On Construction of Hybrid Logistic Regression-Naïve Bayes Model for Classification

Yi Tan & Prakash P. Shenoy
School of Business, University of Kansas

Moses W. Chan & Paul M. Romberg
Advanced Technology Center, Lockheed Martin Space Systems

September 7, 2016

Goal of this research

The goal of this study is to examine the construction of such hybrid models from data where we use logistic regression (LR) as a discriminative component, and naïve Bayes (NB) as a generative component. We propose a heuristic which is based on reducing the conditional dependence of the features in NB part of the hybrid model given the class variable.

Outline

1. Introduction

In machine learning, some classifiers are much easier to understand and interpret than others： Logistic Regression and Naïve Bayes Model

1. Few parameters
2. Scale well to high dimensions
3. Be trained very efficiently
4. Robust method

1. Introduction

Logistic Regression :

$$\ln\left(\frac{P(Y=y_j \mid \mathbf{x})}{1-P(Y=y_j \mid \mathbf{x})}\right) = \beta_{0j} + \sum_{i=1}^{n} \beta_{ij} x_i \qquad (0.1)$$

Discrete-valued dependent variable (J. H. Aldrich and F. D. Nelson, 1984; H. W. David, 1989; S. Menard, 1995 discussed about multinomial logistic regression)
F. Rijmen, in 2008, modeled a logistic regression model as a Bayesian network.

1. Introduction

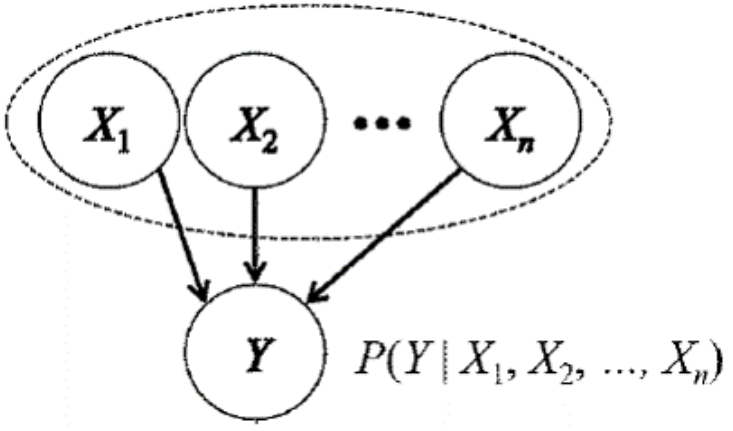

Figure: Logistic Regression as a Bayesian Network.

Logistic regression assumes a parametric form for the distribution $P(Y|X)$. The function for conditional odds of $Y = 1$ given $X_1, \ldots, X_n$ can be derived as:

$$odds(Y = 1 \mid x_1, \ldots, x_m) = \exp(\beta_{0j} + \sum_{i=1}^{n} \beta_{ij} x_i) \qquad (0.2)$$

1. Introduction

Advantages of Logistic Regression:
1. Few parameters
2. Handles both continuous and categorical variables
3. Only a conditional model
4. No conditional independence assumptions

Disadvantages of Logistic Regression:
1. Cannot easily handle missing values

1. Introduction

Naïve Bayes is a full model including priors for Y which makes a strong assumption that feature variables are mutually conditionally independent given the class variable.

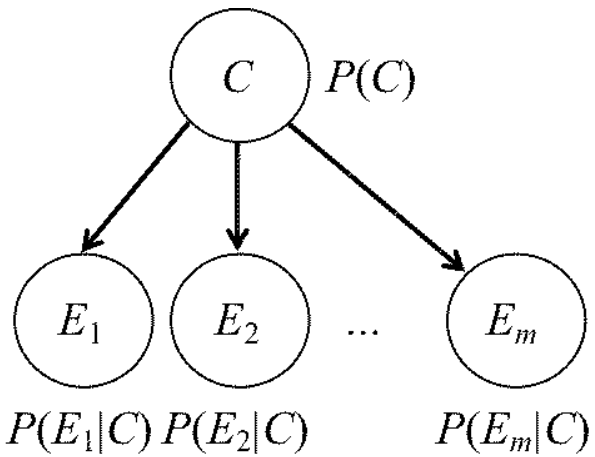

Figure: A NB Model as a Bayesian Network

1. Introduction

Based on the assumption of a NB model, it can be shown that

$$odds(Y = 1 \mid x_1, \ldots, x_m) = odds(Y = 1) \prod_{i=1}^{m} lr(x_i, Y = 1), \quad (0.3)$$

where $lr(x_i, Y = 1) = \frac{P(x_i|Y=1)}{P(x_i|Y=0)}$. If a feature is not observed, we can regard its likelihood ratio as equal to 1.

1. Introduction

Advantages of Naïve Bayes:
1. Few parameters
2. Least sensitive to missing data (Peng, 2005)

Disadvantages of Naïve Bayes:
1. Need discretization for continuous data with non-parametric distribution

1. Introduction

Ng and Jordan (2001) showed that, there can often be two distinct regimes of performance as the training set size increased, one in which each algorithm (LR and NB Model) does better :

1. A LR model has a lower asymptotic accuracy (as the number of training instances becomes large) compared to NB.
2. A NB model approaches its asymptotic error much faster than a LR model.

2. Hybrid Logistic Regression- Naïve Bayes Model

Kang and Tian (2006) introduce a hybrid discriminative-generative classifier where the discriminative component is LR, and the generative component is NB or tree-augmented NB (TAN).

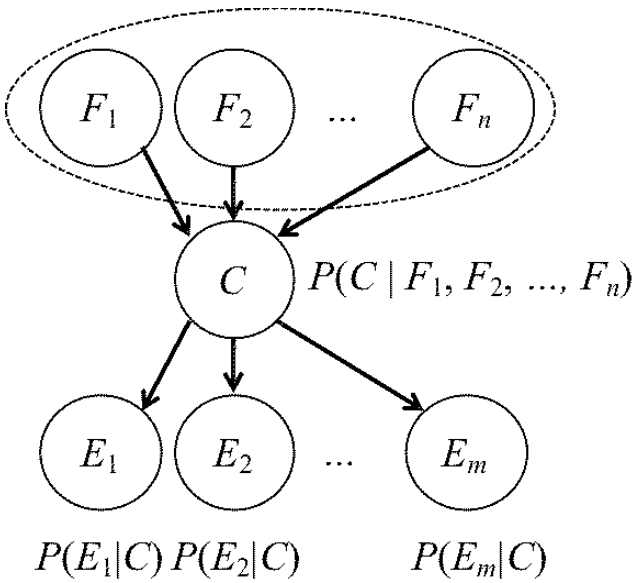

Figure: A Hybrid LR-NB Model as a Bayesian Network

2. Hybrid Logistic Regression- Naïve Bayes Model

The posterior distribution of C given $\mathbf{F} = \mathbf{f}$ and $\mathbf{E} = \mathbf{e}$ is computed by replacing the prior odds of $C = 1$ by the posterior odds from LR part:

$$odds(C = 1|\mathbf{e}, \mathbf{f}) = \exp(\beta_0 + \sum_{i=1}^{n} \beta_i f_i) \prod_{j=1}^{m} lr(e_j, C = 1) \quad (0.4)$$

2. Hybrid Logistic Regression- Naïve Bayes Model

Advantages of Hybrid Model:

1. Simplicity
2. Handle missing values
3. Handle numeric features
4. Flexible to different sizes of data

3. Heuristic to find structure of a hybrid model

Method for constructing a hybrid logistic regression- Naïve Bayes Model ：
Step1： We firstly select features based on the estimated Markov boundary of the class variable C.
Step2： We propose a heuristic which reduces the conditional dependence between the features in the NB part given class variable C to select the LR and NB parts of the hybrid model (Rubinstein and Hastie, 1997).

3. Heuristic to find structure of a hybrid model

Algorithm 1 Find structure of a hybrid model

input: A set of labelled instances.

output: A hybrid network structure with identified *LR-part* and *NB-part*.

1: Find the Markov boundary of class variable C using all 4 constraint-based method with all different conditional independence tests.
2: Set *NB-part* $= \cup MB(C)$ and *LR-part* $= \emptyset$
3: do while $|$*NB-part*$| > 2$
4: Let (X_i, X_j) denote the pair of features that maximizes $norMI(X_i; X_j \mid C)$ where $X_i, X_j \in$ *NB-part*
5: If $max\, norMI \geq 0.05$ then
6: Let $X_s \in$ *NB-part* $\setminus \{X_j\}$ be the attribute that maximizes $norMI(X_i; X_s \mid C)$
7: Let $X_t \in$ *NB-part* $\setminus \{X_i\}$ be the attribute that maximizes $norMI(X_j; X_t \mid C)$
8: If $norMI(X_i; X_s \mid C) < norMI(X_j; X_t|C)$, then *NB-part* $=$ *NB-part* $\setminus \{X_j\}$,
9: *LR-part* $=$ *LR-part* $\cup \{X_j\}$
10: If $norMI(X_i; X_s \mid C) > norMI(X_j; X_t \mid C)$, then *NB-part* $=$ *NB-part* $\setminus \{X_i\}$,
11: *LR-part* $=$ *LR-part* $\cup \{X_i\}$
12: else end do;
13: If $|$*NB-part*$| = 2$ and $norMI(X_i; X_j \mid C) \geq 0.05$, then
14: pick one at random and add to *LR-part*.
15: else end algorithm

4. Experimental Analysis

We conduct experiments on 21 different machine learning datasets from the UCI Machine Learning Repository.
Experimental Setup：
1. Randomly divided each datasets into two parts： training (90%) and testing (10%)
2. Implement our algorithm to identify the model structure and trained the corresponding hybrid model
3. Compare the hybrid model with pure LR model and pure NB model by their prediction accuracies
4. Repeat the entire procedure 100 times.

4. Experimental Analysis

Table: A Summary of 21 Bench-Mark Datasets

Dataset	*# Features*	*# Numeric*	*# Categorical*	*# Instances*	*# Classes*	*Missing Values?*
Pima Indians Diabetes	8	8	0	768	2	yes
Adult Census Income	14	6	8	48,842	2	yes
Credit Approval	15	6	9	690	2	yes
Glass Identification	10	10	0	214	6	no
Hypothyroid	19	7	12	3,163	2	yes
Statlog Vehicle Silhouettes	18	18	0	846	4	no
Wine	13	13	0	178	3	no
Bank Marketing	19	9	10	41,188	2	yes
Banknote Authentication	4	4	0	1,372	2	no
Car Evaluation	6	0	6	1,728	4	no
Chronic Kidney Disease	24	11	13	400	2	yes
Blogger	5	0	5	100	2	no
Breast Tissue	9	9	0	106	6	no
Congressional Voting Records	16	0	16	435	2	yes
Connectionist Bench	60	60	0	208	2	no
Default of Credit Card Clients	23	14	9	30,000	2	no
Ecoli	7	5	2	336	8	no
Mushroom	22	0	22	8,124	2	yes
Nursery	8	0	8	12,960	3	no
Qualitative Bankruptcy	6	0	6	250	2	no
EEG Eye State	14	14	0	14,980	2	no

4. Experimental Analysis

Table: Summary of Results: Avg Est. Markov Blanket Size, Avg Structure of Hybrid Models, and Avg Accuracies of Models, in units of % (SE in parenthesis). Highest accuracies are in boldface.

Dataset	*# Features*	*# MB*	*# LR-part*	*# NB-part*	*Acc. Hybrid*	*Acc. LR*	*Acc. NB*
Pima Indians Diabetes	8	4.47	0.89	3.58	**80.80** (0.45)	77.95 (0.43)	**80.89** (0.46)
Adult Census Income	14	9.87	3.14	6.73	82.12 (0.06)	**84.68** (0.06)	80.73 (0.08)
Credit Approval	15	8.62	3.07	5.55	**85.76** (0.49)	**85.62** (0.49)	**85.18** (0.47)
Glass Identication	10	5.47	3.49	1.98	**66.49** (0.42)	62.36 (0.39)	63.14 (0.60)
Hypothyroid	18	4.46	1.30	3.16	98.24 (0.08)	97.85 (0.07)	**98.59** (0.07)
Statlog Vehicle Silhouettes	18	17.15	14.48	2.67	78.74 (0.49)	**80.35** (0.40)	65.15 (0.49)
Wine	13	9.82	5.84	3.98	96.52 (0.21)	95.03 (0.22)	**97.90** (0.26)
Bank Marketing	19	12.92	7.91	5.01	88.24 (0.05)	**88.84** (0.04)	83.57 (0.07)
Banknote Authentication	4	3.00	2.00	1.00	96.43 (0.20)	**98.92** (0.08)	92.79 (0.19)
Car Evaluation	6	5.00	0.99	4.01	85.79 (0.30)	**92.52** (0.20)	85.79 (0.30)
Chronic Kidney Disease	24	10.29	6.08	4.21	97.26 (0.36)	**98.80** (0.10)	93.16 (0.69)
Blogger	5	2.04	1.04	1.00	68.24 (0.67)	**69.64** (0.57)	66.48 (0.73)
Breast Tissue	9	7.09	5.42	1.67	66.90 (0.64)	**67.80** (0.64)	66.00 (0.68)
Congressional Voting Records	16	7.39	4.58	2.81	94.06 (0.30)	**95.24** (0.30)	92.76 (0.30)
Connectionist Bench	60	7.48	3.61	3.88	71.48 (0.44)	68.65 (0.45)	**72.36** (0.44)
Default of Credit Card Clients	23	6.93	4.30	2.63	81.72 (0.06)	**82.07** (0.06)	80.50 (0.06)
Ecoli	7	6.00	3.79	2.21	**85.57** (0.57)	**85.77** (0.53)	83.73 (0.66)
Mushroom	22	12.53	10.60	1.93	99.98 (0.009)	**99.99** (0.003)	92.96 (0.099)
Nursery	8	8.00	1.00	7.00	90.29 (0.09)	**92.45** (0.07)	90.29 (0.09)
Qualitative Bankruptcy	6	4.54	1.16	3.38	**99.64** (0.10)	**99.56** (0.10)	**99.64** (0.10)

4. Experimental Analysis

In a pairwise comparison between hybrid and pure LR models, hybrid models outperform LR for 6 datasets, are tied with LR for 3 datasets, and do worse than LR for 12 datasets.

In a pairwise comparison between hybrid and pure NB models for the 21 datasets, hybrid models outperform NB for 12 datasets, are tied with NB for 5 datasets, and do worse than NB for 4 datasets.

4. Future work

1. New heuristic by considering other factors, such as missing values of features, number of observations and number of parameters to be estimated.

2. Sensitivity analysis.

Thank you!