COMS 4721 Spring 2016: Homework #2

Peter Darche - pmd2139@columbia.edu
Discussants: none

March 2, 2016

Problem 1

The Perceptron convergence theorem stated in the homework can be thought of as the normalization of the statement given in class. Accordingly, by dividing $\boldsymbol{w}_*$ by $\|\boldsymbol{w}_*\|_2$ you get 1 or $\boldsymbol{u}_*$. Similarly, because we're normalizing, γ becomes $\frac{1}{\|\boldsymbol{w}_*\|_2}$ or the minimum margin. The shortest distance from a data point $\boldsymbol{x}$ to the homogeneous hyperplane with normal vector $\boldsymbol{w}_*$ is the minimum margin $\frac{1}{\|\boldsymbol{w}_*\|_2}$

Problem 2

A Because the generative model uses gaussian class conditional distributions, subtracting the mean will only serve to shift all of the distributions, but not change their probabilities, so centering will not affect the classifier.

B Like the generative model, centering the data will also not affect the Euclidean distance 1-NN classifier. This is because the algorithm only cares about the relative differences in Euclidean distances between points and these relative distances don't change when you change the mean. Again, changing the mean shifts the data, but it preserves the relative distances between points so the classifier is not affected.

C For decision trees, centering would not affect the classifier because the algorithm is looking for the point to maximally reduce uncertainty and so checks all the possible boundaries. Because the locations of boundaries aren't affected by shifting the data, the algorithm ins't affected.

D Because in EMR we care about the shape of the distribution, rather than the centering, changing the centering won't affect the classifier.

Problem 4

Cross-validation error rates for all methods:

Avg Perceptron	Log Regression	QDA	LDA	Avg Perceptron Exp	Log Regression Exp
0.1389	0.0815	0.1637	0.1076	0.3255	0.0756

Training error rate of the classifier learned by the selected method (and state which method was chosen):

Method	Training Error
Logistic Regression with expanded features	0.0456

Test error rate for the learned classifier:

Method	Testing Error
Logistic Regression with expanded features	0.0749