Cosa ho fatto io:

Although SVM is used to classified data belonging to two classes, we developed an algorithm based on OVO (One Vs One) that allowed us to classify the diphthongs. The algorithm works with a L1O cross-validation. After the removing of a sample the dataset is divided in three subsets, each of that generated removing all the samples belonging to one class of diphthong. After that, for all the subsets has been tuned a model selecting the best value for the cost. Then we proceeded assigning the label to the sample removed at the beginning following this rule: get the output of the three models and assign the label that appear the most.

The algorithm is repeated until all the samples have been used for the test set. The result we obtained is 179 correct classifications and 7 missclassifications for a ratio of 0,96 correct labels assigned. The most interesting part is the fact that SVM didn’t use gaussianity assuption. Given the good results we thought that a classification based on gaussianity could be exploited anyway.

Vostro lavoro:

**GAUSSIANTY**

Obviously the data provided to us were far from gaussian. We tried to recover gaussianity using both a parametric algorithm (box-cox transformation) and mahalanobis distance. While the box-cox didn’t give us a satisfactory result (we knew that should be hard having 44 variables), mahalanobis help a little but the cost was to remove from the dataset a large number of the 186 samples. For this reasons we decided to go on without the assumption of gaussianity and exploit what has been got from mahalanobis only in case the following results wouldn’t be good enough.

**CLASSIFICATION**

Even if we didn’t get gaussianity we decided to try to use classical classifier to reach our main goal. Unfortunately, given the fact that the number of samples for diphthong “OY” were less then the dimesionality of the dataset, it wasn’t possible to use a QDA classifier. But seen the good results coming from SVM (that does not use gaussian assumption), we decided that the samples were classifiable anyway and so we tried with LDA. What we got was outstanding, only 7 smaples have been missclassified! We computed the value for the APER and the AER. The values obtained are: APER = 0.022 and AER = 0.054.

**CLUSTERING**

We also used unsupervised learning techniques. The reason was to know if it was possible to identify the three different groups also without looking at the label of each sample. Also in this case we reached our goal using kmeans. We tried also with hierarchical clustering, but we the results weren’t good, while with kmeans we have been able to split the dataset in three parts that contained the samples divided by diphthong. (non ho runnato il tuo codice perciò non so bene come siano gli errori in questa parte)

All this part gave us more evidences about the possibility of try to use supervised learning techniques without gaussianity, given the fact that samples are so distant between classes.