Brian Liu
Professor Zhuowen Tu
COGS 118A
16 December 2023

Comparisons Between Classification Methods

a) Abstract

There are many supervised learning methods, we are going to compare and evaluate some of them here including : logistic regression, bagged-logistic regression, support vector matrix, bagged support vector matrix, decision tree, bagged decision tree, naive bayes, and lastly calibrated naive bayes. In this experiment we experiment with the effects of classifiers and the bagged counterparts as well as naive bayes and the effects of dataset calibration on the classifier. In An Empirical Evaluation of Supervised Learning in High Dimensions it is shown that bagged classifiers and calibrated classifiers perform better, so we are testing if that holds true with our datasets.

b) Introduction

We will be exploring the performance of different classifiers on three different dataset. The first dataset being the adult dataset which deals with a balanced type of features of both categorical and empirical, where we will try to determine if given the features if we are able to correctly classify whether the adults make above 50k salary. Next will be the abalone dataset, which is only empirical data and we will be seeing if we could detect whether it is a male or female given the features. Lastly will be about mushrooms which are full of only categorical data and we will be seeing if we can correctly classify if the mushroom is edible or not. In this study we will be using different classifiers on these three datasets and we will be testing for multiple ideas. If bagging makes an improvement over our base classifiers, and which base classifier does the best amongst our base classifiers on empirical data, categorical data, and a dataset containing both, as well which bagged classifier does the best. As well test if naive-bayes performs better if our dataset is calibrated as in “An Empirical Evaluation of Supervised Learning in High Dimensions”, Caruana states that calibrated can affect data in a negative way for classifiers such as logistic regression, but their results also suggested that calibration doesn’t really affect their findings and “never hurts”.

c) Methodology

Training Procedure

We will be performing the following learning algorithms alongside their bagged equivalent. We will be using sklearn to perform our classifiers. In each experiment, we will be using datasets of size 20/80, 50/50, 80/20. In each experiment we will be using 3-fold cross validation as well and we will be reporting the accuracy score and f1 score of the found best hyper parameters and tuning.

Learning Algorithms

- Logistic Regression (LR)
    - We will be using LogisticRegression(solver = "liblinear", class_weight = "balanced") from sklearn with regularization terms [.001,.01,.1,1,10].
- Bagged Logistic Regression (B-LR)
    - We will be using BaggingClassifier(estimator=log_reg, n_estimators=10,random_state=0) with n=10 as our estimators and random state = 0.
- Linear Support Vector Machine (SVM)
    - We will be using LinearSVC(class_weight="balanced")
- Bagged Linear Support Vector Machine (B-SVM)
    - We will be using BaggingClassifier(estimator=LinearSVC(class_weight="balanced"), n_estimators=10,random_state=0) with n=10 as our estimators and random state = 0.
- Decision Trees (DT)

        - We will simply use DecisionTreeClassifier(random_state=0, class_weight="balanced") with the random state being at 0 and we will fix the maximum depth parameter from the range of 3 to 20 and perform 4 fold cross validation.
    - Bagged Decision Tree (B-DT)
        - We will just do the same thing, but bagged and with n=10 as our estimator count.
    - Naive Bayes (NB)
        - We will simply use gaussian naive bayes
    - Calibrated Naive Bayes (C-NB)
        - We will calibrate our gaussian naive bayes with isotonic regression

Datasets

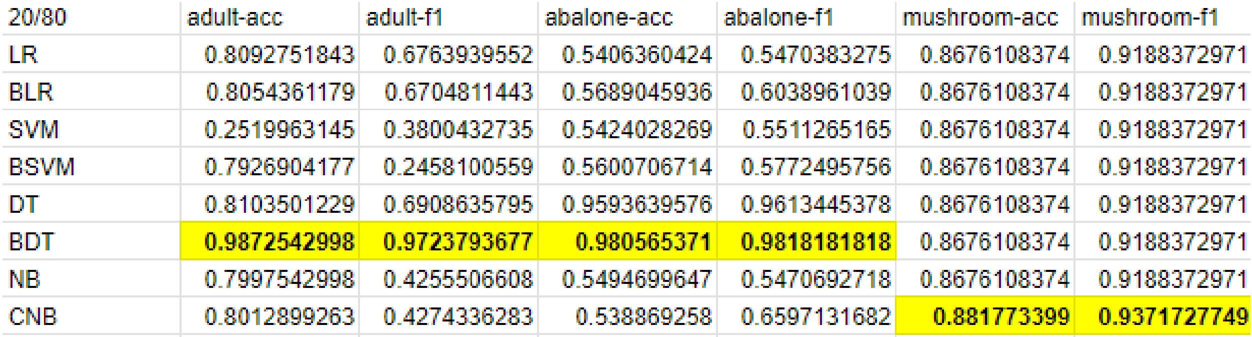
We will be dealing with 3 datasets : adult, abalone, and mushrooms. We have chosen these 3 datasets based on the features they provided, because mushrooms only contain categorical data, abalone contains only empirical data, and adult contains a mix of both. We are interested in this because we are wondering if the type of data will play a difference in affecting our classifier. Though we won't be able to make any conclusions, it's just a fun thought experiment that we can do alongside our current experiment.

- **Adult**
    - This is a dataset of size 32560 with features : age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, relationship, capital-gain, capital-loss, hours-per-week, native-country, income. We will be using the following features to determine if given the features if we can predict whether the adult makes <=50k or >50k yearly.
- **Mushroom**
    - This is a dataset of size 32560 with features: 'cap-shape', 'cap-surface', 'cap-color', 'bruises', 'odor', 'gill-attachment', 'gill-spacing', 'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'stalk-surface-above-ring', 'stalk-surface-below-ring', 'stalk-color-above-ring', 'stalk-color-below-ring',

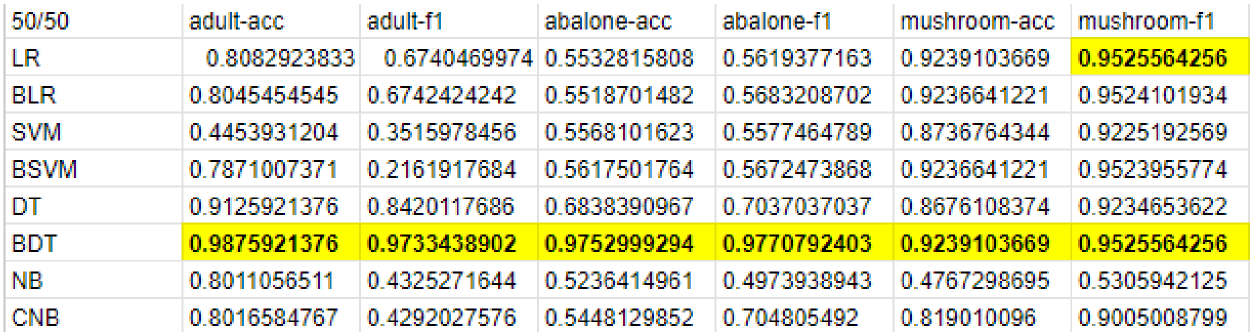
    'veil-type', 'veil-color', 'ring-number', 'ring-type', 'spore-print-color', 'population', 'habitat', 'edible'. We will be using the following features to determine if the mushroom is edible or poisonous from the Agaricus and Lepiota Family. As well this dataset is determined only with categorical data. Each species is labeled with varying edibility, but we parse the dataset to either edible or not edible. We will clean this data set and encode the categorical data using one-hot encoding. Surprisingly, with this many categorical data, it is very easy to accurately predict (100% accuracy) for nearly every classifier if a mushroom is edible or not, so I've decided to drop a lot of the columns to only ['spore-print-color', 'population', 'habitat'] to make things interesting and allow me to experience one-hot encoding.

- **Abalone**
    - This is a dataset of size 2834 with features: 'Length', 'Diameter', 'Height', 'Whole weight', 'Shucked weight', 'Viscera weight', 'Shell weight', 'Rings', 'Sex'. We will be using the following features to determine the sex of the abalone. As well this dataset is determined only with empirical data aside from its sex. Each abalone is also labeled is either male, female or infant as infants are not sex discernable so we will be removing the infants and examining only the male and female abalones.

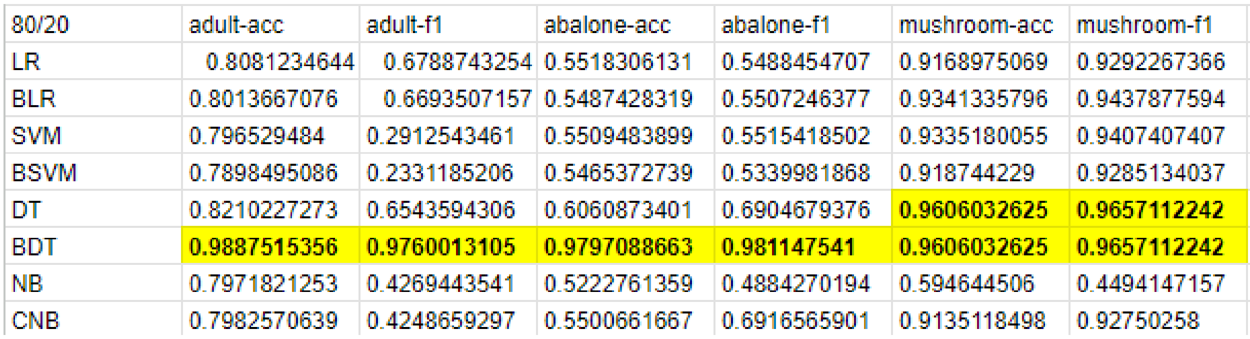
d) experiment



20/80	adult-acc	adult-f1	abalone-acc	abalone-f1	mushroom-acc	mushroom-f1
LR	0.8092751843	0.6763939552	0.5406360424	0.5470383275	0.8676108374	0.9188372971
BLR	0.8054361179	0.6704811443	0.5689045936	0.6038961039	0.8676108374	0.9188372971
SVM	0.2519963145	0.3800432735	0.5424028269	0.5511265165	0.8676108374	0.9188372971
BSVM	0.7926904177	0.2458100559	0.5600706714	0.5772495756	0.8676108374	0.9188372971
DT	0.8103501229	0.6908635795	0.9593639576	0.9613445378	0.8676108374	0.9188372971
BDT	**0.9872542998**	**0.9723793677**	**0.980565371**	**0.9818181818**	0.8676108374	0.9188372971
NB	0.7997542998	0.4255506608	0.5494699647	0.5470692718	0.8676108374	0.9188372971
CNB	0.8012899263	0.4274336283	0.538869258	0.6597131682	**0.881773399**	**0.9371727749**



50/50	adult-acc	adult-f1	abalone-acc	abalone-f1	mushroom-acc	mushroom-f1
LR	0.8082923833	0.6740469974	0.5532815808	0.5619377163	0.9239103669	**0.9525564256**
BLR	0.8045454545	0.6742424242	0.5518701482	0.5683208702	0.9236641221	0.9524101934
SVM	0.4453931204	0.3515978456	0.5568101623	0.5577464789	0.8736764344	0.9225192569
BSVM	0.7871007371	0.2161917684	0.5617501764	0.5672473868	0.9236641221	0.9523955774
DT	0.9125921376	0.8420117686	0.6838390967	0.7037037037	0.8676108374	0.9234653622
BDT	**0.9875921376**	**0.9733438902**	**0.9752999294**	**0.9770792403**	**0.9239103669**	**0.9525564256**
NB	0.8011056511	0.4325271644	0.5236414961	0.4973938943	0.4767298695	0.5305942125
CNB	0.8016584767	0.4292027576	0.5448129852	0.704805492	0.819010096	0.9005008799



80/20	adult-acc	adult-f1	abalone-acc	abalone-f1	mushroom-acc	mushroom-f1
LR	0.8081234644	0.6788743254	0.5518306131	0.5488454707	0.9168975069	0.9292267366
BLR	0.8013667076	0.6693507157	0.5487428319	0.5507246377	0.9341335796	0.9437877594
SVM	0.796529484	0.2912543461	0.5509483899	0.5515418502	0.9335180055	0.9407407407
BSVM	0.7898495086	0.2331185206	0.5465372739	0.5339981868	0.918744229	0.9285134037
DT	0.8210227273	0.6543594306	0.6060873401	0.6904679376	**0.9606032625**	**0.9657112242**
BDT	**0.9887515356**	**0.9760013105**	**0.9797088663**	**0.981147541**	**0.9606032625**	**0.9657112242**
NB	0.7971821253	0.4269443541	0.5222761359	0.4884270194	0.594644506	0.4494147157
CNB	0.7982570639	0.4248659297	0.5500661667	0.6916565901	0.9135118498	0.92750258

Surprisingly, I was able to get very accurate predictions, even with varying dataset sizes with my bag decision tree, which lines up with Caruana's paper as they have DT as one of the better classifiers. It was also interesting seeing some classifiers equal to each other, which on further research seems to happen when you are using binary featured data, which is mostly what the mushroom dataset as it is categorical according to Dr Alim-Marvasti in their article “Converging Support Vector Classifiers and Logical Regression”. As well it could be the fact that our 20/80 dataset wasn't large enough as this dataset was small and simplistic enough where if I didn't remove any categories it would 100% accurately predict if a mushroom was edible. So either this dataset was too easy or too small, it caused this kind of result to occur. As well, we could definitely see that NB did perform better when it is calibrated which isn't surprising to see as it matches what Caruana's paper mentions and same thing with bagging where it does improve the accuracy.

e) conclusion

In conclusion, we can see that our data agrees with Caruana's paper and it does seem that Bag-DT is one of the better performing classifiers. As well you can definitely see the underfitting and overfitting problem where you can see the training data and testing data mean change as you change datasizes.It was also surprising to see decision trees perform well on some of the datasets, because it was also one of the poorer performing classifier in Caruana's paper. It was surprising to see CNB perform as it didn't perform as well in their paper. In the future what I would do better is perform more of these predictions on more datasets, because 3 datasets can't really tell you much. Also we saw that calibrating NB does indeed increase the performance and bagging does also increase the performance of the classifier, which also agrees with the paper as they claimed bagging and calibration doesn't really hurt to do.

f) references.

Caruana, Rich, and Alexandru Niculescu-Mizil. "An empirical comparison of supervised learning algorithms." *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 2006, https://doi.org/10.1145/1143844.1143865.

https://scikit-learn.org/stable/index.html

https://towardsdatascience.com/scikit-learn-1-1-comes-with-an-improved-onehotencoder-5a1f939da190

https://towardsdatascience.com/support-vector-classifiers-and-logistic-regression-similarity-97ff06aa6ec3

g) bonus points

I think I deserve just a little bonus points for implementing one-hot encoding for my mushroom dataset. It was difficult to do , but fun to learn and for the amount of classifiers I did for the fun of it.