Machine Learning, Fall 2019: Project 1

Liam Li G48502460

**Head:**

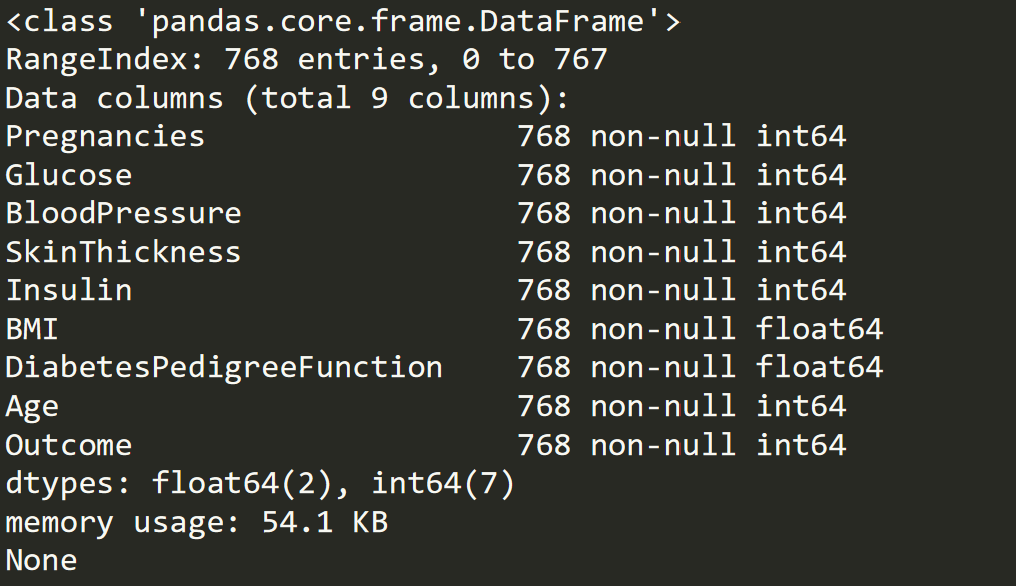
Python: pandas and matplotlib modules for visualization

Sublime text 3

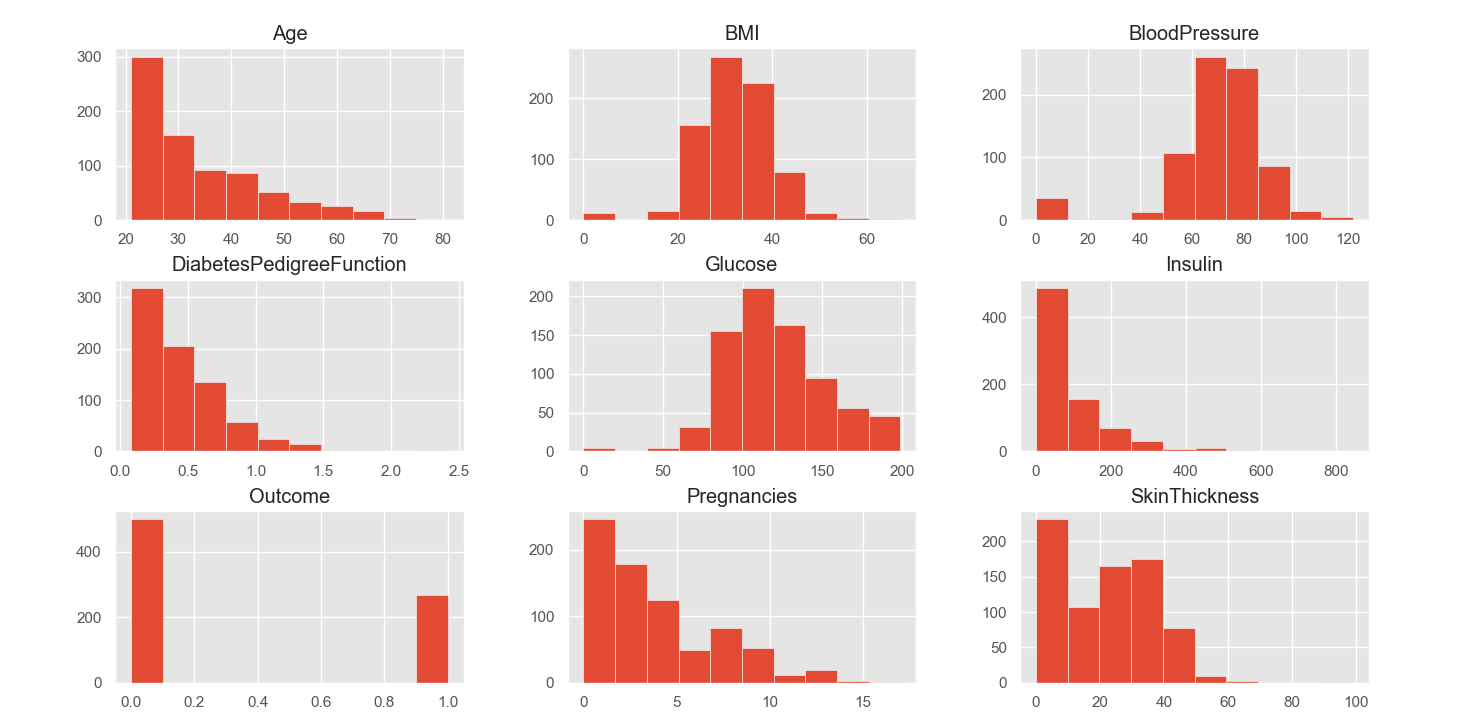
Pima Dataset

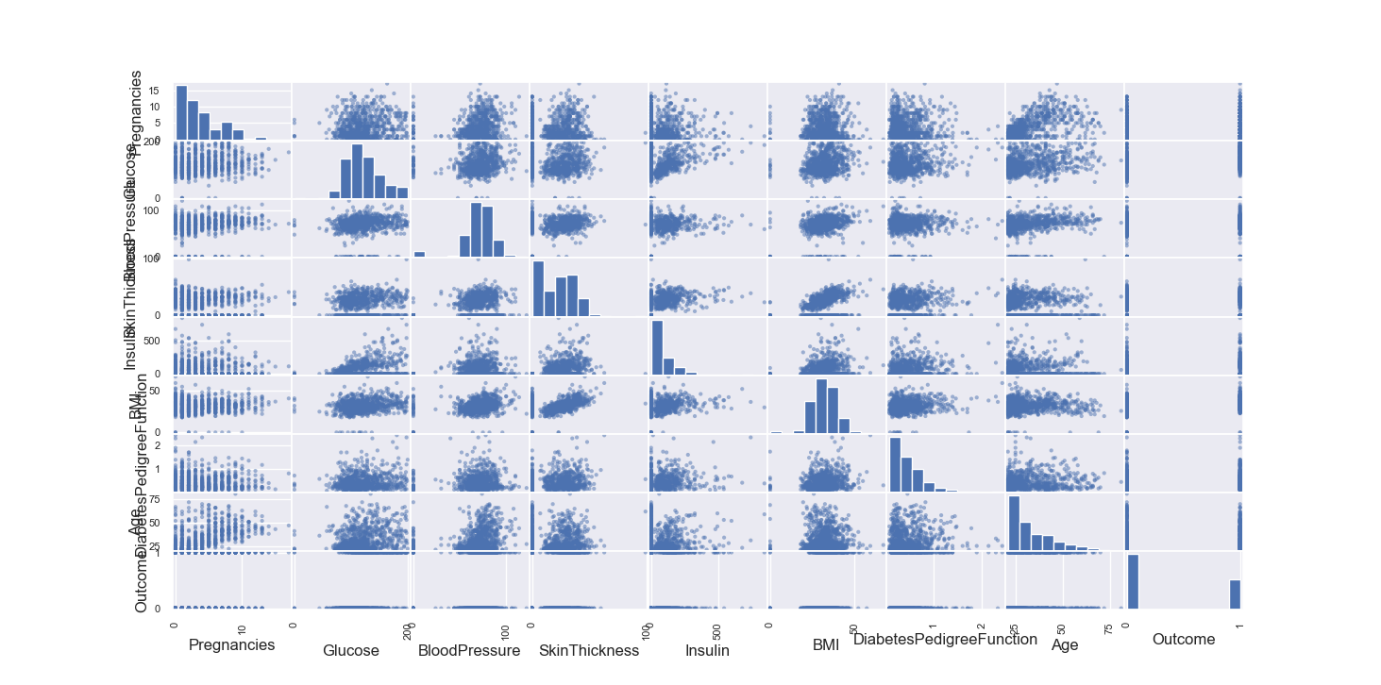
**Dataset details:**

This dataset contains eight medical factors that may affect diabetes, and one outcome, which indicates that whether the unit has diabetes. More details show in the image below.



The histogram represents the distribution of each column.

****

****

For this dataset, I split the data into training set, validation set and test set as 80%, 10% and 10%.

**Algorithm description:**

This project is a classifer based on K-NN algorithm.

First, It transfers csv data file into a 2D array by readfile() function.

Then it splits data into into training set, validation set and test set as 80%, 10% and 10% by split() and split\_scale(), split\_scale() function is used to guarantee the proportion of labels in each splitted data is same as the proportion in the whole dataset.

It doesn’t include feature scaling considering this dataset is relatively small.

Three distance metrics are used to compare the result, which are Manhattan distance(p=1), Euclidean distance(p=2), and p=4.

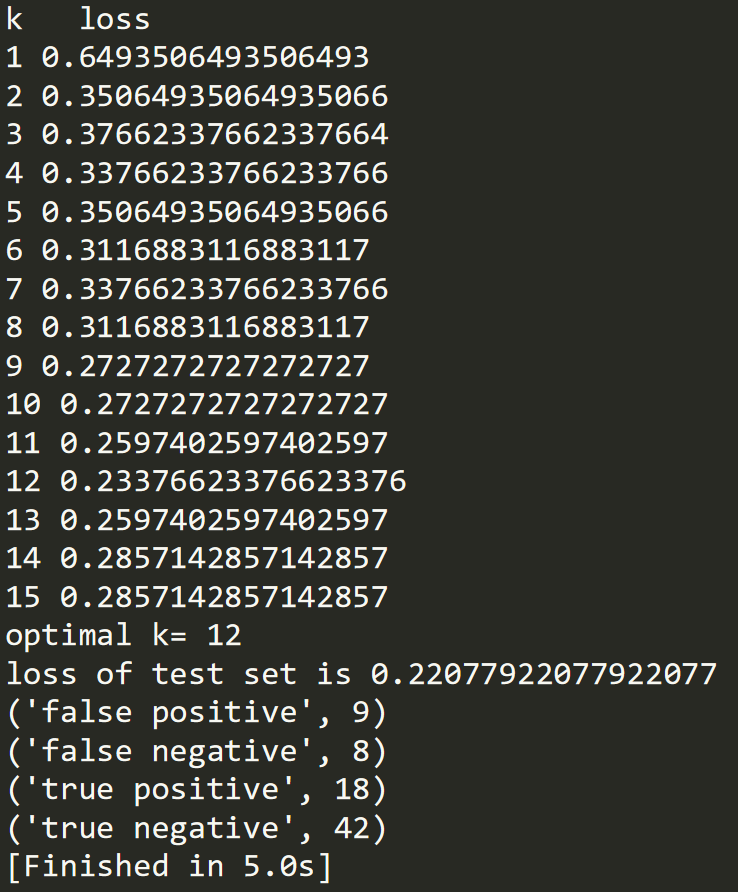
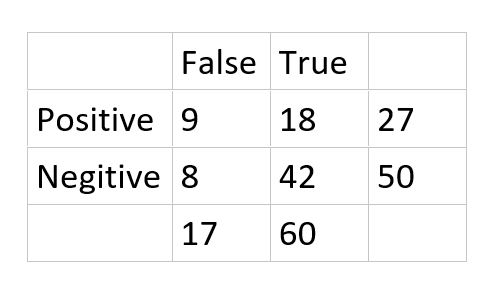
Knn(target,d,k) returns a prediction given by the majority of k neighbors.

Loss(data,k) returns the error rate on data.

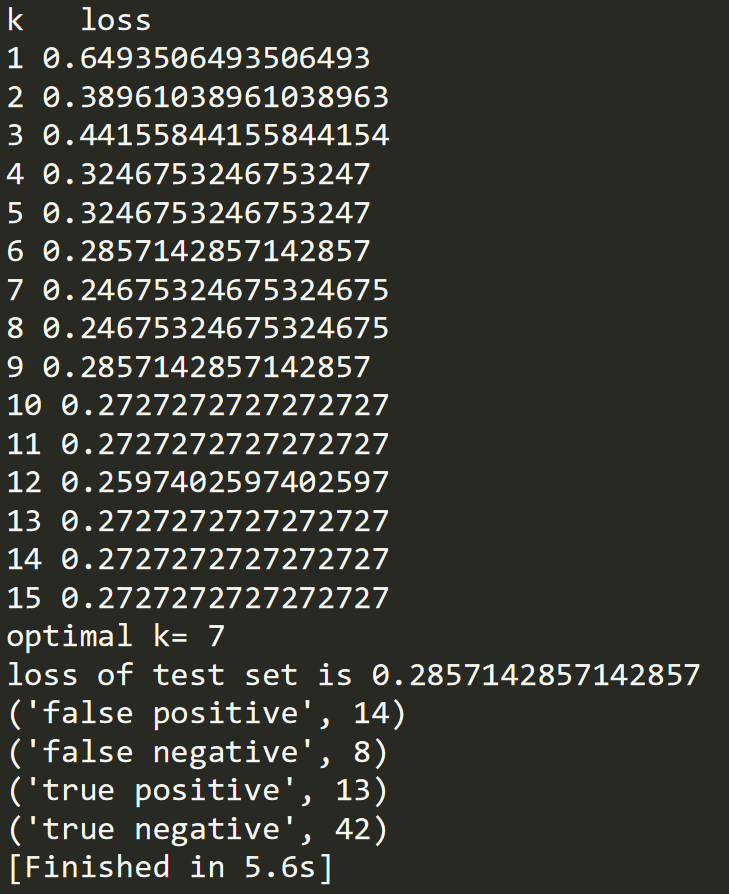
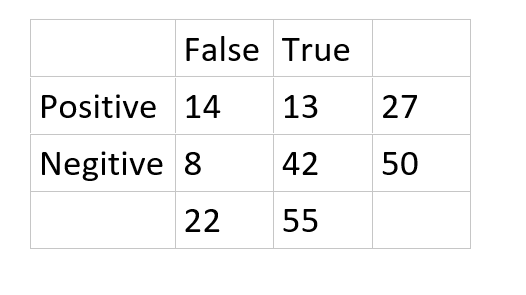
Evaluation(data,ks) shows different error rate when k in range(1,ks=15), by observation of the result we can obtain the optimal k for the classfier.

**Algorithm results:**

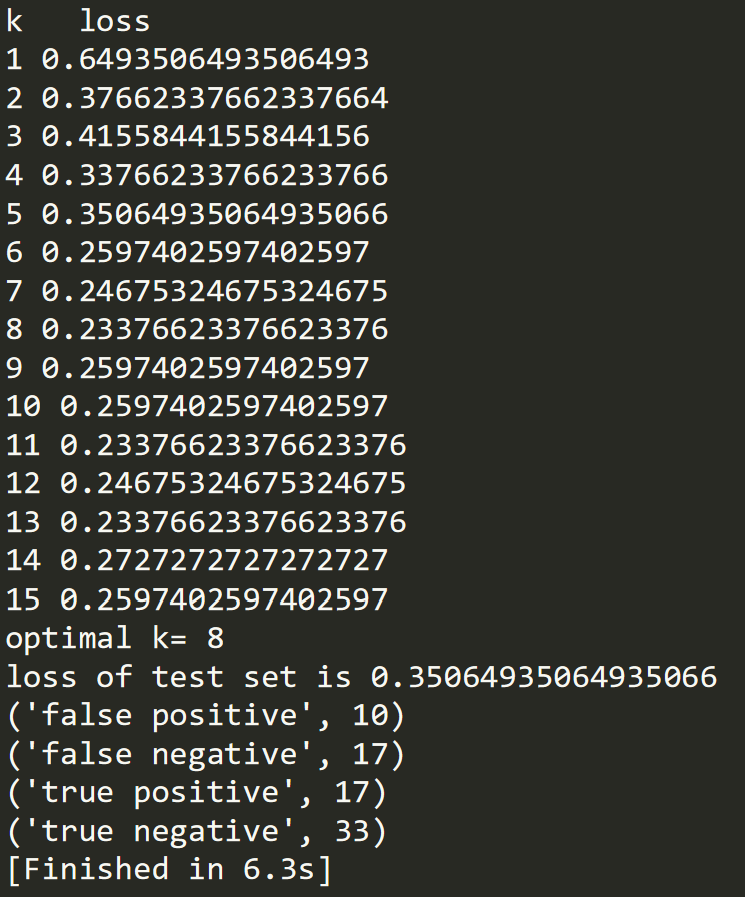
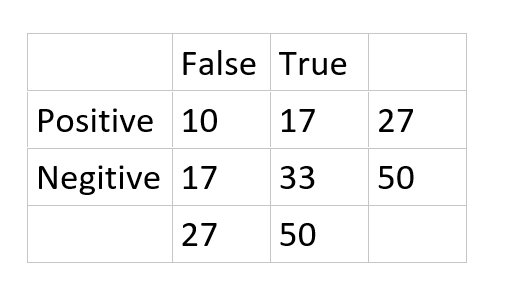
Manhattan distance(p=1)

Euclidean distance(p=2)

p=4.

**Runtime**

Time complexity is O(nd), n is the size of training data, d is the number of features.

Wall-clock time is presented in the images of the result part (including loading data file and splitting data).

MNIST Dataset

**Dataset details:**

“The data files train.csv and test.csv contain gray-scale images of hand-drawn digits, from zero through nine.”

“Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.”

 ![图片包含 物体

描述已自动生成](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAABwAAAAcCAAAAABXZoBIAAAAV0lEQVR4nGNgZBhgwNzbzYxTkvPfPw4kLhM+gzAkk/BJ+pFtLG7Jv7vxSP5agEeSxRKPJZz//u0gz0E0kpyMKsmCwpNlxCPJ8J94O1GB/b9/dsSrpgkAADYxDnZH2YttAAAAAElFTkSuQmCC)    ![图片包含 剪贴画

描述已自动生成](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAABwAAAAcCAAAAABXZoBIAAABI0lEQVR4nMWRMUsDQRCFn2fCwpGoGIs0KS4mIHZWKST3D6wVUgh2Rwo7wTqdhUEkEhD8CYKFlaWNiAZsrvAQDAFBJSQatVFfLM7NZvYH6DT7hm9n9s0skvjPSARPtdqCUkqpCZvN3XAUGzZsGMarVNxLs6yP72YIeBXAbb3Jwjp5qXWxDFkJYF+LKIpP5zefzOCzb9vQptMvaHtAIfAAHERtcSlNPuR2Th9jtyeegG6HvDOzHMv/2CRJtg593w/eyVkBp844OCrNAABuyZK0llxa1HKL3B0fBVheCbW8+LCHajzPj/Q9e66oRGY7b5LpMQCgSPb2qjkgX21+DZkS0FkPSb52uwOSvFbWq6uh2UIWMLsFACSctQLKOAfq/aFt+I/iB7ijiEfij/wjAAAAAElFTkSuQmCC)    

        ![图片包含 剪贴画

描述已自动生成](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAABwAAAAcCAAAAABXZoBIAAABG0lEQVR4nM2RIUhDYRSFj6bB9piCC7KgICgIY8m1gVGEIYZVTVuxDIMYnNNksVgURWwKaza3LIIWi0VQlD3FBcOi7zm+f4bN6f63aPCmCx/n3nvOVUj/uBJnb6ZVzi6Eg8gpeoAB6tmozarQgVCUJA10Yf5Akh4uXkYy0yJ981sYfoRy8nZSUugZ0j1Tx19htt0uvuPP9O7chbshSVppwJp10DzwNCeVfDh0LBjdbsJnpdKE63jQ6HrHynGfEKSUMaZlTKlfQsvfCVUjNhu7Apq54U0fzm2Yh4+TUUmJBn7Sgvew0+4ysCVJGuzCmlSPSVJsSbq0lDmgtpqKF1yCY52jn5ftBaxECq4HBryNgBVJmjp1DfsT/dAf1hcy45xcNRwviAAAAABJRU5ErkJggg==) 

![图片包含 剪贴画

描述已自动生成](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAABwAAAAcCAAAAABXZoBIAAABEElEQVR4nM3PsUsCcRjG8QdxqUmCCkoassClodaGyCGIqKC1wSUqikIOIXTsFoem6M9ojaI1KIRoiQsqoqUhbAjJTg/7aoMi97u7OXqm5+UzvO8r/YMsX7Va7eNUpE25ADxOhmnUesFOpy/gPmTWA7AnzVd5D1D/aRPeblOSzvhIGBbbBA6TkqQ5yBu4A5SG5cd4DzPSSaHd6bOBjTPfkO32sSfcJZ8NnuNZse5QwLx2C5xer4Hts5VPvPVOTWQ9KI37sAJHnYeG7oDGokxckyRNAzhJv+V+KMclqfgFbPQZf6zWuZY0YdehuR2TmQrV3YWbV8DJKZhiHQCeL0dCJhUbQG1/IIIkHeDmM9H0p/kFBAmKsAFkVzQAAAAASUVORK5CYII=)   ![图片包含 剪贴画

描述已自动生成](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAABwAAAAcCAAAAABXZoBIAAABFklEQVR4nM2QvS+DYRTFT9VHYqggFklpNBIlYsAmIZGUmXQ3NTFYsVh0kHTtgsQg/gSDqaMwdGlEKkRIOpDSD6W06v2xtHhfj83gTPc5v+fm3nPl1v+Sy/7sz+akYIveDn5+XXm5WTh5sMCKOZFv65VPOdhsHsA6j4zswqEDhqC4szEmDWQoDDng+GPCJ8m9WsaaN68/eQQsGpE/ZkFuwmVAns0SsNdlQG3hawDu5gwwVI/45K05jV+wWrndTyWleGuzoXW4R5L0zLQ5iNSwVMVvcwK99aopDPnu7yxaKi5LkrzRDFxO2RrfuQ92BGa2L7JwtdZun5OA0zRA5SzicS7ReVyGciG1Pmr3a1cc7FM6+VuAv9YH/gKEwdvuVPQAAAAASUVORK5CYII=) ![图片包含 家具, 座位

描述已自动生成](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAABwAAAAcCAAAAABXZoBIAAAAlElEQVR4nM2RwQ3DMAhFSdUx7K4RsUfKHFX2aBfpHjFzBO/hHApWGpMol0p9F2yeQV8yXOG/oWl1uWwtOjLqhHxZo9CnzuSsTVq5d2TWJo/RCaS9l6AjDX60MohJfG7fxzKRLi6zNe3HRBAhZc5DADbZ1VmMMMAd5I3h1gSpjGU/LeQ27Rral6meHClHEg5lOiV/xALz7iOr2RmdlAAAAABJRU5ErkJggg==)  ![图片包含 剪贴画

描述已自动生成](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAABwAAAAcCAAAAABXZoBIAAABEUlEQVR4nNXQzyvDcRzH8deGLbv4MSVDKYXy6ybloEkc5OK49jcoBwe5KHcuDsuJMD+KXFworSg3LZSL+VHzK2qjhpXnh4vDvrbPzcXr+ujd691LHv1ZStcOg1achXWbLaXZKytM3nPDU2Vh8y1jdi13JVuQrLD0jWfI+i02meG422KhFBcdFlOE7IAkqT+RmAg7bc6YqCR/77MxxuzUO/CAu3KpZxO4hiMHtryzIgU/DFNtNXH2c82zSKrLNfrGw7Db3Zn8nMnF6hsW5D3hVs2+QXiVJBX/4ONLXZPCrXKFhjQiTTveabgnUrQBxE7hst3twEActueBL0g3/p4gcAYAhqux/IFqVwGI9VXl27/INyBufLgvv0caAAAAAElFTkSuQmCC)   

**Algorithm description:**

This project is a classifer based on K-NN algorithm.

First, It transfers train.csv data file into a 2D array by readfile() function. Due to limited time and computing performance, I shrink the dataset 15 times smaller. The size of training data is 2240.

Then it splits data into into training set and validation set as 80%, 20% by split(), and test set is originally separated as test.csv.

Distance metrics is Euclidean distance(p=2).

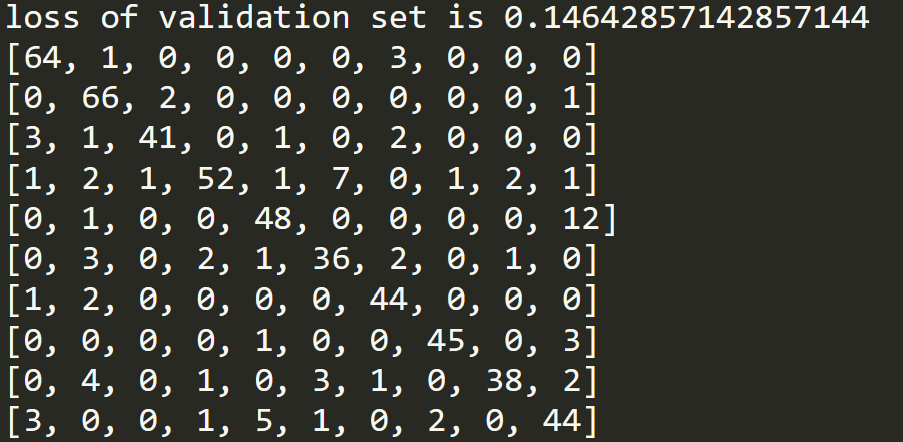
Knn(target,d,k) returns a prediction given by the majority of k neighbors.

Loss(data,k) returns the error rate on data.

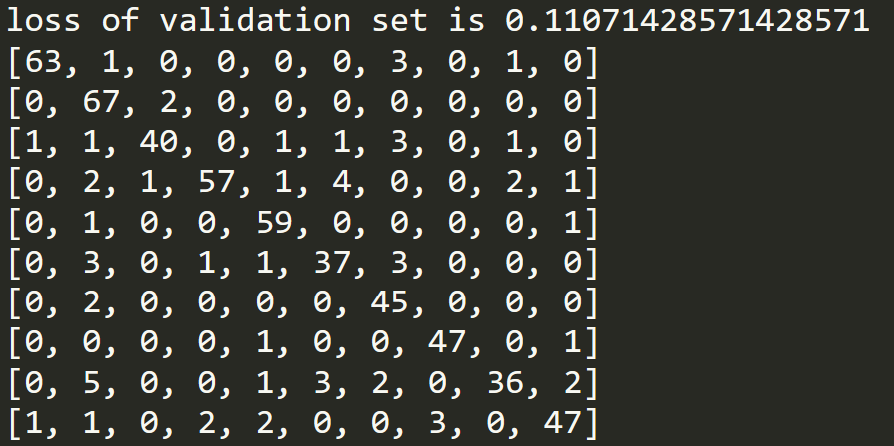
Evaluation(data,ks) shows different error rate when k in (3,6,9,12,15), by observation of the result we can obtain the optimal k for the classfier.

**Algorithm results:**

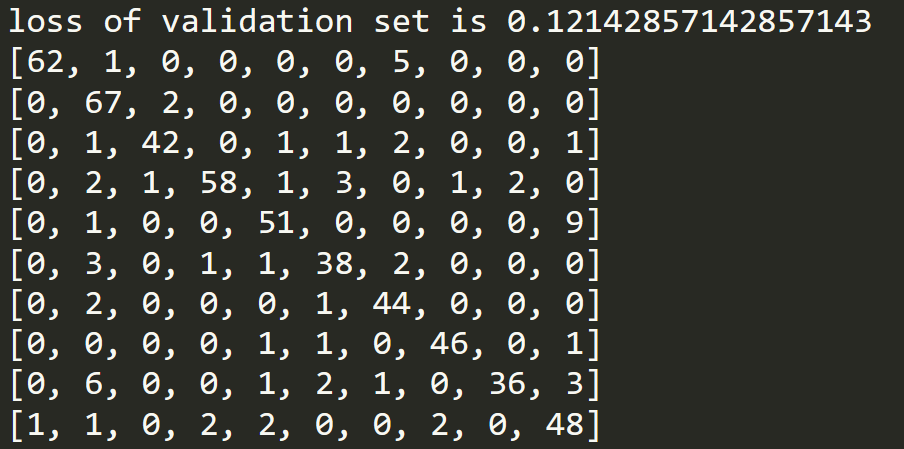
When k=3, loss is 15%. Matrix[i][j] is the number of the scenario that actual digit is I and predicted digit is j. The size of validation set is 560.

****

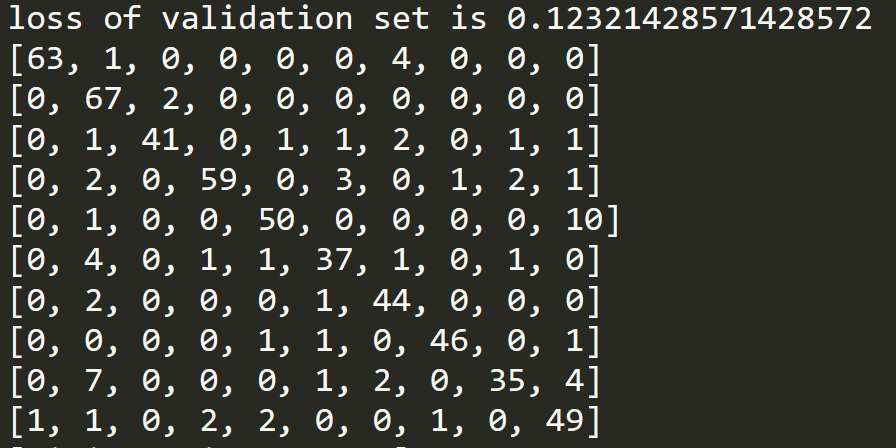
When k=6, loss is 11%. Matrix[i][j] is the number of the scenario that actual digit is I and predicted digit is j. The size of validation set is 560.



When k=9, loss is 12%. Matrix[i][j] is the number of the scenario that actual digit is I and predicted digit is j. The size of validation set is 560.



When k=12, loss is 12%. Matrix[i][j] is the number of the scenario that actual digit is I and predicted digit is j. The size of validation set is 560.

****

**Runtime**

Time complexity is O(nd), n is the size of training data, d is the number of features.

Wall-clock time is presented in the images of the result part(including loading data file and splitting data).

Reference

1. <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
2. <https://www.kaggle.com/c/digit-recognizer/data>
3. <https://www.kaggle.com/shrutimechlearn/step-by-step-diabetes-classification-knn-detailed>