CS 445/545

Machine Learning

Winter 2016

Homework 4: Naive Bayes Classification

Due Thursday Feb. 25, 2016

In this homework you will use Gaussian Naïve Bayes to classify the Spambase data from the UCI ML repository (the same dataset you worked with in Homework 3).

For this homework, use all the data (4,601 instances). The full data set has approximately 40% spam, 60% not-spam.

1. Create training and test set:
Split the data into a training and test set. Each of these should have about 2,300 instances, and each should have about 40% spam, 60% not-spam, to reflect the statistics of the full data set.

2. Create probabilistic model. (Write your own code to do this.)

- Compute the prior probability for each class, 1 (spam) and 0 (not-spam) in the training data. As described in part 1, $P(1)$ should be about 0.4.

- For each of the 57 features, compute the mean and standard deviation in the training set of the values given each class.

C. Run Naïve Bayes on the test data. (Write your own code to do this.)

- Use the Naïve Bayes algorithm to classify the instances in your test set, using

$$P(x_i \mid c_j) = N(x_i; \mu_{i,c_j}, \sigma_{i,c_j})$$

where

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Because a product of 58 probabilities will be very small, we will instead use the log of the product. Recall that the classification method is:

$$class_{NB}(\mathbf{x}) = \underset{class}{\operatorname{argmax}} \left[P(class) \prod_i P(x_i \mid class) \right]$$

Since

$$\underset{z}{\operatorname{argmax}}\, f(z) = \underset{z}{\operatorname{argmax}} \log f(z)$$

we have:

$$class_{NB}(\mathbf{x}) = \underset{class}{\operatorname{argmax}} \log \left[P(class) \prod_i P(x_i \mid class) \right]$$
$$= \underset{class}{\operatorname{argmax}} \left[\log P(class) + \log P(x_1 \mid class) + \ldots + \log P(x_n \mid class) \right]$$

In your report, include a short description of what you did, and your results: the accuracy, precision, and recall on the test set, as well as a confusion matrix for the test set. Write a few sentences describing your results, and answer these questions: Do you think the attributes here are independent, as assumed by Naïve Bayes? Does Naïve Bayes do well on this problem in spite of the independence assumption? Speculate on other reasons Naïve Bayes might do well or poorly on this problem.

Here is what you need to turn in:

- Your report (just needs to be a paragraph, along with accuracy, precision, recall, and confusion matrix).
- Your well-commented code.

How to turn it in (read carefully!):

- Send these items in electronic format to mm@pdx.edu by 2pm on the due date. No hard copy please!
- The report should be in pdf format and the code should be in plain-text format.
- Put "MACHINE LEARNING HW 4" in the subject line.

If there are any questions, don't hesitate to ask me or e-mail the class mailing list.

Policy on late homework: If you are having trouble completing the assignment on time for any reason, please see me before the due date to find out if you can get an extension. Any homework turned in late without an extension from me will have 5% of the grade subtracted for each day the assignment is late.