Part 1: Naive bayes

1.Load the kinematics dataset as measured on mobile sensors from the file “run_or_walk.csv”. List out the columns in the dataset.

2. Let the target variable ‘y’ be the activity and assign all the columns after it to ‘x’.

3.Using Scikit-learn fit a Gaussian Naive Bayes model and observe the accuracy. Generate a classification report using scikit learn.

4.Repeat the model once using only the acceleration values as predictors and then using only the gyro values as predictors. Comment on the difference in accuracy between both the models.

Part 2: SVM

1. Load the data from “college.csv” that has attributes collected about private and public colleges for a particular year. We will try to predict the private/public status of the college from other attributes.

2. Use LabelEncoder to encode the target variable in to numerical form and split the data such that 20% of the data is set aside for testing.

3. Fit a linear svm from scikit learn and observe the accuracy.
[Hint: Use Linear SVC]

4. Preprocess the data using StandardScalar and fit the same model again and observe the change in accuracy.
[Hint: Refer to scikitlearn’s preprocessing methods]

5. Use scikit learn’s gridsearch to select the best hyperparameter for a non-linear SVM, identify the model with best score and its parameters.
[Hint: Refer to model_selection module of Scikit learn]

Project: Domain Media

Challenge/requirement
Motion Studios is the largest Radio production house in Europe. Their total revenue $ 1B+. Company has launched a new reality show – "The Star RJ". The show is about finding a new Radio Jockey who will be the star presenter on upcoming shows. In first round participants have to upload their voice clip online and the clip will be evaluated by experts for selection into the next round. There is a separate team in the first round for evaluation of male and female voice. Response to the show is unprecedented and company is flooded with voice clips.

You as a ML expert have to classify the voice as either male/female so that first level of filtration is quicker.

Fields in Data

• meanfreq: mean frequency (in kHz)
• sd: standard deviation of frequency
• median: median frequency (in kHz)
• Q25: first quantile (in kHz)
• Q75: third quantile (in kHz)
• IQR: interquantile range (in kHz)
• skew: skewness (see note in specprop description)
• kurt: kurtosis (see note in specprop description)
• sp.ent: spectral entropy
• sfm: spectral flatness
• mode: mode frequency
• centroid: frequency centroid (see specprop)
• peakf: peak frequency (frequency with highest energy)
• meanfun: average of fundamental frequency measured across acoustic signal
• minfun: minimum fundamental frequency measured across acoustic signal
• maxfun: maximum fundamental frequency measured across acoustic signal
• meandom: average of dominant frequency measured across acoustic signal
• mindom: minimum of dominant frequency measured across acoustic signal
• maxdom: maximum of dominant frequency measured across acoustic signal
• dfrange: range of dominant frequency measured across acoustic signal
• modindx: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
• label: male or female

Business benefits

Since "The Star RJ" is a reality show, time to select candidates is very short. The whole success of the show and hence the profits depends upon quick and smooth execution