EE 559 **Overview of Project Datasets** 4/18/2020

v1.1 (updated 4/20/2020)

## 1. Wireless Indoor Localization Dataset

Contributed by: Zihang

**Summary description:** An indoor WiFi system has 7 WiFi routers at various locations. The goal is to use the WiFi signal strength received from the 7 routers, to predict the location of a user (who is in one of 4 rooms). Thus, there are 7 input variables (features) and 4 classes (user locations).

**Data Description [1]**: 8 columns: column 1~7 refer to the measured signal strength (dB, integers) at 7 wireless sensors (WS1-WS7 below) (routers); column 8 is the user location (class). There are 500 data points for each location (class).

Full dataset [1] has N_Total=2000, C=4, No. of features=7. All features are integer-valued; there are no missing values.

Sample data for user localization using wireless signal strength [2]:

| WS1 | WS2 | WS3 | WS4 | WS5 | WS6 | WS7 | Class |
|---|---|---|---|---|---|---|---|
| −64 | −56 | −61 | −66 | −71 | −82 | −81 | 1 |
| −68 | −57 | −61 | −65 | −71 | −85 | −85 | 1 |
| −17 | −66 | −61 | −37 | −68 | −75 | −77 | 2 |
| −16 | −70 | −58 | −14 | −73 | −71 | −80 | 2 |
| −52 | −48 | −56 | −53 | −62 | −78 | −81 | 3 |
| -49 | -55 | -51 | -49 | -63 | -81 | -73 | 3 |

**Required training and test sets:** Separate testing and training sets are being extracted and will be posted on D2L; they are D_Test (with N_Test = 400) and D_Train (with N_Train = 1600). For the class project, you are required to use these sets as partitioned, so that everyone's test-set accuracy can be compared fairly. (You are free to further divide D_Train as you wish for cross-validation, etc.)

**Comment**: Theoretically, the dataset should be linear separable if the locations are not really close to each other and the wireless routers and the geometry are not all symmetric. However, results obtained by the paper referenced on the UCI website imply that the dataset is probably not linearly separable.

**Tip:** this dataset is pretty straightforward to use for classification. If you want to extend the problem, here are a few suggestions. (i) You can try putting more effort into feature-space dimensionality, such as introducing new features as nonlinear functions of the given features (nonlinear mapping), and/or reducing dimensionality of feature space (e.g., with linear transformations or other feature selection). (ii) Perform additional analysis. If you have some knowledge of the problem domain, what could be done to improve the features, or how might one

improve the localization in a future system? (iii) You can add a confidence measure to the classification outputs, and see how quickly the accuracy improves when a threshold is placed on the confidence measure (i.e., data points that are below some minimal confidence measure are put into a "reject" class); we will provide more information on confidence measures (probably in Discussion 14). (iv) Follow ideas of your own that use this data, are relevant to the problem, and use pattern recognition techniques; it is advisable to check with a TA or professor to see if they think it's a reasonable idea. In particular, Zihang has some background in this topic area, so could be helpful if your questions or ideas relate specifically to the domain of the problem.

**References**


[1] Dataset information: http://archive.ics.uci.edu/ml/datasets/Wireless+Indoor+Localization

[2] Jayant G Rohra, Boominathan Perumal, Swathi Jamjala Narayanan, Priya Thakur, and Rajen B Bhatt, 'User Localization in an Indoor Environment Using Fuzzy Hybrid of Particle Swarm Optimization & Gravitational Search Algorithm with Neural Networks', in Proceedings of Sixth International Conference on Soft Computing for Problem Solving,2017, pp. 286-295.

## 2. Hand Postures (from motion capture) Dataset

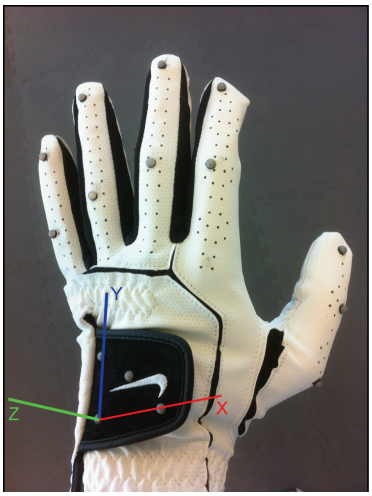
Contributed by: Pratyusha and Fernando.

**Summary description:** "5 types of hand postures from 12 users were recorded using unlabeled markers on fingers of a glove in a motion capture environment" [1]. A Vicon capture camera system was used in an indoor environment to obtain the data. The pattern recognition problem is, given the position of the markers on the glove, predict the hand posture.

**Challenging aspects:** Because the markers are unlabeled, the input data (features) only record marker positions relative to a reference point on the glove. Which glove-finger each marker was on, (and which marker it is on the glove-finger), is not known. Additionally, some of the markers are typically occluded by other fingers or other portions of the hand, so that the number of markers recorded varies from one input data point to another. Also, one subject contributes many data points (many captures of a hand posture), so extra care must be taken in the creation of training, validation, and test sets.

**Comment:** There are different ways to define the pattern recognition problem on this dataset. For our class project, we will treat this as a user-independent posture recognition problem (as has been done in the literature). So the goal is to be able to classify the posture, for (any) users that have not been trained on.

**Data description:** Full dataset posted on the UCI ML Archive [1] has N_Total=78095, C=5, from 14 users*. Number of input attributes per data point: varies from 9 to 36. See below for the datasets you are required to use.




The picture shows "the glove used as the data source…. The axes of the local coordinate system based upon the rigid pattern are shown" [2]. The large dots are the markers. The reference point used for the local coordinate system is at the origin of the axes shown.

**Required training and test sets:** A downsampled training set and separate test set have been posted on D2L. *Two of the users (users 4 and 7) have been removed before preparing all the D2L-posted datasets, because of the small number of data points for each of those users. Note

also that there is no user 3. The required D2L-posted datasets are D_Test which has data from 3 users (with N_Test = 21099), and D_Train which has data from the other 9 users with 1500 data points per user (N_Train = 13500). (The first column in each of the D2L datasets is the record (data point) number from the original UCI-posted dataset; you are unlikely to need these and can safely omit them.) For the class project, you are required to use D_Train and D_Test as partitioned, so that everyone's test-set accuracy can be compared fairly. (You are free to further divide D_Train as you wish for cross-validation, etc.)

**Optional training set:** You can optionally also train and on a larger set (same users as D_Train but more data points (posture captures) per user). This larger training set includes all the data available in the original UCI ML Archive dataset, for users in D_Train. This set is also posted on D2L as D_Train_large (with N_Train_large = 56125). Note that in D_Train_large, the number of data points per user varies. Of course, the compute time will be substantially larger for these, so you may need to revise your code or approach.

**Tips on dealing with challenging aspects**

**Tip 1:** Different data points recorded for a given user are likely highly correlated. Thus, to treat the data correctly, one must keep track of the "user ID" in dataset handling, and ensure that data from any one user is not shared amongst training, validation, and test sets. Thus, whenever the data is split, any user in a validation set cannot also have data points in a training set or the test set; similarly, any user in a test set cannot also have data points in a training or validation set. So data is first split by user ID, then stratified by percent representation of each class (as usual). Because of the small number of users, for validation (e.g., model selection), it is recommended to use cross-validation with the leave-one-user-out technique (each validation set has all the data from just 1 user).

**Tip 2:** The data will not work well in its raw form, because the number of features varies from data point to data point, and further the maker number (0 for $x_0, y_0, z_0$, etc.) has no significance (because the markers are unlabeled). To convert the data to a usable form, it is recommended to define new features that are extracted from the given data. Here are examples of 13 features that you could define (and feel free to create and define more features of your own), for each data point: number of recorded markers, mean $x$ of marker locations, mean $y$ of marker locations, mean $z$ of marker locations; standard deviation of $x$ of marker locations (similarly for $y$ and $z$); maximum $x$ of marker locations (similarly for $y$ and $z$), and minimum $x$ of marker locations (similarly for $y$ and $z$).

**References**


[1] Dataset: http://archive.ics.uci.edu/ml/datasets/Motion+Capture+Hand+Postures

[2] Andrew Gardner, "Datasets for Motion-Capture-Based Hand Gesture Recognition", http://www2.latech.edu/~jkanno/datadescription-1.pdf