Interpretable Boosted Naive Bayes Classification

Greg Ridgeway,
David Madigan,
Thomas Richardson
Department of Statistics

John O'Kane
Department of Orthopedics

University of Washington

Introduction

- Classification - from a set of observable features choose among a discrete set of class labels
- Interpretability - the quality of a model that exposes its reasoning process in a way that a person could understand

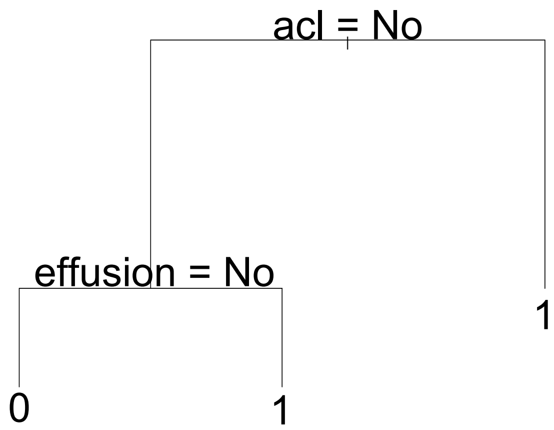
Classification models

$$h : \text{features} \rightarrow \text{class label}$$

- Written digit recognition
- Automated medical diagnosis
- Credit approval
- Remote sensing

Interpretability

- Tree models



- Logistic regression

$$\log \frac{P(Y=1 \mid \text{acl}, \text{effusion})}{P(Y=0 \mid \text{acl}, \text{effusion})} = \beta_0 + \beta_1 \cdot 1_{acl} + \beta_2 \cdot 1_{effusion} + \varepsilon$$

Naïve Bayes Classification

Probabilistic Classification

$$P(Y=y \mid X_1=x_1,\ldots,X_d=x_d)=\frac{P(\underline{X} \mid Y=y)P(Y=y)}{P(\underline{X})}$$

The naïve Bayes assumption

$$P(\underline{X} \mid Y=y)=P(X_1=x_1 \mid Y=y)\cdots P(X_d=x_d \mid Y=y)$$

Estimation

- Probability estimates are trivial

$$\hat{P}(X_j = x_j \mid Y = y) = \frac{\text{count}(X_j = x_j \cap Y = y)}{\text{count}(X_j = x_j)}$$

- Estimation is linear in the number of predictors and the number of observations

Interpretability

Consider the log-odds in favor of $Y=1$

$$\log \frac{P(Y=1 \mid \underline{X})}{P(Y=0 \mid \underline{X})} = w_0 + \sum_{j=1}^{d} w_j (X_j)$$

- Positive w_j are evidence in favor of $Y=1$
- Negative w_j are evidence in favor of $Y=0$

Evidence balance sheets

Evidence in favor of knee surgery		Evidence against knee surgery	
Female	+8	Prior evidence	-10
Knee is unstable	+88	Age 50	-12
Knee locks	+172	No effusion	-62
Tender med JL	+49	Negative McMurray's	-38
Total positive evidence	+317	Total negative evidence	-122
Total evidence		+195	
Probability of knee surgery		88%	

Boosting algorithms

1. Learn a classifier from the data
2. Upweight observations poorly predicted, downweight observations well predicted
3. Refit the model using the new weighting
4. After *T* iterations, have each model vote on the final prediction.

AdaBoost algorithm

Freund & Shapire (1997)

- AdaBoost defines a particular reweighting scheme and a voting method for merging the classifiers
- AdaBoost decreases bias and variance in many settings - Bauer and Kohavi [1998]
- Boosted naïve Bayes tied for first place in the 1997 KDD Cup

AdaBoost

- Extremely dense voting scheme

$$P(Y=1\mid x)=\frac{1}{1+\prod_{t=1}^{T}\beta_t^{2r(x)-1}} \qquad r(x)=\frac{\sum_{t=1}^{T}(\log\frac{1}{\beta_t})P_t(Y=1\mid x)}{\sum_{t=1}^{T}(\log\frac{1}{\beta_t})}$$

- Destroys interpretability

Regaining Interpretability

Rewriting the voting scheme…

$$\log \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)} = \sum_{t=1}^{T} (\log \beta_t) \left(1 - 2 \left(1 + e^{-\log \frac{P_t(Y=1|X)}{P_t(Y=0|X)}} \right)^{-1} \right)$$

Substitute Taylor expansion…

$$\frac{1}{1+e^{-x}} = \tfrac{1}{2} + \tfrac{1}{4}x - \tfrac{1}{48}x^3 + O\left(x^5\right)$$

Regained Interpretability

$$\sum_{t=1}^{T} \alpha_t \log \frac{P_t(Y=1)}{P_t(Y=0)} + \sum_{j=1}^{d} \sum_{t=1}^{T} \alpha_t \log \frac{P_t(X_j \mid Y=1)}{P_t(X_j \mid Y=0)}$$

$$= \text{boosted prior weight of evidence} + \sum_{j=1}^{d} \text{boosted weight of evidence from } X_j$$

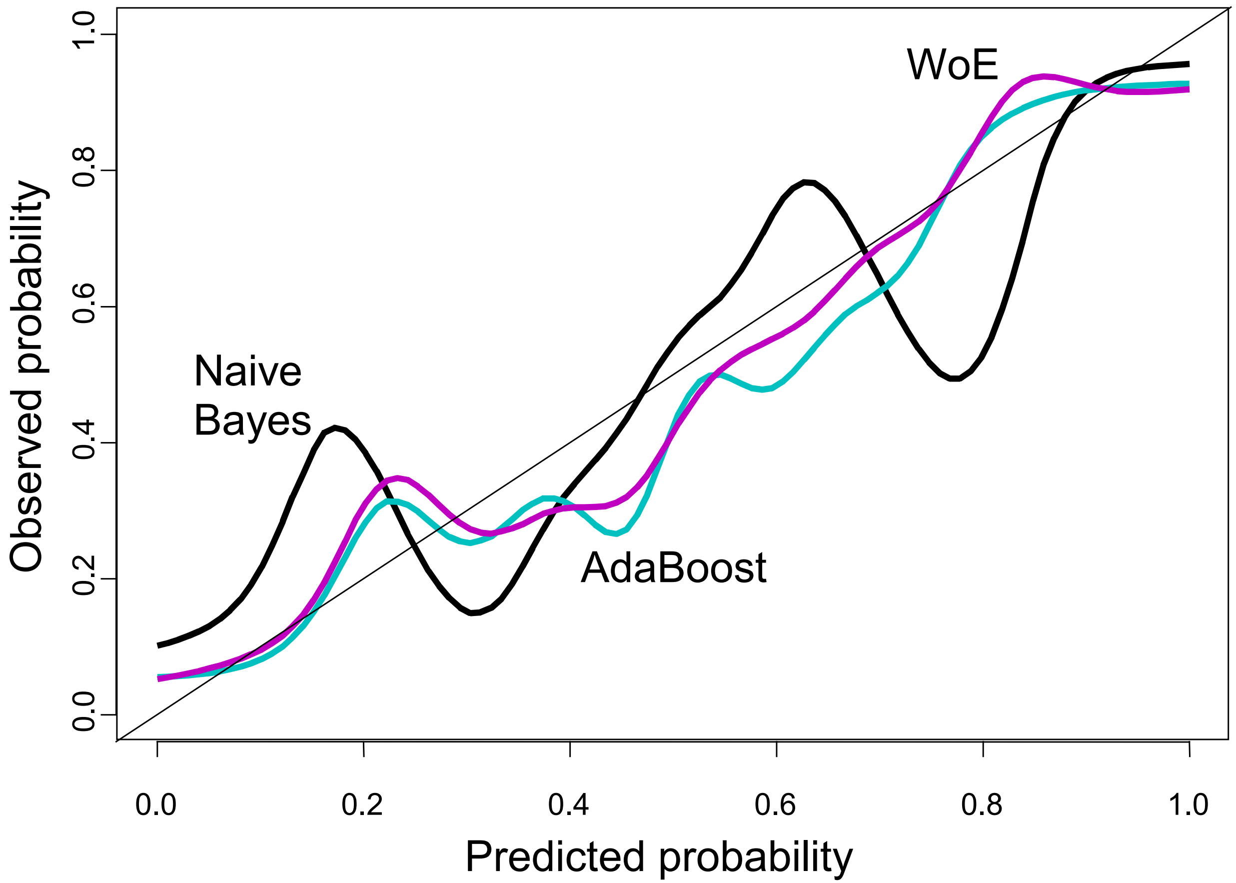
- Boosting biases parameter estimates
- Adjusts naïve Bayes' for over-optimism

Misclassification rates

	Naïve Bayes	AdaBoost	Weight of evidence
Knee diagnosis	14.0%	13.8%	13.4%
Diabetes	25.0%	24.4%	24.4%
Credit approval	16.8%	15.5%	15.5%
CAD	18.4%	18.3%	18.3%
Breast tumors	3.9%	3.8%	3.8%

- Boosting offers modest improvement
- Actual AdaBoost and approximation are close

Calibration



Future directions

- Search for other boosted models that are interpretable
- Further investigation of the effect of boosting on calibration
- Synthesis of boosting and likelihood methodology - Friedman, *et al* [1998]

Conclusions

- Naïve Bayes is a simple, efficient, and interpretable classifier
- Boosting improves the naïve Bayes classifier but does not necessarily sacrifice its interpretability
- Boosting may improve calibration of probabilistic classifiers