Final Project Specifications

CMPE 650
kNN Overview

• **K-Nearest Neighbors (kNN)** is a relatively simple ML algorithm in which an object is classified by a majority vote of its k nearest neighbors. Given a new input, its nearest neighbors are computed using some distance metric such as Euclidean distance.

• Basic Idea:
  – During a *training phase*, training feature vectors are stored along with their corresponding class label.
  – During a *testing phase*, test feature vectors are provided. For each new feature, the k nearest neighbors are found and the majority class label is assigned to this feature.
**KNN Euclidean classifier**

- Finds K- nearest neighbors to the test input and decides based on the majority vote of the neighbors.
- utilizes Euclidean distance

\[
d_1 = \sqrt{(x_{Test-f1} - x_{Train-f1_1})^2 + (x_{Test-f2} - x_{Train-f2_1})^2 + \cdots + (x_{Test-fm} - x_{Train-fm_1})^2}
\]

\[
d_2 = \sqrt{(x_{Test-f1} - x_{Train-f1_2})^2 + (x_{Test-f2} - x_{Train-f2_2})^2 + \cdots + (x_{Test-fm} - x_{Train-fm_2})^2}
\]

\[\vdots\]

\[
d_n = \sqrt{(x_{Test-f1} - x_{Train-f1_n})^2 + (x_{Test-f2} - x_{Train-f2_n})^2 + \cdots + (x_{Test-fm} - x_{Train-fm_n})^2}
\]
Simulation Setup for EEG recordings

• EEG recording (extracted features) for one patient for 22 channels, 29172 lines:

• Every 22 channels correspond to one second of data.
  – There are total $29172/22 = 1326$ seconds of data (or group of data)
  – You can treat each group as one 9x22 feature vector or treat each channel as one feature vector, then do the classification per channel and then do majority vote or threshold-based decision.

• Training and Test for machine learning:
  – Take $2/3$ of the database as training and $1/3$ as test.
  – e.g take the first $29172/3 = 9724$ lines of the code for test and the rest use for training (i.e $29172 - 9724 = 19448$). This means that there will be total $9724/22 = 442$ seconds of test data and $1326 - 442 = 884$ for training.
Simulation Setup for EEG recordings

• For every set of that 442 group, run the KNN
  – For each feature test vector compute the distance from all 29172*2/3=19448 training vectors
  – Find the K min distances among calculated ones.
  – Look at the majority vote of labels for that K min distances