Classifier Technology and the Illusion of Progress

Elizabeth Sweeney

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

January 29, 2015

Hand, David J. "Classifier technology and the illusion of progress." Statistical Science 21.1 (2006): 1-14.

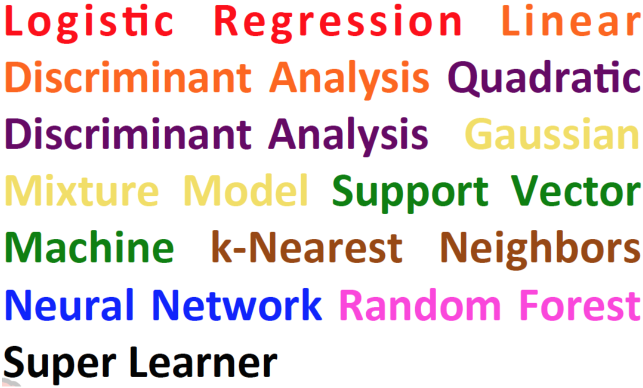
The Classification Task

1. We have random variables (X, Y) such that $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$.
2. Let $g : \mathbb{R}^d \to \{0, 1\}$
3. The error probability is $P_g\,(X, Y) = \mathbb{P}\{g(X) \neq Y\}$
4. The classification task is to find g that minimizes $P_g\,(X, Y)$
5. The functions g are called “classifiers”

The Supervised Classification Task

1. In practice we have a training sequence $\xi_n = ((X_1, Y_1), (X_2, Y_2), ...(X_n, Y_n))$ where (X_k, Y_k) are assumed to be iid from (X, Y)
2. We then estimate Y from $g_n(X, \xi_n)$, minimizing $\mathbb{P}\{g_n(X, \xi_n) \neq Y\}$

Classifiers



Logistic Regression Linear Discriminant Analysis Quadratic Discriminant Analysis Gaussian Mixture Model Support Vector Machine k-Nearest Neighbors Neural Network Random Forest Super Learner

"A large number of comparative studies have been conducted in attempts to establish the relative superiority of these [classification] methods. This paper argues that these comparisons often fail to take into account important aspects of real problems, so that apparent superiority of more sophisticated methods may be something of an illusion"

(Hand et al., 2006)

Outline

1. Marginal Improvements
2. Design Sample Selection
3. Problem Uncertainty
4. Interpreting Empirical Comparisons

Marginal Improvements

1. Large gains in predictive accuracy are won using relatively simple models – potential gains decrease in size as the modeling process is taken further.
2. Extra accuracy is achieved by modeling "minor" aspects of the distribution

Marginal Improvements: Linear Regression Example

A simple regression case, with response variable y and predictors $\mathbf{x} = (x_1, ...x_d)^T$. The correlation matrix for $\left(\mathbf{x}^T, y\right)$ is

$$\Sigma = \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right] = \left[\begin{array}{cc} (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^T & \tau \\ \tau^T & 1 \end{array} \right] \tag{1}$$

So the correlation between each pair of predictors is ρ and between each predictor and the response is τ. We assume $\tau, \rho \geq 0$.

Marginal Improvements: Linear Regression Example

The conditional variance of y given $\mathbf{x}$ is

$$V(d) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \tag{2}$$

as

$$\Sigma_{11}^{-1} = \frac{1}{1-\rho}\left\{\mathbf{I} - \frac{\rho\mathbf{1}\mathbf{1}^T}{1+(d-1)\rho}\right\} \tag{3}$$

we have that

$$V(d) = 1 - \frac{d\tau^2}{1-\rho} + \frac{\rho d^2\tau^2}{(1+(d-1)\rho)(1-\rho)} \tag{4}$$

Marginal Improvements: Linear Regression Example

Let $X\left(d+1\right)$ be the decrease in the conditional variance of y given $\mathbf{x}$ after adding another predictor to the model.

$$\begin{aligned} X\left(d+1\right) &= V\left(d\right) - V\left(d+1\right) \qquad (5) \\ &= \frac{\tau^2}{1-\rho} + \frac{\rho\tau^2}{1-\rho}\left[\frac{d^2}{1+(d-1)\rho} - \frac{(d+1)^2}{1+d\rho}\right] \qquad (6) \end{aligned}$$

Marginal Improvements: Linear Regression Example

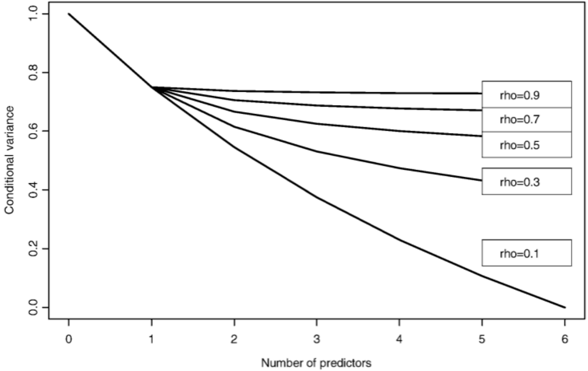
$$X\left(d+1\right)=\frac{\tau^2}{1-\rho}+\frac{\rho\tau^2}{1-\rho}\left[\frac{d^2}{1+(d-1)\rho}-\frac{(d+1)^2}{1+d\rho}\right] \tag{7}$$

Case #1: The predictor variables are uncorrelated, i.e. $\rho = 0$

$$X\left(d=1\right)=\tau^2 \tag{8}$$

Marginal Improvements: Linear Regression Example

Case #2: The predictor variables are correlated, i.e. $\rho > 0$



Marginal Improvements: Effectiveness of Simple Classifiers

Comparison of Fisher's Linear Discriminant Analysis (LDA) and the "best" method in the literature on 10 datasets.

Marginal Improvements: Effectiveness of Simple Classifiers

Fisher's Linear Discriminant Analysis (Hastie et al., 2009)

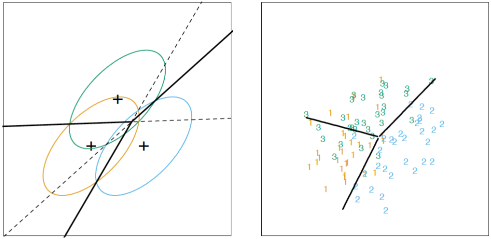


FIGURE 4.5. *The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing 95% of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former). On the right we see a sample of 30 drawn from each Gaussian distribution, and the fitted LDA decision boundaries.*

Marginal Improvements: Effectiveness of Simple Classifiers

1. Best method misclassification rate (m_T)
2. LDA misclassification rate (m_L)
3. Default rate (m_0) – classify all points to the class with the highest prior probability
4. Prop linear

$$\frac{m_0 - m_L}{m_0 - m_T} \tag{9}$$

Marginal Improvements: Effectiveness of Simple Classifiers

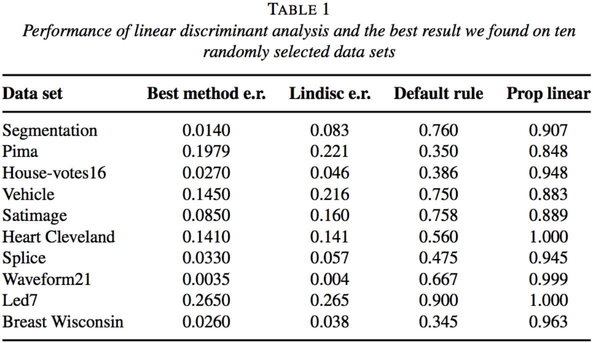


TABLE 1

Performance of linear discriminant analysis and the best result we found on ten randomly selected data sets

Data set	Best method e.r.	Lindisc e.r.	Default rule	Prop linear
Segmentation	0.0140	0.083	0.760	0.907
Pima	0.1979	0.221	0.350	0.848
House-votes16	0.0270	0.046	0.386	0.948
Vehicle	0.1450	0.216	0.750	0.883
Satimage	0.0850	0.160	0.758	0.889
Heart Cleveland	0.1410	0.141	0.560	1.000
Splice	0.0330	0.057	0.475	0.945
Waveform21	0.0035	0.004	0.667	0.999
Led7	0.2650	0.265	0.900	1.000
Breast Wisconsin	0.0260	0.038	0.345	0.963

Marginal Improvements: Neural Networks

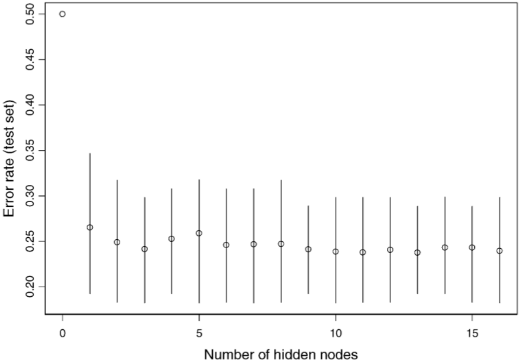


FIG. 2. *Effect on misclassification rate of increasing the number of hidden nodes in a neural network to predict the class of the sonar data.*

Design Sample Selection

1. In classification, we make the assumption that the data in the design set are randomly drawn from the same distribution as those to be classified in the future.
2. Issue arise with **population drift** and **sample selectivity bias**.

Design Sample Selection: Population Drift

1. Population distributions are non-stationary.
2. We make the assumption that our test sample is the same as the sample to be encountered in the future.
3. Caution against putting too much weight in a classifier on one variable, as the relationship may change over time.

Design Sample Selection: Population Drift Example 1

The validation data consists of labels “good” and “bad” and the value of 17 predictor variables for 92,258 customers taking our unsecured personal loans with 24-month term given by a major UK bank during the period of January 1st 1993 to November 30th 1997, with 8.86% of customers belonging to the “bad” class.

The model is trained on data just preceding this time period.

Design Sample Selection: Population Drift Example 1

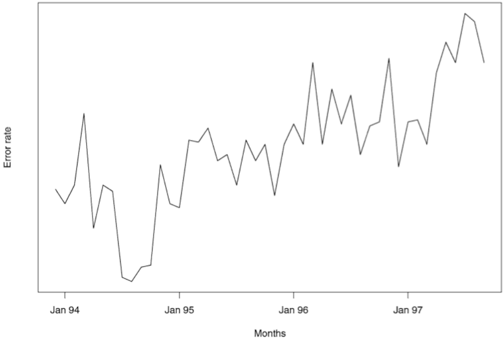


FIG. 4. *Evolution of misclassification rate of a classifier built at the start of the period.*

Design Sample Selection: Population Drift Example 2

A set of $60,000$ customers.

For the design set customers $1, 3, 5, 7, ..., 4999$ were used.

The the classifier is applied to alternate customers, beginning with the second, up to the $60,000^{th}$ customer. (i.e. different customers were used for designing and testing)

Design Sample Selection: Population Drift Example 2

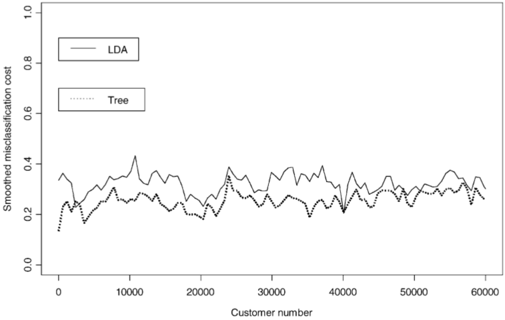


FIG. 5. *Lowess smooths of cost-weighted misclassification rate for a tree model and LDA applied to customers* $2, 4, 6, \ldots, 60{,}000$.

Design Sample Selection: Sample Selectivity Bias

1. Why should we model the subtler aspects of the distributions if the design sample is drawn from a distribution distorted in some way from the original sample
2. Example: Data sampled from one hospital, model to be generalized to other hospitals

Problem Uncertainty

Other ways in which classification can go awry....

1. Error in Class Labels
2. Arbitrariness in the Class Definition
3. Optimization Criteria and Performance Assessment

Problem Uncertainty: Error in Class Labels

Assumption of classical supervised classification paradigm is that there are no errors in the class labels. Let $p\left(1|\mathbf{x}\right)$ and $p\left(2|\mathbf{x}\right)$ be the true posterior class probabilities. Let δ be the small proportion of each class incorrectly believed to come from the other class. Let $p^*(1|\mathbf{x})$ denote our apparent posterior probability for class 1. Then we will have

$$p^*(1|\mathbf{x}) = (1-\delta)p\left(1|\mathbf{x}\right) + \delta p\left(2|\mathbf{x}\right) \tag{10}$$

Problem Uncertainty: Error in Class Labels

Let $r(x)$ be the true odds ratio

$$r(x) = \frac{p(1|\mathbf{x})}{p(2|\mathbf{x})} \tag{11}$$

then our apparent odds ratio is

$$\begin{aligned} r^*(x) &= \frac{p^*(1|\mathbf{x})}{p^*(2|\mathbf{x})} \qquad (12) \\ &= \frac{r(\mathbf{x}) + \epsilon}{\epsilon r(\mathbf{x}) + 1} \qquad (13) \end{aligned}$$

with $\epsilon = \frac{\delta}{1-\delta}$

Problem Uncertainty: Error in Class Labels

Let the true optimal decision surface be $r(x) = k$. Then the optimal decision surface when errors are present will be $r^*(x) = k^*$, then $k^* = \frac{(k+\epsilon)}{(\epsilon k+1)}$. In the case of equal misclassification costs $k = k^* = 1$, so this does not matter. But, when classification costs are not equal this will impact your classification.

Problem Uncertainty: Arbitrariness in Class Definition

Assumption of supervised classification is that the classes are well defined – which may not always be true (especially when defining the classes by thresholding a continuous variable)

1. Consumer credit: definition of a person who ”defaults”

Problem Uncertainty: Optimization Criteria and Performance Assessment

The difference between the (1) criterion used to choose the model (2) the criterion used to evaluate its performance and (3) the criterion which actually matters in real applications.

Interpreting Empirical Comparisons

1. Different categories of users may be expected to obtain different rankings of classification methods in different studies
2. People may have a “favorite classifier”
3. Classification methods may do well on a dataset just by chance

Interpreting Empirical Comparisons: Ping-Pong Theorem

“If we reveal to Professor Breiman the performance of our best model and gave him our data, then he could develop an algorithmic random forest, which would outperform our model. But, if he reveals to us the performance of his model, then we could develop a segmented scorecard which would outperform his model”

– Hoadley

References

Hand, D. J. et al. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.