Machine Learning with Domestic Energy Use Data

Sam Stern (s1134468)

March 24, 2015

Abstract

Contents

Chapter 1

Results

This section discusses the quantitative evaluation methods used to determine the potential for each of the classifiers to reveal household characteristics and then analyses the results from training and running each classifier.

1.1 Evaluation Methods

For each classifier, a *confusion matrix* (CM) is produced using the MATLAB tool `confusionmat`, which, for a K class classification problem, returns a $K \times K$ matrix where each element (i, j) contains the number of times an instance of class i has been classified as j. The diagonal elements elements of CM contain the number of instances of households that have been classified correctly for each class. [1]

The accuracy of a classifier is defined as the sum of the diagonal elements of CM, divided by the total number of samples,S.

$$ACC = \frac{\sum_{i=1}^{K} CM_{i,i}}{S}$$

This is compared to the accuracy of performing a random guess (RG), which assigns a household to one of the K classes at random.

$$ACC_{RG} = \frac{1}{K}$$

To account for the imbalances in classes, we also calculate the most probable class (MPC) which uses knowledge of the prior probability of each class in the training data to find a baseline by assigning all samples to the most probable class.

$$ACC_{MPC} = \frac{argmax(S^K)}{S}$$

where S^K is the number of samples from the test data that are in class K.

For socio-economic classification problem, the ordinal structure of the classes should also be taken into account i.e it is worse for our classifier to predict a household of social grade B as D, then it is to predict it as C1 or A. Therefore, the *accuracy within* n[2].

Particularly for unbalanced classes, reporting the accuracy alone is not satisfactory in determining the quality of a classifier. The obvious and well known example being; constructing a classification problem where 99% of instances are in class A and only 1% in class B. A classifier that simply predicts all new data as class A would be correct 99% of the time, but would still not be a good classifier.

A widely applied method for evaluating a classifier is to compute the *true positive rate* (TPR) and *true negative rate*(TNR). The TPR gives the proportion of positives that are correctly identified as being positive, while the TNR gives the porportion of negatives that are correctly identified as negative.

$$TPR = \frac{TP}{TP + FN} \qquad\qquad TNR = \frac{TN}{TN + FP}$$

From these statistics, it is common to plot an ROC curve, which is a plot of the TPR against the *false positive rate*(FPR), which is defined as 1-TNR. The evaluation criterion (the area under the ROC curve) is preferred over the accuracy, particularly when considering unbalanced classes as the impact of skewness can be analysed [3]. To create the ROC curve, a value is found for each classifier which acts as the threshold above which an instance is classified as positive. Typically for logistic regression, this is the probability of an instance being assigned to class 1.

This is not as straight forward for random forests and knn as they are not probabilistic classifiers. Probabilities can, however be generated from the classifier results. For random forest the decision boundary may be the ratio of number of trees that vote in favor of assigning an unseen instance to class 1 and the total number of trees. In knn it is the number of nearest neighbors that are of class 1 divided by the total number of nearest neighbors.

In computing the ROC curve to evaluate the binary classification task of discriminating between households with and without children is straight forward, it is straightforward te determine which class is 'positive' and which is 'negative' (has children is positive). However for multi-class classification it is unclear what is 'positive' and what is 'negative'. When evaluating their socio-economic classifier, Beckel et. al. group nearby groups together and then use a one-versus-all approach[4, 5]. A similar method is used, analogous to the *accuracy within n* method described above, where classes within n are considered positive and all else are negative.

The final metric that is presented is the Matthews correlation coefficient (MCC) which is a value between -1 and +1 representing the correlation between the predicted and true values of a binary classifier. A MCC of -1 indicates that that there is no correlation between the predicted a true class while a value of +1 would indicate a perfect classifier while a value of 0 means it is no better or worse than a random guess. MCC is a worthwhile metric to report as it gives a value to the performance without inflating the imbalances in class sizes [6]. The MCC can be calculated using the formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

1.2 Feature Selection

As explained in section **??**, SFS was used to determine which features are of greatest value for each classifier. What was noticed, however, is that when running the SFS algorithm multiple times, the selected features were not always the same for each of the classifiers (even with cross validation). This was seen particularly in random forest (as might be expected for an algorithm that uses bootstrap aggregation). Therefore, the feature selection algorithm was run multiple times and the features that appeared most often were used. To evaluate the feature selection method, two additional sets of features were made (one for each classification problem) by choosing features based on how they appeared to seperate classes in **??**. All classification models are evaluated using the features found by SFS as well as the features found manually. The lists of features used for each classifier are outlined in tables 1.1 and 1.2.

Children

SFS			**Manual**
Log Reg	**KNN**	**Rand Forest**	
Monday Daytime	Monday Evening	Thursday Total	Month Total
Tuesday Evening	Monday Night	Sunday Night	Sunday Daytime
Wednesday Night	Wednesday Evening	Monday Morning	Saturday Evening
Thursday Daytime	Friday Morning	Monday Daytime	Thursday Variance
Friday Morning	Saturday/Weekday Ratio	Monday Evening	Monday Morning
Tuesday Variance	Sunday Variance	Tuesday Daytime	Friday Evening
Thursday Variance	Monday Variance	Friday Night	Saturday Variance
Saturday Variance	ρ(Monday Thursday)	Thursday Variance	Monday Total
ρ(Monday Tuesday)	ρ(Monday Friday)	Friday Variance	Saturday Total
ρ(Wednesday Thursday)	ρ(Tuesday Wednesday)	ρ(Monday Wednesday)	ρ(Monday Tuesday)

Table 1.1

Social Grade

	SFS			**Manual**
Ord Log Reg	**Nom Log Reg**	**KNN**	**Rand Forest**	
Monday Morning	Friday Total	Friday Total	Sunday Total	Monday Night
Tuesday Morning	Monday Morning	Tuesday Morning	Wednesday Total	Tuesday Variance
Tuesday Daytime	Tuesday Morning	Wednesday Night	Monday Morning	Monday Total
Friday Night	Tuesday Daytime	Thursday Morning	Monday Night	Tuesday day
Friday Variance	Wednesday Morning	Friday Morning	Wednesday Morning	Sunday Night
Saturday Variance	Wednesday Daytime	Saturday Evening	Wednesday Evening	Thursday Night
ρ(Monday Tuesday)	Wednesday Evening	Sunday Total	Friday Night	ρ(Tuesday,Friday)
ρ(Monday Friday)	Fridayday Morning	Thursday Variance	Thursday Variance	Thursday Total
ρ(Wednesday Fridayday)	Friday Night	Friday Variance	ρ(Monday Friday)	Saturday Night
First Fourier Feature	Saturday Daytime	ρ(Wednesday, Thursday)	First Fourier Feature	Friday Evening

Table 1.2

1.3 Classifiers

Here the results of the classifiers are presented of running each model on unseen data as outlined in 1.1.

1.3.1 Discriminating Between Households With and Without Children

Bibliography

[1] JERZY STEFANOWSKI. Data mining - evaluation of classifiers. Poznan University of Technology.

[2] Lisa Gaudette and Nathalie Japkowicz. title = Evaluation Methods for Ordinal Classification,. In *Advances in Artificial Intelligence*.

[3] Willem Waegeman, Bernard De Baets, and Luc Boullart. Roc analysis in ordinal regression learning. *Pattern Recognition Letters*, 29(1):1–9, 2008.

[4] Christian Beckel, Leyna Sadamori, and Silvia Santini. *Towards automatic classifi-cation of private households using electricity consumption data*, pages 75–86. ACM, 2013.

[5] Christian Beckel, Leyna Sadamori, Thorsten Staake, and Silvia Santini. Revealing household characteristics from smart meter data. *Energy*, 78:397–410, 2014.

[6] David M W Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation. Technical report, University of South Australia, 2007.