PPHA 30545 Machine Learning: Problem Set 4

1. The objective of this question is to practice fitting and interpreting an LDA model.

We will be using data from the Open Access Corpus of Scientific, Technical, and Medical Content (OA-STM) Corpus, put together by Elsevier Labs. The dataset contains 110 papers from STM disciplines, with more information available here: https://github.com/elsevierlabs/OA-STM-Corpus.

Be sure to download and save the full set of text files either from Canvas or from the original repository here: https://github.com/elsevierlabs/OA-STM-Corpus/tree/master/SimpleText/SimpleText_auto.

(a) Load the data as a corpus. For doing so, use the following command (leveraging the `tm` package):

```
library(tm)
texts <- file.path("[path]/SimpleText_auto") # add local file path!
docs <- VCorpus(DirSource(texts))
```

(b) Clean the data. This implies transforming all characters to lowercase and removing stop words, punctuation, and any other words that will not generate meaningful content for identifying the topics. Think about words that are likely common in academic papers (e.g., table, figure, results). Also think about combining forms of the same word (e.g., genes and gene). Be sure to justify your decisions.

(c) Present the 50 most frequently used words in the corpus in an informative way. This can include a table of results or a word cloud.

(d) Fit a topic model on the corpus setting k equal to 2, 3, 5, 8, and 10. Present the topics for each value of k and interpret the topics. In your opinion, which of the selected values of k yield the most meaningful coherence for each topic?

(e) Optimize the hyperparameters of the LDA model using 10-fold cross-validation. Present the topics from the best model and explain your results. Hint: The `topicmodels` package may be helpful.

2. We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form $\beta_0 + \beta X_1 + \beta_2 X_2 = 0$. We now investigate a non-linear decision boundary.

(a) Sketch the curve

$$(1 + X_1)^2 + (2 - X_2)^2 = 4.$$

(b) On your sketch, indicate the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

as well as the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 \leq 4.$$

(c) Suppose that a classifier assigns an observation to the blue class if

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

and the red class otherwise. To what class is the observation (0, 0) classified? (-1,1)? (2,2)? (3,8)?

(d) Argue that while the decision boundary in (c) is not linear in terms of X_1 and X_2, it is linear in terms of X_1, X_1^2, X_2, and X_2^2.

3. Generate a simulated two-class data set with 100 observations and two features in which there is a visible but non-linear separation between the two classes. Show that in this setting, a support vector machine with a polynomial kernel (with degree greater than 1) or a radial kernel will outperform a linear support vector classifier (i.e., with a trivial kernel that just keeps the inner product) on the training data. Which technique performs best on the test data? Make plots and report training and test error rates in order to back up your assertions.

Hint: (1) Create a random initial dataset which lies along the parabola $y = 3 * x^2 + 4$, then separate the two classes by translating them along the y-axis. (2) Create both training and testing data frames by taking half of positive and negative classes and creating a new `z` vector of `0` and `1` for classes. (3) Train a model.

4. In this problem, you will use support vector approaches in order to predict whether a given car gets high or low gas mileage based on the `Auto` data set.

(a) Create a binary variable that takes on a 1 for cars with gas mileage above the median, and a 0 for cars with gas mileage below the median.

(b) Fit a support vector classifier (with just the inner product as the kernel) to the data with various values of `cost`, in order to predict whether a car gets high or low gas mileage. Report the cross-validation errors associated with different values of this parameter. Comment on your results.

(c) Now repeat (b), this time using SVMs with radial and polynomial basis kernels, with different values of `gamma` and `degree` and `cost` (from the `e1071` R package). Comment on your results.

(d) Make some plots to back up your assertions in (b) and (c).

Hint: Use the `plot` function to create plots displaying pairs of variables at a time. For example, where `svmfit` contains your fitted model, and `dat` is a data frame containing your data, you can type: `plot(svmfit, dat, x1~x4)` in order to plot just the first and fourth variables. However, you must replace `x1` and `x4` with the correct variable names. To find out more, type `?plot.svm`.