ISSS610
Applied Machine Learning

Assignment 2

Ensemble Learning

[Question 1] Compare Logistic Regression and Naïve Bayes Classifiers

(1) In this assignment, you are given the lending club dataset on the approve/reject classification of personal loans. You are also given the data pre-processing part in .ipynb format to start with. In this question, you are to build two binary classifiers, one is a logistic regression classifier and the other is a naïve Bayes classifier. Justify your choice of β (not more than 100 words), compare the two classifiers and report their evaluation metrics in F_β and AUC under ROC in the table below (code required, 6 marks). [Hint: the threshold should be chosen to maximize F_β for a fixed β.]

	Logistic Regression with C = __	Naïve Bayes
Precision		
Recall		
Chosen beta		
F_β		
AUC under ROC		

[Question 2] Classification on Clusters

(2) It is sometimes possible to do a clustering before classification, as clustering is a grouping of the samples. In this question, let us train Gaussian mixture models before training logistic regression and naïve Bayes classifier. For a sample $\boldsymbol{x}$, let $p(k|\boldsymbol{x})$ be the posterior probability distribution of $\boldsymbol{x}$ over all clusters $k \in K$, where K is the set of all clusters. Classification models are then trained for each cluster with samples with posterior probability higher than average, e.g., higher than 0.2 if there are 5 clusters. With trained classification models, $p(C|k, \boldsymbol{x})$ – the class probability distribution of sample $\boldsymbol{x}$ can be computed as if it is clustered to cluster k, for all $\boldsymbol{x}$. Therefore, the class probability $p(C|\boldsymbol{x})$ can be obtained as follow

$$p(C|\boldsymbol{x}) = \sum_{k \in K} p(C|k, \boldsymbol{x}) p(k|\boldsymbol{x})$$

Use 5, 10, 20 clusters to build 3 GMM clustering models, and then build 6 classifiers for each pair of a clustering model and one classification model chosen from logistic regression or naïve Bayes classifiers. Report their evaluation metrics using the same beta in the table below (code required, 6 marks).

		Logistic Regression with C = __	Naïve Bayes
5 clusters	Precision		
	Recall		
	F_β		
	AUC		
10 clusters	Precision		
	Recall		
	F_β		
	AUC		
20 clusters	Precision		
	Recall		
	F_β		
	AUC		

[Question 3] Ensemble Learning

(3) Build a bagging model for the approve/reject classification (code required, 3 marks).

(4) Build a boosting model for the approve/reject classification (code required, 3 marks).

(5) Build a stacking model for the approve/reject classification (code required, 3 marks).

(6) Compare the performance of the three models using the same beta, and give your short comments (not more than 200 words, 4 marks).

	Bagging	Boosting	Stacking
Base classifier			
Precision			
Recall			
F_β			
AUC			

[Question 4] Ensemble Learning with External Data

You are given some external data of income bracket indicators for each region of 5-digits zip code. In our original data, we have only the first three digits of zip code, which is at county level, rather than town level. Therefore, we need to aggregate the income brackets to estimate the income of loan applicants. Build ensemble models again and see if there is any improvement.

(7) Build a bagging model for the approve/reject classification (code required, 3 marks).

(8) Build a boosting model for the approve/reject classification (code required, 3 marks).

(9) Build a stacking model for the approve/reject classification (code required, 3 marks).

(10) Compare the performance of the three models using the same beta, and give your short comments (not more than 200 words, 4 marks).

	Bagging	Boosting	Stacking
Base classifier			
Precision			
Recall			
F_β			
AUC			