* **Preprocess: (missing features on data set)**
* Method

Method1:

* Overview

We used N-fold cross validation to partition the labeled 198 instances into training set and testing set. The Gaussian Naïve Bayes was implemented to train a classifier based on training set, ending up with trained MLE parameters, such as class frequency, mean, and covariance matrix. Classifying testing set instances with Naïve Bayes classifier, we could get testing errors and confusion matrix for the evaluation.

* Specification:

Naïve Bayes.pdf page 2-3, algorithm part.

Method2:

* Introduction:

Semi-supervised learning is a class of [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) tasks and techniques that also make use of unlabeled [data](https://en.wikipedia.org/wiki/Data) for training – typically a small amount of [labeled data](https://en.wikipedia.org/w/index.php?title=Labeled_data&action=edit&redlink=1) with a large amount of unlabeled data.

The learner has both labeled training dataand unlabeled training data , and learns a predictor Where F is the hypothesis space. The predictor learned by semi-supervised methods usually predicts future test data better than that learned by supervised learning which only considers the labeled training data.

* Motivation:

In real world applications, labeled data are relatively hard to get while unlabeled data are cheap. The target label {y} requires human annotation which takes a long time, and these experiments require a lot of resources including experienced experts and special devices. Therefore, a trend to utilize the surplus unlabeled data together with scarce labeled data is desirable.

* Learning Approach:

Self-Training Algorithm

1. Given a small amount of initial labeled training data, the learner trained a classifier using the Gaussian Naïve Bayes model mentioned above and obtained a MLE .
2. Apply the classifier on the unlabeled data  and assign the target label {y’} for them by computing the posterior probability .
3. Augment the original training data with those new labeled data, and then we got.
4. Repeat step 1 to obtain the new MLE 

* **Result**
* **Evaluation:**

We checked the training errors and testing errors respectively, and calculated the expected accuracy. In addition, we looked at confusion matrix and account for sensitivity and specificity as well. Compare the result between Naïve Bayes classifier and Semi-supervised learning one.

References

[1] Diana Dumitru. ”Prediction of recurrent events in breast cancer using the Naive Bayesian classification”

Annals of University of Craiova (2009), Vol 36(2).

[2] Chapelle, Olivier; Schlkopf, Bernhard; Zien, Alexander (2006). Semi-supervised learning. Cambridge,

Mass.: MIT Press.ISBN978-0-262-03358-9.

[3] Zhu, Xiaojin. [Semi-Supervised Learning](http://pages.cs.wisc.edu/~jerryzhu/pub/SSL_EoML.pdf) University of Wisconsin-Madison.

[4] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2/3):103–134, 2000.

More specifics about 3.2.3 would be useful. Why were these not good? Why did you use these to begin with? Was this tried in previous literature? What sort of results did they get? (Points deducted in rubric)