Homework 2
CS 436/580L: Introduction to Machine Learning

Instructor: Arti Ramesh

Instructions

1. You can use either C/C++, Java or Python to implement your algorithms.
2. **Your implementations should compile on remote.cs.binghamton.edu.**
3. Make sure remote.cs.binghamton.edu has the packages that you require before starting to implement.
4. This homework requires you to **implement** the algorithms. Using existing packages for the algorithms is not allowed.
5. Your homework should contain the following components:
   (a) README.txt file with detailed instructions on how to compile and run the code.
   (b) Code source files
   (c) Type-written document containing the results on the datasets and answers to point estimation questions.
6. Refer to submission instructions on myCourses on naming conventions. If you fail to adhere with the submission instructions points **will** be deducted.

1 Naive Bayes for Text Classification

In this question, you will implement and evaluate Naive Bayes for text classification.

0 Points Download the spam/ham (ham is not spam) dataset available on myCourses. The data set is divided into two sets: training set and test set. The dataset was used in the Metsis et al. paper [1]. Each set has two directories: spam and ham. All files in the spam folders are spam messages and all files in the ham folder are legitimate (non spam) messages.

40 points Implement the multinomial Naive Bayes algorithm for text classification described here: `http://nlp.stanford.edu/IR-book/pdf/13bayes.pdf` (see Figure 13.2). Note that the algorithm uses add-one laplace smoothing. Ignore punctuation and special characters and normalize words by converting them to lower case, converting plural words to singular (i.e., "Here" and "here" are the same word, "pens" and "pen" are the same word). Normalize words by stemming them using an online stemmer such as `http://www.nltk.org/howto/stem.html`. Make sure that you do all the calculations in log-scale to avoid underflow. Use your algorithm to learn from the training set and report accuracy on the test set.

10 points Improve your Naive Bayes by throwing away (i.e., filtering out) stop words such as "the" "of" and "for" from all the documents. A list of stop words can be found here: `http://www.ranks.nl/stopwords`. Report accuracy for Naive Bayes for this filtered set. Does the accuracy improve? Explain why the accuracy improves or why it does not?

2 Point Estimation

20 points Derive maximum likelihood estimators for

1. parameter p, Bernoulli(p) sample of size n.
2. parameter p based on a Binomial(N, p) sample of size n. Compute your estimators if the observed sample is (3, 6, 2, 0, 0, 3) and N = 10.
3. parameters a and b based on a Uniform (a, b) sample of size n.
4. parameter μ based on a Normal(μ, σ^2) sample of size n with known variance σ^2 and unknown mean μ.
5. parameter σ based on a Normal(μ, σ^2) sample of size n with known mean μ and unknown variance σ^2.
6. parameters (μ, σ^2) based on a Normal(μ, σ^2) sample of size n with unknown mean μ and variance σ^2.

30 points You are given a coin and a thumbtack and you perform the following experiment: toss both the thumbtack and the coin 100 times. You get 60 heads and 40 tails for the coin, 70 heads and 30 tails for the thumbtack. You put Beta priors Beta(1,1), Beta(40, 60), Beta(30, 70), Beta(100, 100), Beta(1000, 1000), and Beta(100,000, 100,000) on the coin and thumbtack, respectively. (For both the coin toss and thumbtack toss experiments, you put these three priors, respectively.)

1. Derive the MLE and MAP estimates for the coin and the thumbtack.
2. With the help of figures identify how the different priors affect the estimated parameter values. Follow the point estimation example in the lectures and illustrate in a similar manner in your answer. For

full credit, a curve for each scenario should be shown along with an explanation of the curve in terms of the different values in the question (number of heads and tails recorded and parameters of the Beta prior). Record changes in the curve for the different combinations and include intuitive explanations for the changes.

3. True or False: As you have collect more data instances by tossing the coin/thumbtack, the MLE estimate will approach the MAP estimate. Explain. [Answers without explanation will receive no credit.]
4. True or False: The MLE of the coin and the thumbtack are different but their MAP estimates approach the same value when we use a larger prior (think Beta(100,000, 100,000) and larger). Explain. [Answers without explanation will receive no credit.]

What to Turn in

- Your code
- README file for compiling and executing your code.
- A detailed write up that contains:
  1. The accuracy on the test set.
  2. Detailed answers to point estimation questions showing the complete derivation (can be hand-written and scanned). No points will be given for answers that are incomplete.

References

[1] V. Metsis, I. Androutsopoulos and G. Paliouras, "Spam Filtering with Naive Bayes - Which Naive Bayes?". Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006), Mountain View, CA, USA, 2006.