INTELLIGENCE SYSTEM DEVELOPMENT
PTIT – D20CNTT, Semester I, 2023
quetd@ptit.edu.vn, tdque@yahoo.com
Thinking: *The more you study, the more you recognize that you understand nothing*

Due Date: 11:30 PM, THỨ HAI 11/09/2023

BÀI TẬP 2: Hiểu biết về xử lý dữ liệu và các thuật toán học máy cơ bản

2.1. Split a set into training and testing sets?
Presenting your knowledge in >3 pages and 5 code examples with running images
https://www.geeksforgeeks.org/how-to-do-train-test-split-using-sklearn-in-python/
https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/
https://towardsdatascience.com/how-to-split-a-dataset-into-training-and-testing-sets-b146b1649830
https://realpython.com/train-test-split-python-data/
https://www.geeksforgeeks.org/how-to-split-a-dataset-into-train-and-test-sets-using-python/

2.2. Cleaning data ([2], pag 107) – writing > 3 pages and running the example code

2.3. Wrangling data: Các bước wrangling data? writing > 3 pages an running examples
https://www.javatpoint.com/data-wrangling
https://www.jobsity.com/blog/a-guide-to-data-wrangling-in-python
https://www.geeksforgeeks.org/data-wrangling-in-python/
https://www.tutorialspoint.com/python_data_science/python_data_wrangling.htm

2.4. Evaluate ML models [TextBook 1.3]:
Presenting your knowledge in > pages and 5 code examples with running images
https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15
https://www.jeremyjordan.me/evaluating-a-machine-learning-model/#:~:text=The%20three%20main%20metrics%20used,the%20number%20of%20total%20predictions
https://www.altexsoft.com/blog/machine-learning-metrics/

2.5. Reading and running to discover k-nearest and k-means ([1.1] Chap 9, 10)

2.6. Sinh viên chạy ví dụ sau đây và giải thích từng dòng code

```
import numpy as np import matplotlib.pyplot as plt from matplotlib.colors import ListedColormap from sklearn import datasets from sklearn.neighbors import KNeighborsClassifier iris = datasets.load_iris() x = iris.data[:,:2] #X-Axis - sepal length-width y = iris.target #Y-Axis - species x_min, x_max = x[:,0].min() - .5,x[:,0].max() + .5 y_min, y_max = x[:,1].min() - .5,x[:,1].max() + .5 #MESH cmap_light = ListedColormap(['#AAAAFF','#AAFFAA','#FFAAAA']) h = .02 xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_max, h)) knn = KNeighborsClassifier() knn.fit(x,y) Z = knn.predict(np.c_[xx.ravel(),yy.ravel()]) Z = Z.reshape(xx.shape) plt.figure()
```

plt.pcolormesh(xx,yy,Z,cmap=cmap_light) #Plot the training points plt.scatter(x[:,0],x[:,1],c=y) plt.xlim(xx.min(),xx.max()) plt.ylim(yy.min(),yy.max())

2.7. Linear regression là gì? Sinh viên chạy ví dụ sau đây và giả thích từng dòng code

import numpy as np import matplotlib.pyplot as plt from sklearn import linear_model from sklearn import datasets diabetes = datasets.load_diabetes() x_train = diabetes.data[:-20] y_train = diabetes.target[:-20] x_test = diabetes.data[-20:] y_test = diabetes.target[-20:] x0_test = x_test[:,0] x0_train = x_train[:,0] x0_test = x0_test[:,np.newaxis] x0_train = x0_train[:,np.newaxis] linreg = linear_model.LinearRegression() linreg.fit(x0_train,y_train) y = linreg.predict(x0_test) plt.scatter(x0_test,y_test,color='k') plt.plot(x0_test,y,color='b',linewidth=3)

2.8. Chạy ví dụ và giải thích từng dòng code

import numpy as np import matplotlib.pyplot as plt from sklearn import linear_model from sklearn import datasets diabetes = datasets.load_diabetes() x_train = diabetes.data[:-20] y_train = diabetes.target[:-20] x_test = diabetes.data[-20:] y_test = diabetes.target[-20:] plt.figure(figsize=(8,12)) for f in range(0,10): xi_test = x_test[:,f] xi_train = x_train[:,f] xi_test = xi_test[:,np.newaxis]

xi_train = xi_train[:,np.newaxis] linreg.fit(xi_train,y_train) y = linreg.predict(xi_test) plt.subplot(5,2,f+1) plt.scatter(xi_test,y_test,color='k') plt.plot(xi_test,y,color='b',linewidth=3)

2.9. Logistic regression la gì? Chạy các ví dụ và giải thích
https://www.geeksforgeeks.org/understanding-logistic-regression/

2.10. SVM là gì? Chạy ví dụ và giải thích từng dòng code

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import svm
x = np.array([[1,3],[1,2],[1,1.5],[1.5,2],[2,3],[2.5,1.5],
 [2,1],[3,1],[3,2],[3.5,1],[3.5,3]])
y = [0]*6 + [1]*5
plt.scatter(x[:,0],x[:,1],c=y,s=50,alpha=0.9)
```

2.11. Chạy ví dụ và gỉai thích từng dòng code

import numpy as np import matplotlib.pyplot as plt from sklearn import svm from sklearn import datasets diabetes = datasets.load_diabetes() x_train = diabetes.data[:-20] y_train = diabetes.target[:-20] x_test = diabetes.data[-20:] y_test = diabetes.target[-20:] x0_test = x_test[:,2] x0_train = x_train[:,2] x0_test = x0_test[:,np.newaxis] x0_train = x0_train[:,np.newaxis] x0_test.sort(axis=0) x0_test = x0_test*100 x0_train = x0_train*100 svr = svm.SVR(kernel='linear',C=1000) svr2 = svm.SVR(kernel='poly',C=1000,degree=2) svr3 = svm.SVR(kernel='poly',C=1000,degree=3) svr.fit(x0_train,y_train) svr2.fit(x0_train,y_train) svr3.fit(x0_train,y_train) y = svr.predict(x0_test) y2 = svr2.predict(x0_test) y3 = svr3.predict(x0_test) plt.scatter(x0_test,y_test,color='k') plt.plot(x0_test,y,color='b') plt.plot(x0_test,y2,c='r') plt.plot(x0_test,y3,c='g')

2.12. Clustering là gì? Giải thích và ví dụ (>3 trang).

https://www.geeksforgeeks.org/clustering-in-machine-learning/
https://www.geeksforgeeks.org/different-types-clustering-algorithm/?ref=ml_lbp
https://www.tutorialspoint.com/machine_learning_with_python/clustering_algorithms_k_means_algorithm.htm