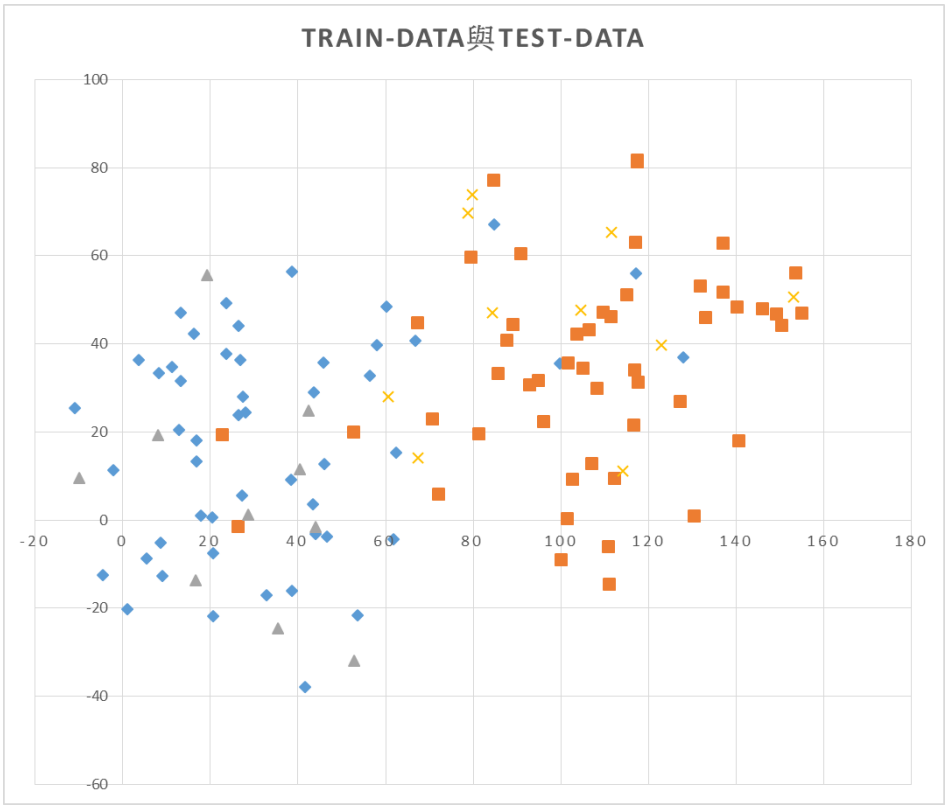
物件導向程式設計及應用 第一次作業

Due Date:2017/10/11 11:59p.m.

※注意事項：方案名稱請取名為學號_HW1(如 N12345678_HW1)，各小題分別開一專案，專案名稱為 HW1_題號(如 HW1_1)。請在程式碼開頭註明作者姓名與學號。函式與類別的宣告和實作需分別置於.h 和.cpp 檔中，請使用 C++語言，禁止使用標準樣板函式庫與任何外部函式庫。未依以上規定將斟酌扣分。繳交時請將整個方案與題目要求之其他檔案製成壓縮檔後準時上傳至 FTP server。遲交一率以零分計算，不提供補交。

本次作業之三個小題將實作三種簡單的機器學習相關演算法，分別是 K-NN、K-means 與 Pocket PLA，並透過 train data 建立模型、test data 比較結果。

題目附件中包含兩個.txt 檔分別是 train-data.txt 與 test-data.txt，train-data.txt 檔中含有 100 個二維座標點與其 label，檔案中一列三個數字依序為 X 座標、Y 座標與該座標之 label，其中 label 標示為 1 與-1 的各有 50 個。test-data.txt 檔中含有 20 個二維座標點與其 label，檔案中一列三個數字依序為 X 座標、Y 座標與該座標之正確 label，其中 label 標示為 1 與-1 的各有 10 個。



- 第一小題：(50%)

  K-NN(K-Nearest Neighbor)是一種簡單的機器學習演算法，其中最基礎的為 1-NN，其概念是尋找與輸入點最接近的 train data 分類作為此輸入點之分類答案[1]。

  由於實際應用中的 train data 數量通常相當龐大，為了增加效能，本題規定須先將 train data 中的座標點建立 K-D Tree 資料結構[2]後，由 K-D Tree 的特性快速尋找各 test data 座標在 train data 中的最鄰近點，由此得到該 test data 的分類結果，並與 test-data.txt 中的 label 做比較即可知道此分類器的準確率如何。

  本題須將 test data 之分類結果輸出成文字檔，命名為 test-result-1.txt，文字檔格式必須與 test-data.txt 中一致(座標在前、分類結果 label 在後)。並以 Excel 製作散佈圖，繪出所有 train data 座標點(不同 label 需有不同顏色)與 test data 座標點(不同 label 需有不同顏色)，其中 test data 分類錯誤的座標點請另外以特殊顏色繪出。將 test-result-1.txt、Excel 檔與另存為圖檔的散佈圖一起上傳。(散佈圖顏色請明顯區分各點群，Excel 檔與圖檔皆命名為 test-result-1)

- 第二小題：(25%)

  K-means 是一種資料分群演算法[3]，目的在將資料分成 K 群並尋找群中心，使得各群中所有的資料點到其所屬的群中心之距離總和最小。在電腦上的實作通常透過疊代尋找區域最佳解。找到群中心之後，讀入新的資料點，判斷此新的資料點與哪一個群中心較近即可做出簡單的分類[4]。

  本題規定先讀入 train-data.txt 中的資料，並取(140.57414041910056 , 18.0524769614646)作為 label -1 的初始群中心、(16.20521473399179 , 42.37661357896893) 作為 label 1 的初始群中心，疊代 50 次得到兩群各自的群中心後，讀入 test-data.txt 做出分類。如第一小題，輸出 test-result-2.txt，以 Excel 製作散佈圖，繪出所有 train data 座標點(不同 label 需有不同顏色)與 test data 座標點(不同 label 需有不同顏色)，其中 test data 分類錯誤的座標點請另外以特殊顏色繪出。將 test-result-2.txt、Excel 檔與另存為圖檔的散佈圖一起上傳。(散佈圖顏色請明顯區分各點群，Excel 檔與圖檔皆命名為 test-result-2)

- 第三小題：(25%)

  PLA(Perceptron Learning Algorithm)是一種簡單的線性二元分類器[5]，用於將可完全二分的資料分割，為了要用在如本題不可完全二分的資料上則須使用 Pocket PLA 疊代方式找最佳解。

  其概念是透過向量的特性不斷修正分割線，盡可能的使分類錯誤點數最少，詳細的理論與公式請見網站與教學影片[6] [7]。

  本題首先要讀入 train-data.txt 中的資料點，計算所有座標點的平均中心座標作為分割線必經的點，之後將 train-data 資料以 Pocket PLA 疊代至少 500 次(每次疊代隨機取點測試與修正)，得到 train-data 的分割線。請在程式主控台視窗上顯示 train data 分割線的方程式與分割錯誤的點數量。

  得到 train data 的分割線後，讀入 test-data.txt 中的資料點，以點在分割線的哪一側做出分類，與前兩小題一樣可以得到準確率並輸出 test-result-3.txt，以 Excel 製作散佈圖，繪出所有 train data 座標點(不同 label 需有不同顏色)與 test data 座標點(不同 label 需有不同顏色)，其中 test data 分類錯誤的座標點請另外以特殊顏色繪出。將 test-result-3.txt、Excel 檔與另存為圖檔的散佈圖一起上傳。(散佈圖顏色請明顯區分各點群，Excel 檔與圖檔皆命名為 test-result-3)