A Discussion of Disease Prediction and Model Validation

Greg Ridgeway
RAND Statistics Group

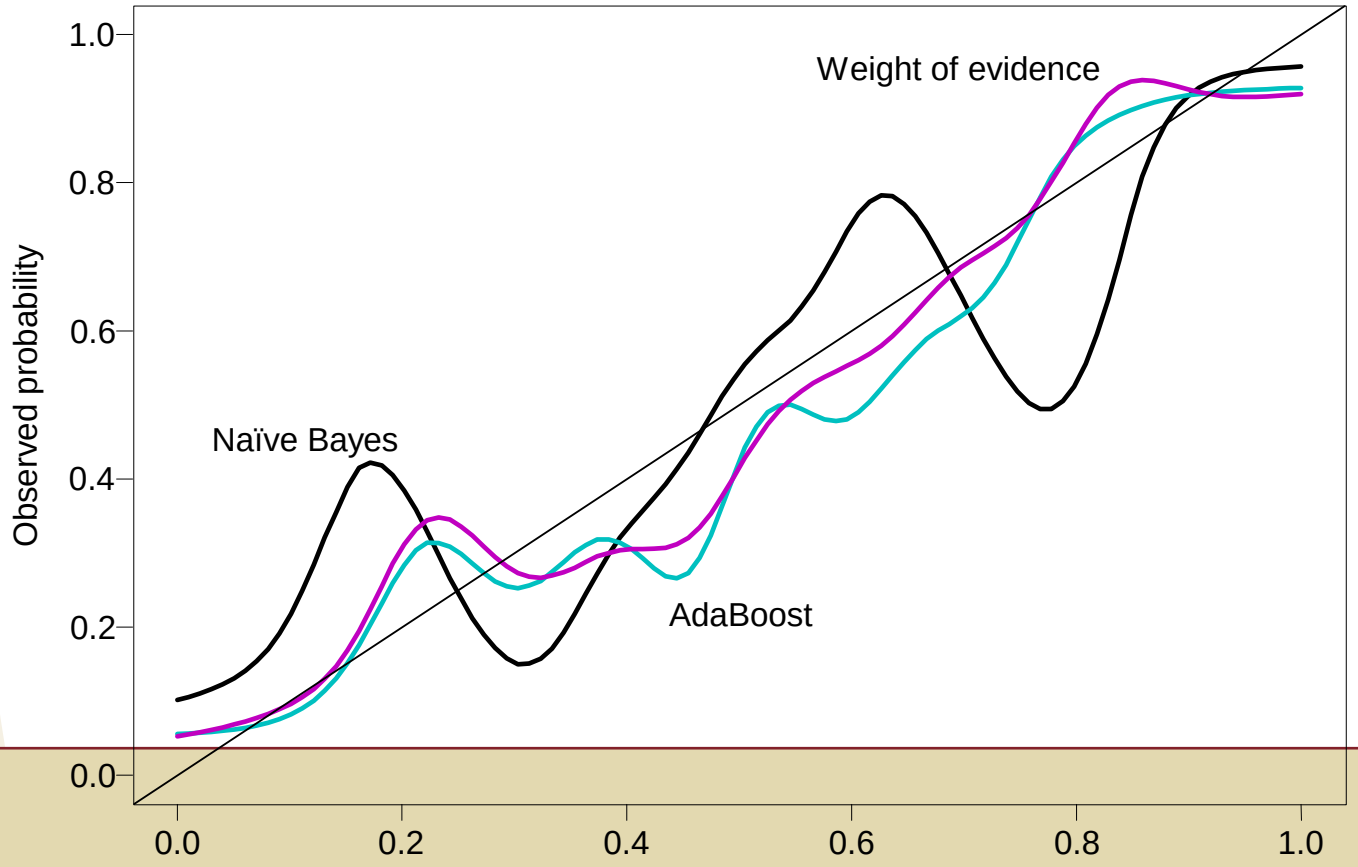
August 13, 2008

A few comments, additions, and views

- Brier score: a useful decomposition of discrimination and calibration
- Modern statistical prediction methods: L_1 shrinkage with lots of covariates
- Weight of evidence scorecard

Gönen: Calibration matters

- Better performance through improved calibration - criticized at a data mining conference as irrelevant
- Yates (1982) suggested that organizational psychologists were too focused on calibration



Brier score

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - p(\mathbf{x}_i))^2$$

$$= \bar{y}(1-\bar{y}) - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 + \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \bar{y}_k)^2$$

$$= \text{uncontrollable variation} + \text{resolution} + \text{calibration}$$

- Resolution is large (very negative) when we can discriminate the 0 outcomes from the 1s, when the average outcome given prediction p_k near 0 or 1
- Calibration is the ability to assign meaningful probabilities to the outcomes.

Sun & Bang: Logistic regression

- Indicate good out-of-sample predictive performance
- Both carefully handled data (e.g. survey weights, correlations)
- Modern statistical prediction methods might squeeze out more signal, especially with $N = 12,000$

L_1 shrinkage and regression

$$\ell(\beta) = \sum_{i=1}^{n} y_i \beta' \mathbf{h}(\mathbf{x}_i) - \log\left(1 + e^{\beta' \mathbf{h}(\mathbf{x}_i)}\right) - \lambda \sum_{j=1}^{J} |\beta_j|$$

- Let $\mathbf{h}(\mathbf{x})$ be piecewise constant functions of the x_js and their interactions, I(BMI$<$20), I(COPD Hx=NA), or I(age$<$40)I(waist$>$90)
- Efron *et al* (2004) LARS essentially showed that boosting essentially implements this, but avoids constructing the full design matrix
- The `gbm` package in R implements this

Spiegelhalter & Knill-Jones (1984)

- Report: ”Despite encouraging results in a research context, statistical systems have had limited practical impact”
- Suggest the failure of statistical systems results from being *too simplistic*, *inapplicable*, and *incomprehensible*
- Present a weight-of-evidence scorecard that accommodates “ignorance” and “conflict of evidence”

Weights of evidence

$$
\begin{aligned}
WOE &= \log \frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})} \\
&= \log \frac{P(Y=1)}{P(Y=0)} + \log \frac{P(X_1=x_1|Y=1)}{P(X_1=x_1|Y=0)} + \ldots \\
&\quad + \log \frac{P(X_d=x_d|Y=1)}{P(X_d=x_d|Y=0)} \\
&= w_0 + w_1(x_1) + \ldots + w_d(x_d)
\end{aligned}
$$

- All of the probabilities can be calculated in a single scan of the dataset
- They also propose shrinking the estimates with a logistic regression $\beta_0 + \beta_1 w_1(x_1) + \ldots + \beta_d w_d(x_d)$

Evidence balance sheet

Evidence in favor of chronic		Evidence in favor of acute	
Tingle	+22	Prior	-199
Depressed	+16	No leg pain	-2
Attention problem	+61	No hearing loss	-7
Insurance company = DMAB	+128	Marital status: Widow	-25
Lawyer	+174	French speaking	-28
Total positive evidence	+401	Total negative evidence	-261
Total evidence	+140		
Probability of chronic whiplash	80%		

Conclusions

- Partnering prediction models with experts and the public offers opportunities to improve care, reduce costs, and avoid errors
- Significant barriers remain. We need to make them capture complexity, widely applicable, and understandable
- What better validation of progress is there than the CBS Early Show?