Interpretable Boosted Naïve Bayes Classification

Greg Ridgeway
David Madigan
Thomas Richardson
Department of Statistics

John O'Kane
Department of Orthopedics

University of Washington

Outline

1. Classification problems
2. Naïve Bayes classifier
3. Weights of evidence
4. Boosting
5. Boosting the naïve Bayes classifier

Classification

Prediction of examples to a discrete set of possibilities

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

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

Selection of $h : \underline{X} \rightarrow Y$

Good choice for *h* minimizes misclassification rate…

$$P(h(X) = Y)$$

where (X,Y) come from their natural distribution.

Common *h*’s

- Decision trees - CART, C4.5
- Naïve Bayes classifier
- Discriminant analysis
- Logistic regression
- Neural network

Naïve Bayes classification

Bayes' Theorem

$$P(Y = y \mid X) = \frac{P(X \mid Y = y)P(Y = y)}{P(X)}$$

$\mathrm{P}(Y{=}1|\underline{X}) \propto$

$\mathrm{P}(Y{=}1)\mathrm{P}(X_1|Y{=}1)\ldots\mathrm{P}(X_d|Y{=}1)$

$\mathrm{P}(Y{=}0|\underline{X}) \propto$

$\mathrm{P}(Y{=}0)\mathrm{P}(X_1|Y{=}0)\ \ldots\mathrm{P}(X_d|Y{=}0)$

Estimation

Probability estimates are trivial when all the *Y*'s are observed.

$$\hat{P}(X_j = 1 \mid Y = 1) = \frac{N[x_{ij} = 1 \mid y_i = 1]}{N[y_i = 1]}$$

Bayesian estimation adds a constant to the numerator and denominator.

Comments on Naïve Bayes

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

Naïve Bayes is robust to violations of the conditional independence assumption.

Naïve Bayes is robust to irrelevant predictors.

Naïve Bayes models are easy to interpret.

Weight of evidence

$$\log \frac{P(Y=1\,|\,X)}{P(Y=0\,|\,X)}$$

$$= \log \frac{P(Y=1)\prod_{j=1}^{d} P(X_j\,|\,Y=1)}{P(Y=0)\prod_{j=1}^{d} P(X_j\,|\,Y=0)}$$

$$= \log \frac{P(Y=1)}{P(Y=0)} + \sum_{j=1}^{d} \log \frac{P(X_j\,|\,Y=1)}{P(X_j\,|\,Y=0)}$$

$$= w_0 + \sum_{j=1}^{d} w_j(X_j)$$

Positive $w_j(X_j)$ is evidence in favor of Y=1.

A negative weight is evidence for Y=0.

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%	

Example evidence balance sheet

Boosting algorithms

1. Initially weight all observations equally
2. Fit a model to the data
3. Upweight observations poorly modeled… downweight well modeled observations
4. Refit the model accounting for the new weighting
5. After *T* iterations each model “votes” on a final prediction with strength proportional to quality

Boosting classification

Model → $h(x)$ = P(Y=y|X=x)

Fit = Estimate the parameters of P(Y=y|X=x)

Quality =

$$\tfrac{1}{N}\sum_{i=1}^{N}\hat{P}(Y = y_i \mid X = x_i)$$

AdaBoost algorithm

(Freund & Shapire, 1997)

Fit model $h_t(x_i) : X \rightarrow [0,1]$.

$$\varepsilon_t = \sum_{i=1}^{N} w_i^{(t)} \left| y_i - h_t(x_i) \right|$$

$$\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t} \qquad w_i^{(t+1)} = w_i^{(t)} \beta_t^{1-\left| y_i - H_t(x_i) \right|}$$

$$h(x) = \frac{1}{1+\prod_{t=1}^{T} \beta_t^{2r(x)-1}}$$

$$r(x) = \frac{\sum_{t=1}^{T} (\log \frac{1}{\beta_t}) h_t(x)}{\sum_{t=1}^{T} (\log \frac{1}{\beta_t})}$$

Comments on AdaBoost

1. F&S prove that on the training dataset

$$\varepsilon \leq 2^{T-1} \prod_{t=1}^{T} \sqrt{\varepsilon_t (1-\varepsilon_t)}$$

if $h(x)$ is a “weak learner”.

2. No reasonable bounds exist for generalization

3. Boosting takes simple and interpretable models and makes them impenetrable

$$
\begin{aligned}
\log\frac{h(x)}{1-h(x)} &= -\log\prod_{t=1}^{T}\beta_t^{2r(x)-1} \\
&= (1-2r(x))\sum_{t=1}^{T}\log\beta_t \\
&= \sum_{t=1}^{T}(\log\beta_t)(1-2P_t(Y=1\,|\,X))
\end{aligned}
$$

Using the fact that

$$
P_t(Y=1\,|\,X) = \frac{1}{1+\frac{P(Y=0|X)}{P(Y=1|X)}} = \left(1+e^{-\log\frac{P_t(Y=1|X)}{P_t(Y=0|X)}}\right)^{-1}
$$

we can rewrite the log-odds of the combined classifiers as a function of the log-odds of each $P_t(\bullet)$:

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

Continued

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

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

A Taylor approximation yields a linear combination of the log-odds from each boosted naïve Bayes classifier:

$$\approx \sum_{t=1}^{T} (\tfrac{1}{2} \log \tfrac{1}{\beta_t}) \log \frac{P_t(Y=1\,|\,X)}{P_t(Y=0\,|\,X)} .$$

Lastly, substitute the naïve Bayes classifier

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

$$= \text{boosted prior weight of evidence} +$$

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

Empirical results

	Naïve Bayes	AdaBoost	Weight of evidence
Knee diagnosis	14.0% (5.0%)	13.8% (5.5%)	13.4% (5.7%)
Diabetes	25.0% (2.0%)	24.4% (2.5%)	24.4% (2.6%)
Credit approval	16.8% (2.0%)	15.5% (2.1%)	15.5% (2.1%)
Coronary artery disease	18.4% (3.0%)	18.3% (3.2%)	18.3% (3.3%)
Breast tumors	3.9% (1.0%)	3.8% (1.0%)	3.8% (1.0%)

Misclassification rates

Final Comments

Empirically evidence shows that in a wide variety of problems and with various base classifiers boosting

- decreases misclassification error,
- reduces variance in unstable classifiers,
- reduces bias

The naïve Bayes classifier is

- robust,
- interpretable

Boosting weights of evidence combines these two into a competitive classifier.

Ongoing research…

Conjecture: Boosting turns non-Bayes risk consistent classifiers into Bayes risk consistent classifiers.

Regression: Project regression dataset into a classification dataset on infinite size. Use the boosted naïve Bayes classifier with the weight of evidence approximation to obtain a flexible and interpretable model.

X_1	X_2	Y
0.6	0.4	0.3
0.8	0.5	0.9
0.1	0.2	0.5

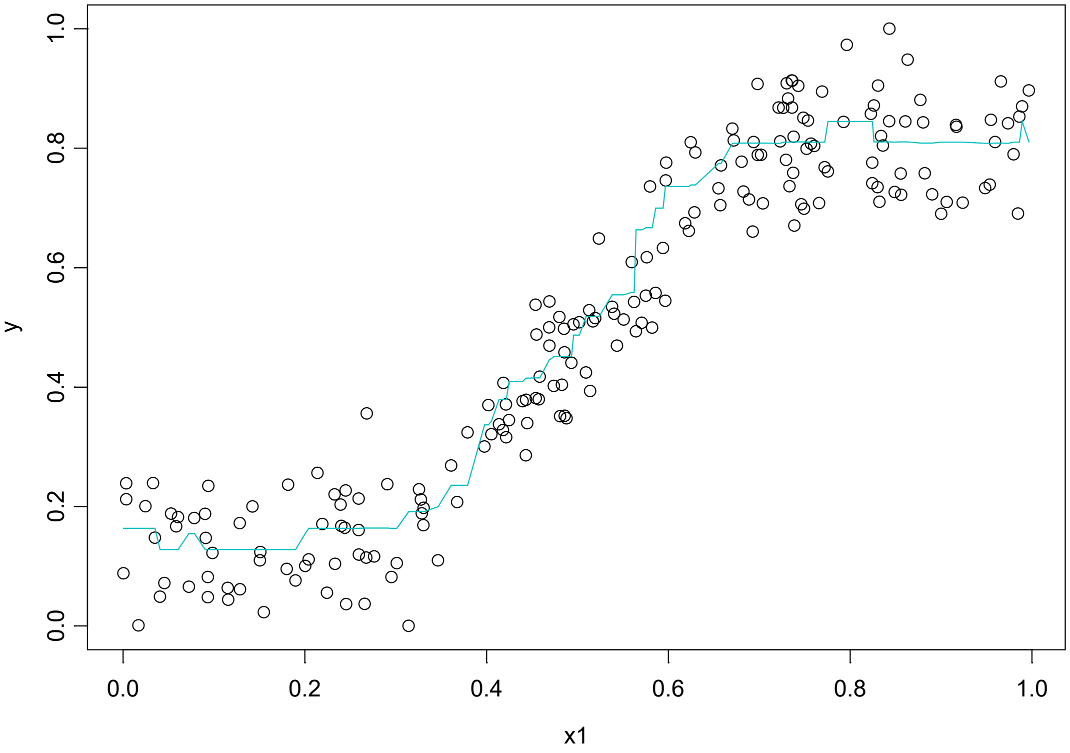
	X_1	X_2	Y	S	Y*
Obs. 1	0.6	0.4	0.3	0.00	0
	0.6	0.4	0.3	0.25	0
	0.6	0.4	0.3	0.50	1
	0.6	0.4	0.3	0.75	1
	0.6	0.4	0.3	1.00	1
Obs. 2	0.8	0.5	0.9	0.00	0
	0.8	0.5	0.9	0.25	0
	0.8	0.5	0.9	0.50	0
	0.8	0.5	0.9	0.75	0
	0.8	0.5	0.9	1.00	1
Obs. 3	0.1	0.2	0.5	0.00	0
	0.1	0.2	0.5	0.25	0
	0.1	0.2	0.5	0.50	1
	0.1	0.2	0.5	0.75	1
	0.1	0.2	0.5	1.00	1

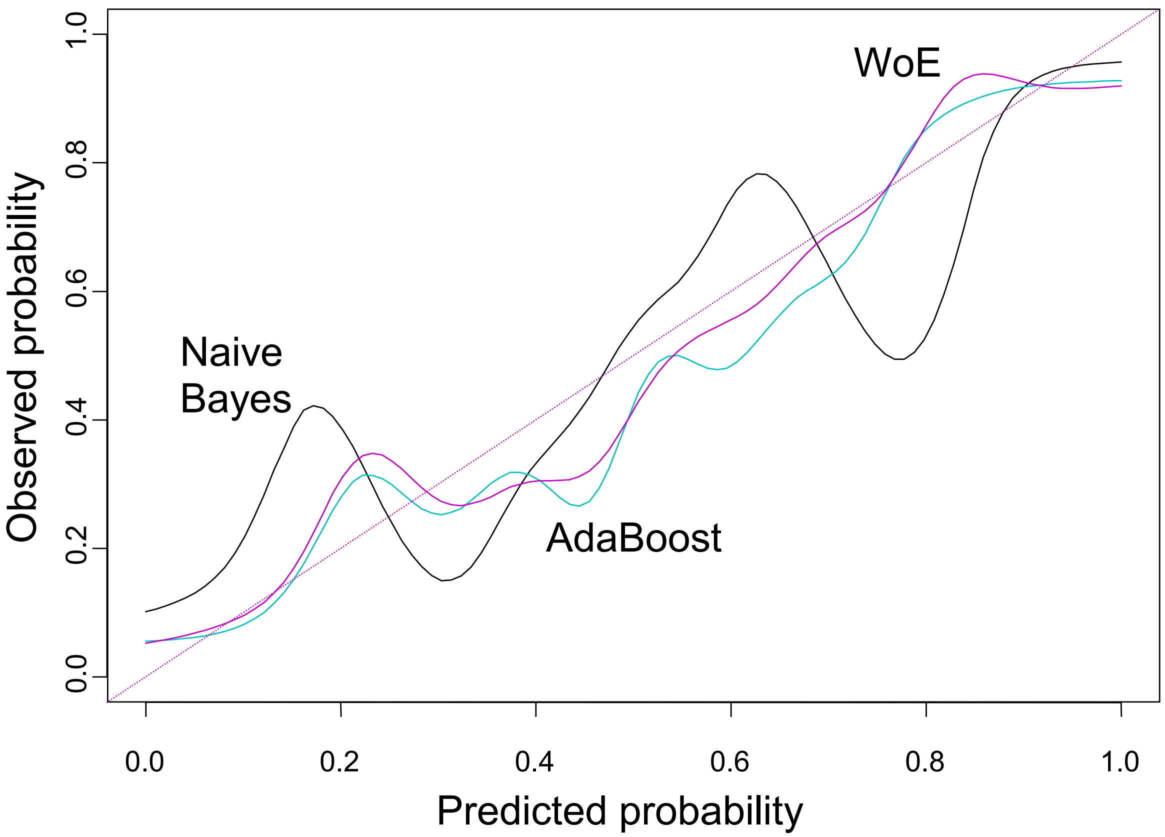
$$\hat{Y} = \inf_{y} \left\{ y : P(Y^* = 1 \mid \underline{X}, y) \geq \tfrac{1}{2} \right\}$$

$$P(Y^* = y^* \mid X_1, \cdots, X_d, S) \propto$$

$$P(Y^* = y^*)P(S \mid Y^* = y^*)\prod_{j=1}^{d} P(X_j \mid Y^* = y^*)$$

$$\hat{Y} = \inf_{y} \left\{ y : \frac{P_S(y \mid Y^* = 1)}{P_S(y \mid Y^* = 0)} \geq \frac{P(Y^* = 0)}{P(Y^* = 1)} \prod_{j=1}^{d} \frac{P(X_j \mid Y^* = 0)}{P(X_j \mid Y^* = 1)} \right\}.$$




Naive Bayes
AdaBoost
WoE
Observed probability
0.0
0.2
0.4
0.6
0.8
1.0
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0