Bayesian Classification of Personal Histories:

An application to the Obesity Epidemic

Christopher R. Stephens , Jose Antonio Borras Gutierrez , and Hugo Flores

C3 - Centro de Ciencias de la Complejidad y Instituto de Ciencias Nucleares, UNAM, 04510 CDMX, Mexico

Facultad de Ciencias, Universidad Nacional Autonoma de Mexico, 04510 CDMX, Mexico

Tech Mileage, 3295 N Drinkwater Blvd, Suite 13, Scottsdale, AZ 85251, USA

AMLTA 2019, Ain Shams University 28-30th March 2019

Bayesian Classifiers

- Want to determine $P(C|\boldsymbol{X})$ for some feature vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$
- As a classifier, if $P(C|\boldsymbol{X}) > P(\bar{C}|\boldsymbol{X})$, where $\bar{C}$ is the set complement of C, then $\boldsymbol{X}$ is considered to be in class.
  - $S(C|\boldsymbol{X}) = \ln\left(\frac{P(C|\boldsymbol{X})}{P(\bar{C}|\boldsymbol{X})}\right) > 0$
- Cannot be determined empirically when $\boldsymbol{X}$ is of high dimension
- Bayes Theorem $P(C|\boldsymbol{X}) = P(\boldsymbol{X}|C)P(C)/P(\boldsymbol{X})$
  - Allows for incorporation of both prior beliefs and information $\boldsymbol{X}$
  - Allows for iteration of this process: prior → posterior → prior → …
  - Need to calculate the likelihood $P(\boldsymbol{X}|C)$

Bayesian Classifiers

- Simplest approximation…
- Assume complete factorization
- $P(\boldsymbol{X}|C) = \prod_{i=1}^{N} P(X_i|C)$
  - Naïve Bayes Approximation leading to Naïve Bayes Classifier
  - Very robust and simple. Surprisingly good performance given strong assumption.
- Many generalizations…
- $P(\boldsymbol{X}|C) = \prod_{i=1}^{N(\xi)} P(\xi_i|C)$
  - Generalized Bayes Approximation leading Generalized Bayes Classifier
  - Schema ξ_i represents a combination of features that are to be considered together
  - The more features in ξ_i the more unreliable is its statistical estimate but the better it takes into account feature correlations – balance
- Can construct statistical diagnostics to determine when GBA is better than NBA – No Free Lunch Theorem (when would we expect one classifier to be better than another?)

Histories as correlated features

- Under what conditions do we expect features to be correlated?
  - Spatial correlations
  - Temporal correlations
- Histories – non-Markovian
- Habits are histories
- $S(C|\boldsymbol{X}) = \ln\left(\frac{P(C|\boldsymbol{X})}{P(\bar{C}|\boldsymbol{X})}\right) = \sum_{i=1}^{N(\xi)} s(\xi_i) + \ln(\frac{P(C)}{P(\bar{C})})$
  - $s(\xi_i) = \ln(\frac{P(\xi_i|C)}{P(\xi_i|\bar{C})})$ is the score associated with the feature combination (history) ξ_i
  - Feature selection carried out using a binomial test
    - $\varepsilon(C|\xi_i) = (N_\xi(P(C|\xi_i) - \mathrm{P}(C))/(N_\xi(P(C)(1 - \mathrm{P}(\mathrm{C})))^{1/2}$

The Obesity Crisis

- Obesity is probably the world's biggest health crisis
  - Leading to excess mortality and morbidity
  - Increasing in spite of world-wide investment in financial and human resources
- Egypt has the highest percentage of obese adults worldwide (NEJM 2017)
- Around 19 million Egyptians, or 35 percent of the adult population, are obese – the highest rate across the globe.
- In addition, over 10 percent, or 3.6 million, of children are also considerably overweight, the study reported.
- Mexico is just as bad!
- Chief risk factors are: malnutrition and sedentariness - as "bad habits"
  - Everyone knows that!
  - They're both extremely multi-factorial
  - Why do people make "bad" decisions; have such "bad" habits?

The Study Data

- Project 42
  - Create the world's "deepest" – multi-factorial, multi-scale - database for the study of obesity and metabolic disorders
  - Over 2000 academics, workers and students from Mexico's largest university (UNAM)
    - Age range 23-85
    - **21% overall obesity rate**
    - **13%/40% obesity rate for academics/workers. Why?**
  - Over 2000 variables (genetic, epidemiological, physiological, psychological, social)
  - Self-reported histories:
    - eating habits, exercise, health, stress, weight
    - "Now" ($t = 2014$), t-1, t-5, t-10, t-20, t-30

- Exercise: number of hours weekly exercise
  - T_{min} = 2.5 h/w (WHO 2018) taken as minimum recommended amount

The Prediction Problem

- C = obesity
- C = academic
- $\boldsymbol{X} = X_1(t-30)X_2(t-20)X_3(t-10)X_4(t-5)X_5(t-1)X_6(t)$ = exercise history
  - $X_i = A$ if $X_i > T_{min}$
  - $X_i = B$ if $X_i < T_{min}$
  - $X_i = *$ if X_i = anything (don't care symbol)
  - E.g. AAAABB is a person who exercised more than the minimum recommended amount 30, 20, 10 and 5 years ago and less than the recommended amount 1 year ago and currently
- Calculate S(C | **X**) using
  - Naïve Bayes Approximation
  - Generalized Bayes Approximation (uses correlation of histories)

Results

Top and bottom drivers for C = obesity

History	ϵ	N_x	N_{cx}	%	Score
A*A*BB	3.56	94	38	40.43	0.73
AAA*B	3.55	91	37	40.66	0.74
AA**BB	3.53	113	44	38.94	0.67
AA**B*	3.40	131	49	37.40	0.60
A***BB	3.23	137	50	36.50	0.57
*A***A	−3.27	157	21	13.38	−0.75
***AAA	−3.27	157	21	13.38	−0.75
AA**AA	−3.51	103	10	9.71	−1.11
*A**AA	−3.61	134	15	11.19	−0.95
****AA	−3.76	193	25	12.95	−0.79

Note that, for obesity, to have changed from good to bad habits is worse than never having had good habits in the first place!

Top and bottom drivers for C = academic

History	ϵ	N_x	N_{cx}	%	Score
*A***A	5.55	157	85	54.14	0.86
*A**AA	5.21	134	73	54.48	0.88
*AA**A	5.13	135	73	54.07	0.86
*A*A*A	5.06	129	70	54.26	0.87
AA	4.97	165	85	51.52	0.76
*BBB**	−4.32	197	37	18.78	−0.77
***BB*	−4.40	267	55	20.60	−0.65
**BBB*	−4.41	207	39	18.84	−0.76
***BBB	−4.41	245	49	20.00	−0.69
***B*B	−4.55	260	52	20.00	−0.69

Good versus bad habit patterns clearly differentiate between academics and non-academics.

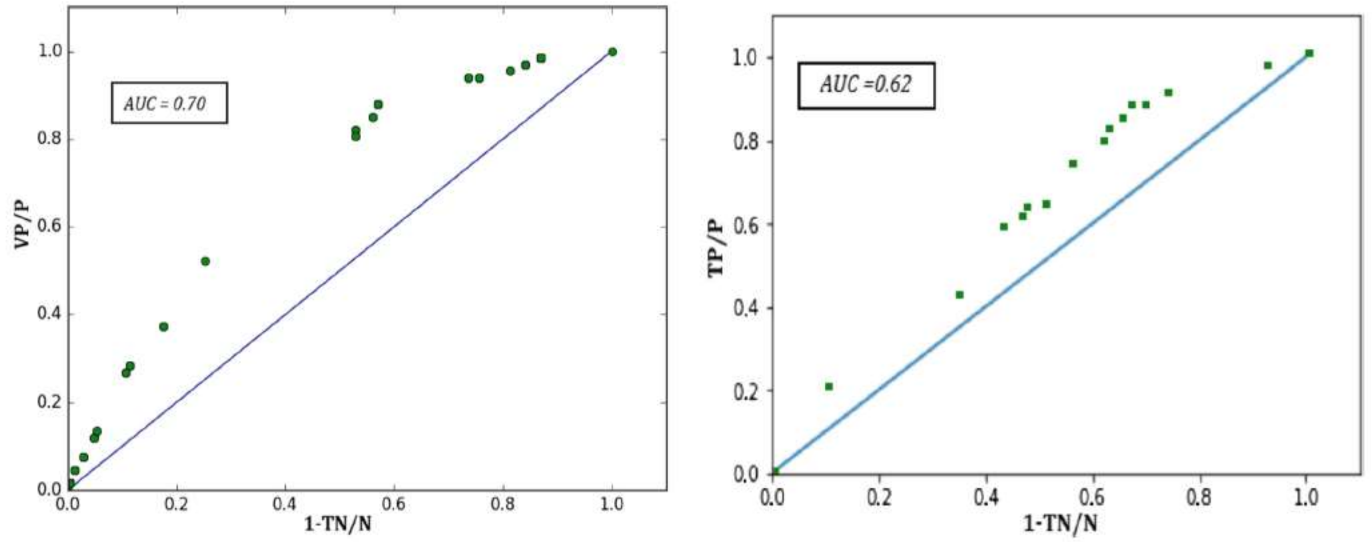
Results

Table 4. Model performance comparison for three models: the NB model, a *Generalised Bayes* model with histories $\#LLLLL$ and a *Generalised Bayes* model with coarse grained histories $\#L**LL$. The column *threshold* refers to the score threshold used for the model to classify predictions. N and P are the number in the no class and class respectively (N = 246, P = 67), while TN and TP are the number of true negatives and true positives associated with each model. PPV is the *positive predictive value*, defined as $PPV = TP/(TP + FP)$ where FP means the number of false positives. $x(1 - TN/N)$ and $y(TP/P)$ are the specificity and sensitivity respectively. Dist is the distance of the point on the ROC curve farthest from the diagonal line corresponding to a random model and Area is the AUC.

ScoreType	Best	Model	Threshold	TN	TP	PPV	x(1-TN/N)	y (TP/P)	dist	Area
Generalised	PPV	#L**LL	0.27	186	32	0.36	0.24	0.48	0.17	0.66
Generalised	dist	#LLLLL	−0.48	106	59	0.30	0.57	0.88	0.22	0.70
Generalised	Area	#LLLLL	−0.48	106	59	0.30	0.57	0.88	0.22	0.70
NB	-	#LLLLL	−0.78	82	60	0.27	0.67	0.90	0.16	0.62

Note that any GNB approximation clearly outperforms the NBA

Results



(a) Generalised (b) NB

Fig. 1. ROC curve with AUC value for (a) the generalised model with $a_0 = \#$, $a_n = A, B$ with $n > 0$ and (b) Naive Bayes.

Table 5. Model performance comparison for an obesity classifier using different degrees of historical information included as Generalized Bayes feature combinations.

Model	Threshold	PPV	x(1-TN/N)	y (TP/P)	dist	Area
#****L	−0.17	0.26	0.60	0.79	0.13	0.59
#***LL	−0.39	0.27	0.67	0.90	0.16	0.60
#**LLL	−0.52	0.27	0.67	0.90	0.16	0.61
#*LLLL	−0.49	0.28	0.60	0.85	0.18	0.67
#LLLLL	−0.48	0.30	0.5	0.88	0.22	0.70

Note that the more historical Information that is included the more accurate the classification. This would not be true if there were no correlations.

Conclusions

- The Generalized Bayes approximation is a way to account for feature variable correlation
- Histories, and in particular human habits, are highly correlated
- Correlated (in time) lifestyle factors (habits) are an important contributing factor to the obesity epidemic
- The GBA leads to more accurate classifiers than the NBA
- The GBA shows that certain exercise patterns are more linked to obesity (losing a good habit is worse than always having had a bad habit)
- Higher education is linked to better health outcomes. One reason for this is that the better educated have better exercise habits.
- The GBA applied to habits is an effective Machine Learning technique that can be applied to multiple problem areas in health and social sciences