HW2 COMS 4721 – Spring 2016

Daniel Kost-Stephenson – dpk2124

March 1, 2016

Question 1

This theorem is the same as what was presented in the lecture because of the relationship between the length of the weight vector and the threshold. In the version of the Perceptron convergence theorem stated in the slides, the threshold is set as one and the norm of the weight vector, $\|\boldsymbol{w}_*\|$, is variable, but the version in the homework fixes the norm of the weight vector, $\|\boldsymbol{u}_*\|$ and varies the magnitude of the threshold, γ. The relationship between $\boldsymbol{w}_*$ and $\boldsymbol{u}_*$ is simply that $\boldsymbol{w}_*$ is a scaled version of $\boldsymbol{u}_*$ because $\boldsymbol{u}_*$ is the normalized weight vector. We can observe that the distance from a hyperplane with weight vector $\boldsymbol{w}_*$ to a point $\boldsymbol{x}$, is given by:

$$\|\boldsymbol{x}\| \cdot \cos\theta = \frac{\langle \boldsymbol{w}, \boldsymbol{x} \rangle}{\|\boldsymbol{w}\|}$$

We can clearly see that decreasing the length of the weight vector would increase the distance between the plane and a point, which is where the equation for our minimum margin stems from:

$$\frac{1}{\|\boldsymbol{w}_*\|}$$

The relationship between the weight vectors from the slides and the homework is simply a certain weight vector divided by its norm to achieve unit length:

$$\frac{\boldsymbol{w}_*}{\|\boldsymbol{w}_*\|} = \boldsymbol{u}_*$$

However, their direction is the same. The relationship between $\|\boldsymbol{w}_*\|$ and γ is inversely proportional; the length of the weight vector decreases as the threshold increases. Using the minimum margin defined above, we can observe that this is equivalent to the following for the perceptron convergence theorem in the homework:

$$\frac{1}{\|\boldsymbol{w}_*\|} = \gamma$$

The minimum distance between a data point $\boldsymbol{x}$, and a homogenous hyperplane with weigh vector $\boldsymbol{w}_*$ is simply 1.

Question 2

Part 1

In the case of a generative model with class conditional Gaussian distributions and a fixed covariance matrix equal to $\boldsymbol{I}$, the learned classifier is affected by centering or scaling. If MLE is used, the parameters are the sample mean and sample variance:

$$\hat{\mu} = \frac{1}{|S|}\sum_{(x,y)\in S} x$$

$$\sigma = \sqrt{\frac{1}{|S|}\sum_{(x,y)\in S}(x-\mu)^2}$$

Since the covariance is already $\boldsymbol{I}$, scaling by the *sample covariance* matrix will indeed affect the result. Centering the data will simply shift everything in space without changing the position of the observations relative to each other. This is a technicality, and the classifier would only be affected if we were scaling by the sample variance instead of the identity matrix. However, the problem states that the observations are scaled by the sample variance and so the learned classifier would be affected.

Part 2

The 1-NN classifier using Euclidean Distance is indeed sensitive to scaling and performing these operations would affect the learned classifier. If the observations are scaled, the relative distance between them will change which would lead to a different classifier because the decision boundary in d space would alter. To illustrate this, consider two distributions, $X \sim N(0,1)$ and $Y \sim N(0,10)$. If they are both scaled, then $X, Y \sim N(0,1)$ and the relative distance between the observations will change because they are concentrated together. Centering the data, however, would not have an effect on the learned classifier because the observations are simply being translated and the relative distance between them would not change.

Part 3

A greedy decision tree algorithm with axis aligned splits would not be affected by centering and scaling. By scaling the data, each dimension is shrunk towards the middle and the proportions between each observations will be unchanged, but the placement of the split along each axis will differ if the data is scaled or not. However, this would not affect the learned classifier because the splits would simply occur at different locations along each axes (e.g. $x > 4$ vs. $x > 0.7$), but the actual observations they separate would be the same. Centering the data would also not change the change the learned classifier because all points and their relative locations would simply be shifted towards the origin.

Part 4

A linear classifier using ERM would also be unaffected by centering and scaling. Scaling would not affect the classifier because the direction of the weight vector would not change but the threshold might (however the actual hyperplane would be the same). Again, centering would shift the data towards the origin so the learned weight vector would be the same as the weight vector learned by the non-centered threshold. Thus, centering and scaling does not affect the learned classifier.

Question 3

This question involved learning six different classifiers on the spam dataset. The six learned classifiers were: averaged perceptron (with 64 passes through the training data), logistic regression, linear discriminant analysis, quadratic discriminant analysis and averaged perceptron and logistic regression with an extended feature map that included quadratic interaction terms ($\mathbb{R}^{57} \rightarrow \mathbb{R}^{1710}$). 10-fold cross validation was performed on each of the six methods to determine the method with the lowest error rate and the results are displayed below:

Method	Error Rate
Averaged Perceptron	0.0773
Logistic Regression	0.0790
LDA	0.1067
QDA	0.1628
Averaged Perceptron w/ extended features	0.0887
Logistic Regression w/ extended features	0.0754

Although the error rates between the logistic regression and averaged perceptron classifiers are similar, logistic regression with an extended feature map was selected as the final classifier. The results of how it performed on the training and test data are displayed below:

Training Error	0.0127
Test Error	0.0755

It should be noted that the results varied from pass to pass meaning that sometimes the averaged perceptron function performed better but for this particular instance (the instance that was reported), the best results were from logistic regression with an extended feature map.