Homework 4
CS 7301: Advanced Machine Learning

Spring 2017

1 EM algorithm [25 points]

In this problem, you will use the EM algorithm and Gaussian mixture models to cluster the data into exactly three classes (or clusters). As discussed in class, in this approach, we assume a Gaussian mixture model with K components for the data and we find the parameters of the model using maximum likelihood. The equations for updating are provided in on the class Web page.

- Download the data from the class website.
- Assume that the number of clusters, K is 3. Implement the EM algorithm for 1-D Gaussian mixture models (assume that the data is an array of doubles or long doubles). Use the algorithm to cluster the given data. Develop several randomized initialization strategies. You can consult the Web for this. However, please cite your source. Run the algorithm multiple times from a number of different initialization points and pick the one that results in the highest log-likelihood (since EM in general only finds local maxima).

  Report the parameters you get for different initializations. Which initialization strategies did you use? How sensitive was the performance to the initial settings of parameters.
- Now assume that variance equals 1.0 for all the three clusters and you only have to estimate the means of the three clusters using EM. Report the parameters you get for different initializations. Which approach worked better, this one or the previous one?

What to turn in for this part:

- Your code. EM for general GMMs and EM for GMMs with known variance.
- A report containing answers to the questions above.

2 Support vector machines. (Extra Credit: 10 points)

- Download LIBSVM, currently the most widely used SVM implementation. Peruse the documentations to understand how to use it. (If you are comfortable with Weka, you can also use LIBSVM with WEKA. However, you may have to modify the data format so that WEKA is able to use it as input).
- Download the new promoters dataset in the LIBSVM format. (Available on the course web page).
- Run LIBSVM to classify promoters. Try different kernels (0-3) and use default parameters for everything else. How does it vary with different choice of kernel?
- Modify your perceptron classifier from homework 3 to apply to this dataset. How does the accuracy compare to what can be obtained using LIBSVM? Why? (If you are not confident about the correctness of your perceptron implementation, use the one in WEKA).

3 Boosting [50 points]

In this part of the homework, you will experiment with Bagging and Boosting. Bagging and AdaboostM1 are available under the ”Meta” category in WEKA. Use the following settings:

- Choose three classifiers of your choice. Example: J48, Logistic regression, Decision stump, etc. If you are using decision trees, turn pruning ON.
- Choose three datasets from the UCI machine learning repository. Note that the following set up does not work with all datasets available there and therefore you will have to choose the three quite carefully. The larger the data size the better (but be reasonable, don’t try very large datasets).
- For Bagging, set numIterations to 30.
- AdaboostM1: set maxIterations to 30. Set weightThreshold to 100000.

You will run the three classifiers by themselves (Vanilla) and then with bagging and boosting on each of the three datasets and report results for 10-fold cross validation in the following table:

```
Dataset1:
Base learner     Vanilla       Bagging        Boosting
Classifier1       xxx          yyy            zzz
Classifier2       xxx          yyy            zzz
Classifier3       xxx          yyy            zzz

Dataset2:
Base learner     Vanilla      Bagging        Boosting
Classifier1       xxx          yyy            zzz
Classifier2       xxx          yyy            zzz
Classifier3       xxx          yyy            zzz

Dataset3:
Base learner     Vanilla       Bagging        Boosting
Classifier1       xxx          yyy            zzz
Classifier2       xxx          yyy            zzz
Classifier3       xxx          yyy            zzz
```

where *xxx*, *yyy* and *zzz* are the error rates; replace them by the error rates that you get. Replace Classifier1, Classifier2, Classifier3 and Dataset1, Dataset2 and Dataset3 by the specific classifiers and datasets chosen.

Repeat the experiment for 2 other settings for number of iterations: 100 and 150.

Answer the following questions (5 points each):

1. Which algorithms+data set combination is improved by Bagging?

2. Which algorithms+data set combination is improved by Boosting?

3. Can you explain these results in terms of the bias and variance of the learning algorithms applied to these domains? Are some of the learning algorithms unbiased for some of the domains? Which ones?

4 K-means clustering on images [25 points]

In this problem, you will use K-means clustering for image compression. We have provided you with two images.

- Display the images after data compression using K-means clustering for different values of K (2, 5, 10, 15, 20).

- What are the compression ratios for different values of K? Note that you have to repeat the experiment multiple times with different initializations and report the average as well as variance in the compression ratio.
- Is there a tradeoff between image quality and degree of compression. What would be a good value of K for each of the two images?

We have provided you java template KMeans.java which implements various image input/output operations. You have to implement the function kmeans in the template. See the file for more details.

What to turn in: (in a single zip file)

1. Your code and datasets (in ARFF format)
2. A README for your compiling/using your code
3. A report (pdf or doc file) containing answers to the questions posed.