : 03694060

Problem 1:



Problem 2:

Using the decision tree from solution 1, X_a = (4.1, −0.1, 2.2) => P(c=1 | X_a , T_1) = 1/1 = 1 and X_b = (6.1, 0.4, 1.3) => P(c=1 | X_b,T_3) = 2/3

Problem 3:

(contd. In next page)

Programming assignment 1: k-Nearest Neighbors classification

In [2]:
```
import numpy as np
from sklearn import datasets, model_selection
import matplotlib.pyplot as plt
%matplotlib inline
```

Introduction

For those of you new to Python, there are lots of tutorials online, just pick whichever you like best :)

If you never worked with Numpy or Jupyter before, you can check out these guides

- https://docs.scipy.org/doc/numpy-dev/user/quickstart.html (https://docs.scipy.org/doc/numpy-dev/user/quickstart.html)
- http://jupyter.readthedocs.io/en/latest/ (http://jupyter.readthedocs.io/en/latest/)

Your task

In this notebook code to perform k-NN classification is provided. However, some functions are incomplete. Your task is to fill in the missing code and run the entire notebook.

In the beginning of every function there is docstring, which specifies the format of input and output. Write your code in a way that adheres to it. You may only use plain python and `numpy` functions (i.e. no scikit-learn classifiers).

Once you complete the assignments, export the entire notebook as PDF using nbconvert (https://nbconvert.readthedocs.io/en/latest/) and attach it to your homework solutions. On a Linux machine you can simply use `pdfunite`, there are similar tools for other platforms too. You can only upload a single PDF file to Moodle.

Load dataset

The iris data set (https://en.wikipedia.org/wiki/Iris_flower_data_set (https://en.wikipedia.org/wiki/Iris_flower_data_set)) is loaded and split into train and test parts by the function `load_dataset`.

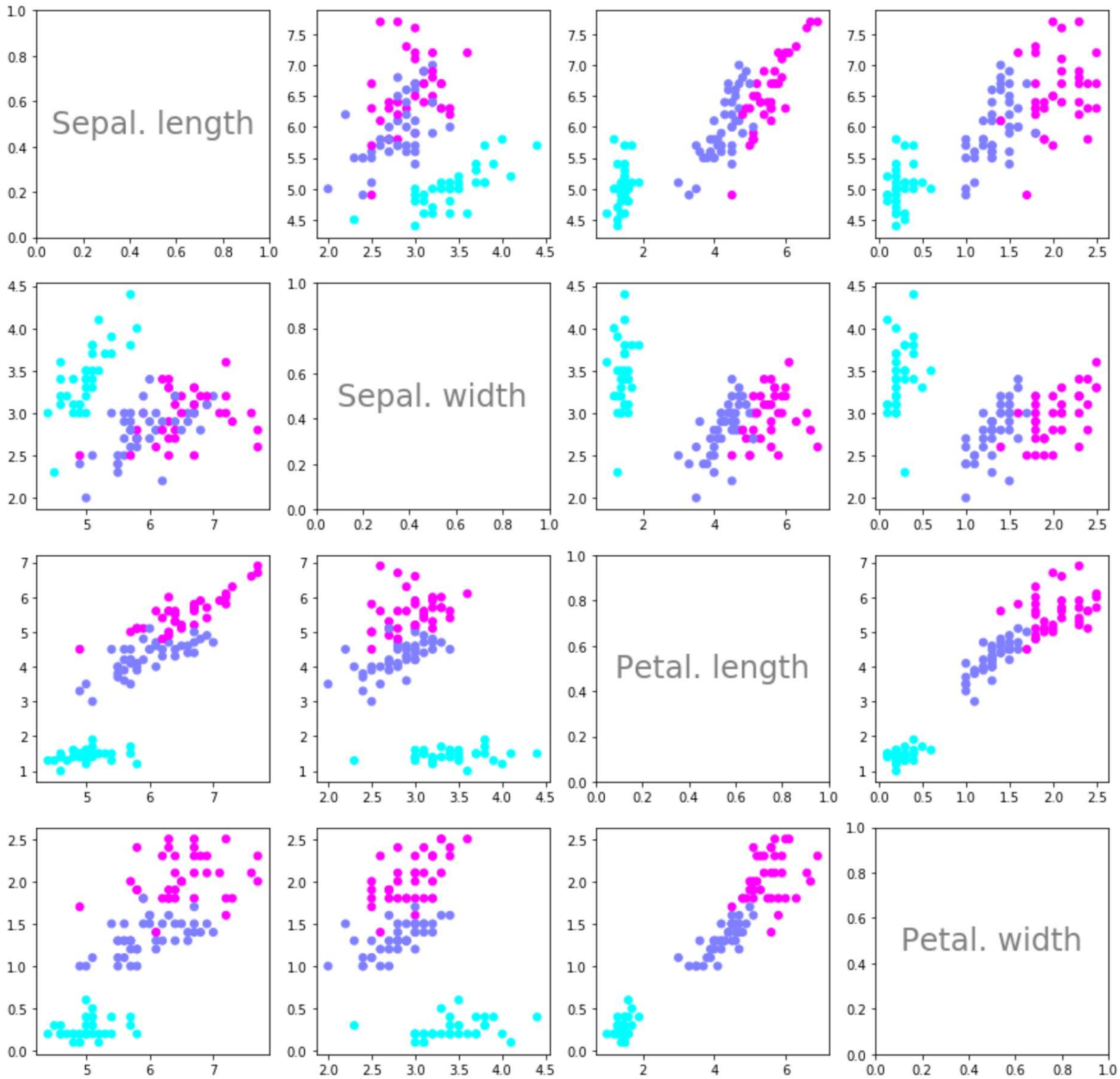
In [3]:
```
def load_dataset(split):
    """Load and split the dataset into training and test parts.

    Parameters
    ----------
    split : float in range (0, 1)
        Fraction of the data used for training.

    Returns
    -------
    X_train : array, shape (N_train, 4)
        Training features.
    y_train : array, shape (N_train)
        Training labels.
    X_test : array, shape (N_test, 4)
        Test features.
    y_test : array, shape (N_test)
        Test labels.
    """
    dataset = datasets.load_iris()
    X, y = dataset['data'], dataset['target']
    X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y,
random_state=123, test_size=(1 - split))
    return X_train, X_test, y_train, y_test
```

In [5]:
```
# prepare data
split = 0.75
X_train, X_test, y_train, y_test = load_dataset(split)
```

Plot dataset

Since the data has 4 features, 16 scatterplots (4x4) are plotted showing the dependencies between each pair of features.

In [6]:
```
f, axes = plt.subplots(4, 4,figsize=(15, 15))
for i in range(4):
    for j in range(4):
        if j == 0 and i == 0:
            axes[i,j].text(0.5, 0.5, 'Sepal. length', ha='center',
va='center', size=24, alpha=.5)
        elif j == 1 and i == 1:
            axes[i,j].text(0.5, 0.5, 'Sepal. width', ha='center', va='center',
 size=24, alpha=.5)
        elif j == 2 and i == 2:
            axes[i,j].text(0.5, 0.5, 'Petal. length', ha='center',
va='center', size=24, alpha=.5)
        elif j == 3 and i == 3:
            axes[i,j].text(0.5, 0.5, 'Petal. width', ha='center', va='center',
 size=24, alpha=.5)
        else:
            axes[i,j].scatter(X_train[:,j],X_train[:,i], c=y_train, cmap=plt.c
m.cool)
```



Task 1: Euclidean distance

Compute Euclidean distance between two data points.

In [20]:
```
import math

def euclidean_distance(x1, x2):
    """Compute Euclidean distance between two data points.

    Parameters
    ----------
    x1 : array, shape (4)
        First data point.
    x2 : array, shape (4)
        Second data point.

    Returns
    -------
    distance : float
        Euclidean distance between x1 and x2.
    """
    # TODO

    diff_squares = 0
    zipped_vector = zip(x1, x2)
    for x in zipped_vector:
        diff_squares += (x[1] - x[0]) ** 2
    return float(math.sqrt(diff_squares))
```

Task 2: get k nearest neighbors' labels

Get the labels of the *k* nearest neighbors of the datapoint *x_new*.

In [14]:
```
import operator
def get_neighbors_labels(X_train, y_train, x_new, k):
    """Get the labels of the k nearest neighbors of the datapoint x_new.

    Parameters
    ----------
    X_train : array, shape (N_train, 4)
        Training features.
    y_train : array, shape (N_train)
        Training labels.
    x_new : array, shape (4)
        Data point for which the neighbors have to be found.
    k : int
        Number of neighbors to return.

    Returns
    -------
    neighbors_labels : array, shape (k)
        Array containing the labels of the k nearest neighbors.
    """
    # TODO
    distances = []
    for x in range(0,len(X_train)):
        dist = euclidean_distance(x_new, X_train[x])
        distances.append((y_train [x], dist))

    distances.sort(key=operator.itemgetter(1))
    neighbors = []

    for y in range(0,k):
        neighbors.append(distances[y][0])
    return neighbors
```

Task 3: get the majority label

For the previously computed labels of the *k* nearest neighbors, compute the actual response. I.e. give back the class of the majority of nearest neighbors. In case of a tie, choose the "lowest" label (i.e. the order of tie resolutions is 0 > 1 > 2).

In [9]:
```
from scipy import stats

def get_response(neighbors_labels, num_classes=3):
    y=stats.mode(neighbors_labels)

    return y[0][0]
```

Task 4: compute accuracy

Compute the accuracy of the generated predictions.

In [15]:
```
def compute_accuracy(y_pred, y_test):
    """Compute accuracy of prediction.
    Parameters
    ----------
    y_pred : array, shape (N_test)
        Predicted labels.
    y_test : array, shape (N_test)
        True labels.
    """
    correct_count=0
    for i in range(0,len(y_pred)):
        if y_pred[i]==y_test[i]:
            correct_count += 1
    accuracy = float((correct_count/len(y_pred)))
    return accuracy
```

In [16]:
```
# This function is given, nothing to do here.
def predict(X_train, y_train, X_test, k):
    """Generate predictions for all points in the test set.

    Parameters
    ----------
    X_train : array, shape (N_train, 4)
        Training features.
    y_train : array, shape (N_train)
        Training labels.
    X_test : array, shape (N_test, 4)
        Test features.
    k : int
        Number of neighbors to consider.

    Returns
    -------
    y_pred : array, shape (N_test)
        Predictions for the test data.
    """
    y_pred = []
    for x_new in X_test:
        neighbors = get_neighbors_labels(X_train, y_train, x_new, k)
        y_pred.append(get_response(neighbors))
    return y_pred
```

Testing

Should output an accuracy of 0.9473684210526315.

In [21]:
```
# prepare data
split = 0.75
X_train, X_test, y_train, y_test = load_dataset(split)
print('Training set: {0} samples'.format(X_train.shape[0]))
print('Test set: {0} samples'.format(X_test.shape[0]))

# generate predictions
k = 3
y_pred = predict(X_train, y_train, X_test, k)
accuracy = compute_accuracy(y_pred, y_test)
print('Accuracy = {0}'.format(accuracy))
```

```
Training set: 112 samples
Test set: 38 samples
Accuracy = 0.9473684210526315
```

In [45]:
```
import pandas as pd
import numpy as np
import matplotlib as plt
from scipy.spatial import distance
import operator
from scipy import stats

def get_response(neighbors_labels, num_classes=3):
    y=stats.mode(neighbors_labels)
    return y[0][0]

filename = '01_homework_dataset.csv'
dataset = pd.read_csv(filename)
X_dataset = dataset.drop(dataset.columns[len(dataset.columns)-1],axis=1)
X_a = [4.1, -(0.1), 2.2]
X_b = [6.1, 0.4, 1.3]
distance_a = []
distance_b = []

for i in range(0,15):
    X_train = X_dataset.loc[i]
    Y_val = dataset.iloc[i,3]
    distance_a.append(((distance.euclidean(X_a,X_train)), X_train,Y_val))
    distance_b.append(((distance.euclidean(X_b,X_train)), X_train,Y_val))

distance_a.sort(key=operator.itemgetter(0))
distance_b.sort(key=operator.itemgetter(0))


neighbors_a = []
neighbors_b = []

for y in range(0,3):
    neighbors_a.append(distance_a[y][2])
    neighbors_b.append(distance_b[y][2])

class_a = get_response(neighbors_a)
class_b = get_response(neighbors_b)

print("Class of Vector Xa is\t" + str(class_a))
print("Class of Vector Xb is\t" + str(class_b))
```

```
Class of Vector Xa is   0
Class of Vector Xb is   2
```

Problem 5:

Using the formula,

$$\hat{y} = \frac{1}{Z} \sum_{i \in N_k(x)} \frac{1}{d(x, x_i)} y_i$$

$$y_i = \{0, 1, 2\}$$

$$Z \Rightarrow \frac{1}{\text{Euclidean distance}}$$

$$\hat{y}_a = \frac{1}{\left(\frac{1}{2.12} + \frac{1}{1.17} + \frac{1}{1.75}\right)} * \left(\frac{1}{2.12} + \frac{2}{1.17} + \frac{0}{1.75}\right)$$

$$= \boxed{1.39}$$

Problem 6:

It is just another classic case of Ambiguity. This can be overcome by using Normalization constant with a distance measure other than Euclidean (e.g. – Mahalanobis, Manhattan etc.)