Project 1: Machine Learning Pipeline

Sarah Wilson SAWI117@JHU.EDU
303-921-7225
Engineering Professionals Computer Science
Johns Hopkins University
Baltimore, MD 21218, USA

1. Introduction

Regression and classification are both common tasks in the realm of Machine Learning. Regression and classification are both supervised learning problems. Supervised learning is where the system is given an input and output and then ask to learn or predict the mapping of input to out.

Regression is used to solve problems where the outcome is a number. An example of a Regression problem would be if a system needed to be built that would predict the price of a car based off of certain attributes of that car, such as, mileage, accident history and age.

Classification is used to solve problems where the outcome is a classifier or string. An example of a Classification problem would be if a system needed to built that would predict if a loan was 'high' or 'low' risk, based off of certain attributes of the person applying for that loan, such as, credit score, previous loan history and income.

In order to implement Regression and Classification algorithms it first must be noted that there needs to be data for these algorithms to run on. A crucial component in Machine Learning is the pre-processing of the data sets that the algorithms are intended to run on. The primary motivation behind this project was to develop a Machine Learning Pipeline that could be used to pre- process multiple unique data sets in order to pass the data to the algorithms. Due to the fact the primary objective was proper data handling the only algorithms that will be discussed in this report are: for Classification problems a Naive Majority predictor and for Regression problems the mean of an attribute in the data sets. These algorithms will be evaluated by using the k-fold cross validation method.

The algorithms implemented for both Classification and Regression are very simple. This leads to the hypothesis that the results from these simplistic algorithms will be highly inaccurate and produce large errors.

Section 2 will discuss more examples of ways that data needs be pre-proccesed before entering the algorithms, the data-sets that were leverages, the algorithms themselves and the k-fold cross validation method. Section 3 will present the results obtained by the Classification task using a Naive Majority predictor and for Regression task taking the mean of an attribute in the data sets. Section 4 will discuss the result that were obtain and compare that to the hypothesis that was outlined in the introduction. This report will conclude in Section 5 with a discussion of lessons learned and areas of possible future work.

2. Algorithms and Experimental Methods

INSERT

Data Sets The following data sets were used during the classification and regression tasks for this project. TABLE X provides a description

Table 1: Data Sets

Set Name	**Description**	**Task Type**	**Predictor**	**Link**
Breast Cancer	Descp	Classification	Diagnosis	Link
Car Evaluation	Descp	Classification	Car Eval.	Link
Congressional Vote	Descp	Classification	Party	Link
Albalone	Descp	Regression	Rings (int)	Link
Computer Hardware	Descp	Regression	PRP (int)	Link
Forest Fires	Descp	Regression	Area (float)	Link

3. Results

The following results were obtained from the Classificaiton task data sets.

Table 2: Classification: Naive Majority Predictor Results

Value 1	**Value 2**	**Value 3**
1	1110.1	a
2	10.1	b
3	23.113231	c

4. Discussion

5. Conclusion

Probabilistic inference has become a core technology in AI, largely due to developments in graph-theoretic methods for the representation and manipulation of complex probability distributions (**?**). Whether in their guise as directed graphs (Bayesian networks) or as undirected graphs (Markov random fields), *probabilistic graphical models* have a number of virtues as representations of uncertainty and as inference engines. Graphical models allow a separation between qualitative, structural aspects of uncertain knowledge and the quantitative, parametric aspects of uncertainty...

Appendix A.

In this appendix we prove the following theorem from Section 6.2:
Theorem *Let u, v, w be discrete variables such that v, w do not co-occur with u (i.e., $u \neq 0 \;\Rightarrow\; v = w = 0$ in a given dataset $\mathcal{D}$). Let N_{v0}, N_{w0} be the number of data points for which $v = 0, w = 0$ respectively, and let I_{uv}, I_{uw} be the respective empirical mutual information values based on the sample $\mathcal{D}$. Then*

$$N_{v0} \;>\; N_{w0} \;\;\Rightarrow\;\; I_{uv} \;\leq\; I_{uw}$$

with equality only if u is identically 0. ■

Proof. We use the notation:

$$P_v(i) \;=\; \frac{N_v^i}{N}, \quad i \neq 0; \quad P_{v0} \;\equiv\; P_v(0) \;=\; 1 - \sum_{i \neq 0} P_v(i).$$

These values represent the (empirical) probabilities of v taking value $i \neq 0$ and 0 respectively. Entropies will be denoted by H. We aim to show that $\frac{\partial I_{uv}}{\partial P_{v0}} < 0$....

References