Bayes Optimality of Wavelet-Based Discrimination

Woojin Chang, Seong-Hee Kim, and Brani Vidakovic
Seoul University and Georgia Institute of Technology

ISBA 2004
Viña del Mar, Chile
May 25, 2004

Overview

- Talk about classifying Y into one of two classes labeled by 0 or 1, by taking into account predictor X.
- Definitions and Notation. Bayes Discriminators
- Wavelet-Based Approximation
- Bayes Optimality (or $\mathbb{L}_2$-Consistency) of the Wavelet-based Classifier.
- Simulations and Paper Production Example

Definitions

- $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$.

$\mu(A) = P(X \in A),\ A \in \mathcal{B};\ \eta(x) = P(Y = 1|X = x) = E(Y|X = x).$

- Pair (μ, η) uniquely determines joint distribution of (X, Y).
- Any function $g : \mathbb{R}^d \to \{0, 1\}$ is a classifier.
- $\boxed{\text{Bayes Classifier: } g^*(x) = \mathbf{1}(\eta(x) > 1/2).}$
- $L(g) = P(g(X) \neq Y)$. [Error, Risk, Misclassification Probability]
- Result: $(\forall g)\ L(g^*) \leq L(g)$. $L^* = L(g^*)$ Bayes Error [Risk, Probability].

Definitions, contd

■ Assume density of X exists, $X \sim f$. Let f_0 and f_1 be class-conditional densities, i.e., densities for $X|Y = 0$ and $X|Y = 1$.

■ Let $\pi = P(Y = 1)$ and $1 - \pi = P(Y = 0)$ be class-probabilities.

■ Function $\alpha(x) = \pi f_1(x) - (1 - \pi) f_0(x)$ has representation $(2\eta(x) - 1) f(x)$.

■ $\boxed{\text{Bayes Classifier: } g^*(x) = \mathbf{1}(\alpha(x) > 0).}$

- $L^* = 1/2 - 1/2 E(|2\eta(X) - 1|)$
- $L^* = \int ((1 - \pi) f_0 \wedge \pi f_1) dx$
- $\pi = 1/2, \quad L^* = 1/2 - 1/4 \int |f_0(x) - f_1(x)| \, dx$

Definitions, contd

- $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training set. Let X be a new observation.
- $g_n(X) = g_n(X, D_n)$, a sequence of classification rules.
- $L_n = P(Y \neq g_n(X, D_n)|D_n)$.

$\mathbb{E}L_n = P(Y \neq g_n(X))$ determined by distribution (X, Y) and classifier g_n.

Classifier g_n is consistent (weakly): $\lim_{n\to\infty} \mathbb{E}L_n = L^*$.

Classifier g_n is consistent (strongly): $\lim_{n\to\infty} L_n = L^*, a.s.$

- Devroye, Györfi, Lugosi (1996)

Fourier Series Classifiers

- Assume $f \in \mathbb{L}_2$. $\quad f \in \mathbb{L}_2 \Rightarrow \alpha = \pi f_1 - (1-\pi) f_0 \in \mathbb{L}_2$.
- Fourier Series Classifier: Classify $X = x$ to be in class 0 if $\sum_{j=1}^{k_n} a_{n,j}\psi_j(x) < 0$. $\quad [\alpha(x) = \sum_{j=1}^{\infty} a_j \psi_j(x)]$
- Trigonometric basis, Legandre polynomials, Hermite functions, Laguerre basis.
- Van Ryzin (1966), Greblicki (1981); Greblicki and Rutkovski (1981), Greblicki and Pawlak (1982, 1983).
- Wavelets?

Benefits of Wavelets: Locality, Fast Calculation of Fourier Coefficients $a_{n,j}$, Control of smoothness.

Consistency result uses $|\psi_j(x)| \leq B$.

Wavelets

■ Multiresolution analysis (MRA) generated by the function ϕ. Functions $\phi_{J,k}(x) = 2^{J/2}\phi(2^J x - k),\ k \in Z\}$ span V_J, a subspace of $\mathbb{L}_2$. The subspaces V_J are nested, i.e., $V_J \subset V_{J+1}$.

■ Wavelets $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), j, k \in Z$ span detail subspaces $W_j = V_{j+1} \ominus V_j$. If $\alpha \in \mathbb{L}_2$,

$$\boxed{\alpha(x) = \sum_{k\in Z} c_{J,k}\phi_{J,k}(x) + \sum_{j\geq J}\sum_{k\in Z} d_{j,k}\psi_{j,k}(x).}$$

■ A *raw* wavelet-based linear classifier, $\hat{g}_J$, is defined as

$$\hat{g}_J(x) = \mathbf{1}(\hat{\alpha}_J(x) > 0),$$

where $\hat{\alpha}_J(x)$ is an estimator of the projection of α on V_J, i.e., an estimator of $\alpha_J(x) = \sum_{k\in Z} c_{J,k}\phi_{J,k}(x)$.

■ The coefficients

$$c_{J,k} = \int_R (2\eta(x) - 1) f(x) \phi_{J,k}(x)\, dx = E[(2\eta(X) - 1)\phi_{J,k}(X)]$$

are estimated by their empirical counterpart

$$\hat{c}^n_{J,k} = \frac{1}{n} \sum_{i=1}^{n} (2Y_i - 1)\phi_{J,k}(X_i).$$

$$\hat{\alpha}_{n,J}(x) = \sum_k \hat{c}^n_{J,k} \phi_{J,k}(x)$$

$$\hat{g}_{n,J}(x) = \mathbf{1}(\hat{\alpha}_{n,J}(x) > 0).$$

■ Let $\hat{L}_n(J) = P(Y \neq \hat{g}_{n,J}(X, D_n)|D_n)$ be the error probability of $\hat{g}_{n,J}$.

■ The estimator $\hat{g}_{n,J}(x)$ is consistent $\longrightarrow$

Theorem 1. Assume that the density for X, f, is compactly supported and belongs to L_∞. Let $J = J(n)$ be the multiresolution level depending on the sample size n in the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Let K be the number of coefficients $\hat{c}^n_{J,k}$ in $\hat{\alpha}_{n,J}(x)$. If

$$J \to \infty \quad \textit{and} \quad \frac{K}{n} \to 0 \quad \textit{as } n \to \infty$$

then the wavelet-based classifier

$$\hat{g}_{n,J}(x) = \mathbf{1}(\hat{\alpha}_{n,J}(x) > 0)$$

is consistent, i.e.,

$$\lim_{n\to\infty} \mathbb{E}\hat{L}_n(J) = L^* \ .$$

Regularized Wavelet Classifier

$$\hat{\alpha}_{n,J}(x) = \sum_{k\in Z} \hat{c}^n_{J,k}\phi_{J,k}(x) = \sum_{k\in Z} \hat{c}^n_{J_0,k}\phi_{J_0,k}(x) + \sum_{J_0\le j<J}\sum_{k\in Z} \hat{d}_{j,k}\psi_{j,k}(x)$$

■ Regularized wavelet representation

$$\tilde{\alpha}_{n,J,\lambda}(x) = \sum_{k\in Z} \hat{c}^n_{J_0,k}\phi_{J_0,k}(x) + \sum_{J_0\le j<J}\sum_{k\in Z} d^*_{j,k}\psi_{j,k}(x)$$

$d^*_{j,k} = (|\hat{d}_{j,k}| - \lambda)_+$; universal threshold $\lambda = \sqrt{2\log K}\ \hat{\sigma}$.

■ Regularized wavelet classifier

$$\tilde{g}_{n,J,\lambda} = \mathbf{1}(\tilde{\alpha}_{n,J,\lambda} > 0).$$

■ Daubechies-Lagarias: Calculating $\phi(x)$ and $\psi(x)$ for any x and any ON wavelet basis.

Theorem 2. Let f, J and K be as in Theorem 1 and let J_0 be multiresolution level such that $J_0 < J$. Let K^* be number of coefficients in detail levels, $J_0 < j < J$. The regularized wavelet-based classifier $\tilde{g}_{n,J,\lambda} = \mathbf{1}(\tilde{\alpha}_{n,J,\lambda} > 0)$ is consistent if

$$K^*(J - J_0)(\max_{J_0 \le j < J} {d^*_{jk}}^2) = o(1), \;\; J, J_0 \to \infty,$$

and

$$\frac{K}{n} \to 0 \quad \text{as } n \to \infty.$$

Empirical Optimality Measures.

■ Empirical errors of classifiers $\hat{g}_{n,J}$ and $\tilde{g}_{n,J,\lambda}$, based on training data set of size n, and evaluated at data $\{(X_j, Y_j), j = 1, \ldots, m\}$:

$$\hat{L}_n(J, m) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}(\hat{g}_{n,J}(X_j) \neq Y_j),$$

and

$$\tilde{L}_n(J, m, \lambda) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}(\tilde{g}_{n,J,\lambda}(X_j) \neq Y_j)$$

Simulation Setup: various m, *Symmlet 8*, $J = 6$ or 7, $J_0 = 3$, and λ universal threshold with the soft-shrinkage policy.

Simulated Data Example: 0 - 1 Discrimination

The training set, $\{(X_i, Y_i), i = 1, \ldots, n\}$, ($n$ is even)

■ The first half: $X_i, i = 1, \ldots, \frac{n}{2}$ sampled from $N(0, 1)$ distribution and $Y_i = 0,\ i = 1, \ldots, \frac{n}{2}$.

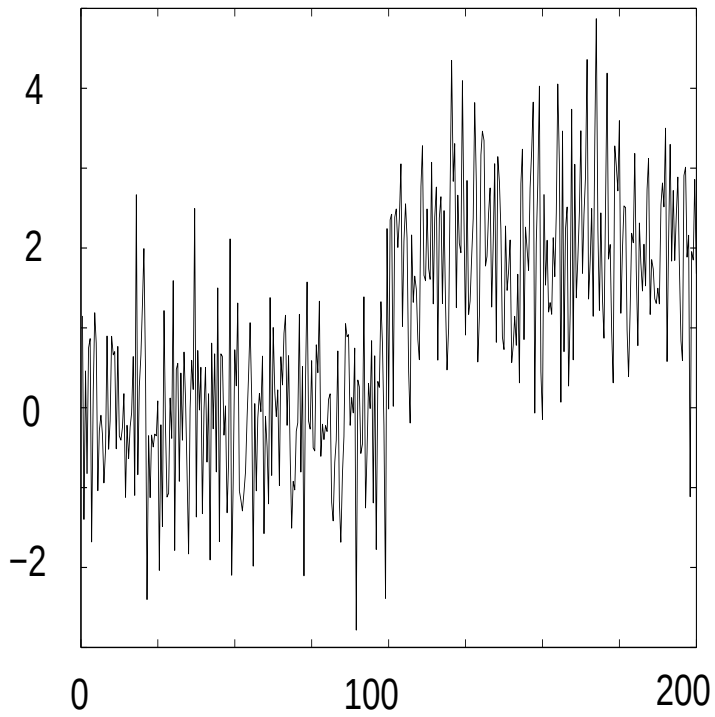
■ The second half: $X_i,\ i = \frac{n}{2} + 1, \ldots, n$ sampled from $N(2, 1)$ distribution and $Y_i = 1, i = \frac{n}{2} + 1, \ldots, n$.

■ The validation set $\{(X_j, Y_j), j = 1, \ldots, m\}$ is generated the same way.

■ Empirical errors $\hat{L}_n(J, m)$,$\tilde{L}_n(J, m, \lambda)$ and the error of the logistic regression classifier,

$$L_n^{\text{logit}}(m) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\Big(\mathbf{1}(f(X_j) > 0.5) \neq Y_j\Big),$$

where f is fitted logistic regression, are compared



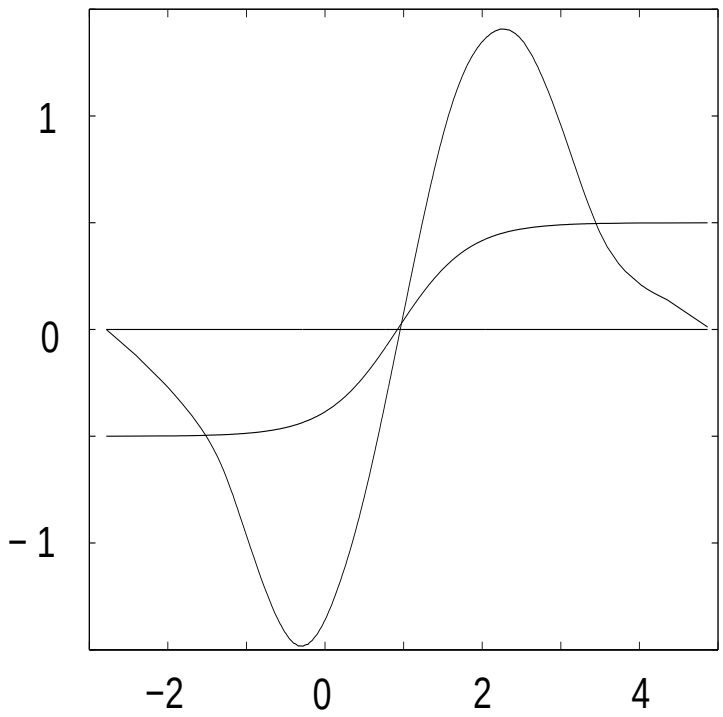


Figure 1: (a) Noisy training data (b) Discriminator functions

Simulational Setup: $m = 200$ (training sample size), $J = 6$ (finest level of detail), and *Symmelet 8* (Daubechies' least asymmetric 8-tap filter).

n	$\hat{L}_n(6, 200)$	$\tilde{L}_n(6, 200, \lambda)$	$L_n^{\text{logit}}(200)$
80	0.272	0.182	0.178
200	0.200	0.179	0.174
400	0.187	0.174	0.171
800	0.169	0.163	0.163
2000	0.160	0.159	0.159

Simulated Data Example: 0 - 1 - 0 Discrimination

■ No automatic logistic regression classifier is possible.

Training data set: $\{(X_i, Y_i), i = 1, \ldots, n\}$, ($n$ is a multiple of 3)

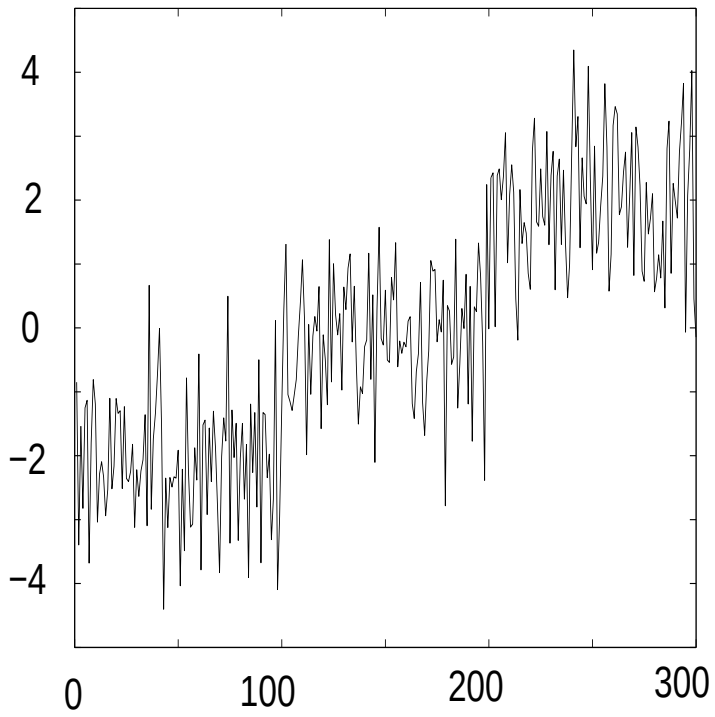
■ The first third of the data: $X_i,\ i = 1, \ldots, \frac{n}{3}$ is generated from $N(-2, 1)$ distribution, with $Y_i = 0,\ i = 1, \ldots, \frac{n}{3}$.

■ The second third of the data: $X_i,\ i = \frac{n}{3} + 1, \ldots, \frac{2n}{3}$ is generated from $N(0, 1)$ distribution, with $Y_i = 1,\ i = \frac{n}{3} + 1, , \ldots, \frac{2n}{3}$.

■ The last third of the data: $X_i,\ i = \frac{2n}{3} + 1, \ldots, n$ is generated from $N(2, 1)$ distribution, and $Y_i = 0,\ i = \frac{2n}{3} + 1, \ldots, n$.

■The evaluation set $\{(X_j, Y_j), j = 1, \ldots, m\}$ is generated in an analogous manner.

Simulational Setup: *Symmlet 8*, $J = 7$, soft with $\lambda = \sqrt{2 \log K}\ \hat{\sigma}$.



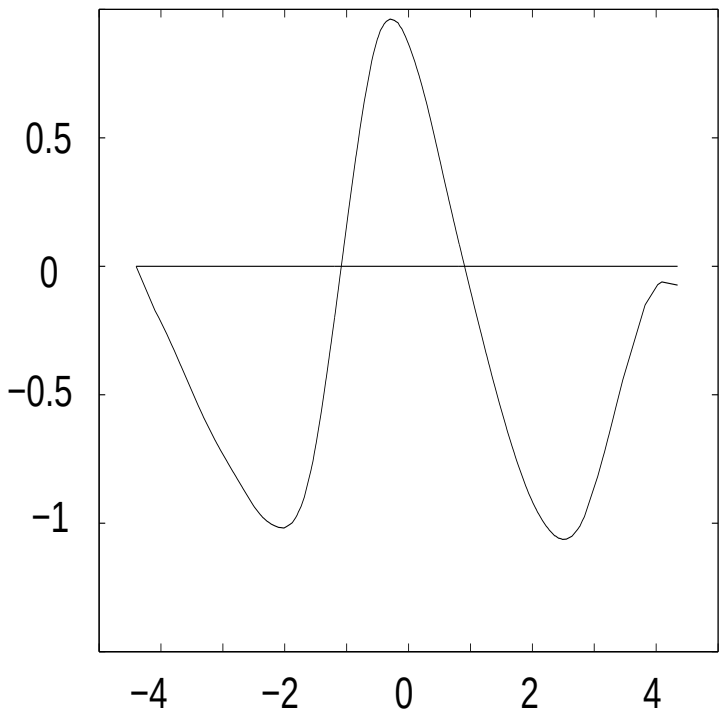


Figure 2: (a) Noisy training data (b) Discriminator function

■ Average empirical errors using training data of size n, $J = 7$, and $m = 300$ training data pairs.

n	$\hat{L}_n(7, 300)$	$\tilde{L}_n(7, 300, \lambda)$
120	0.340	0.213
300	0.288	0.221
600	0.247	0.218
900	0.232	0.212
1200	0.214	0.202

Paper Production Process

■ Data from Pandit and Wu (1993), size 100.

■ B_t and S_t for $t = 1, 2, \ldots, 100$ are the basis weight and the stock flow rate at time t.

$X_t = \hat{B}_t = 0.7B_{t-1} + 0.25S_{t-1}$

$Y_t = \mathbf{1}(39.5 \leq B_t \leq 40)$

We have 99 data points (X_t, Y_t) from the given 100 values of B and S.

■ Predict whether the future basis weight will be "good" or "bad"

■ (X_t, Y_t)'s with odd t: training set

■ The remaining even-index set: validation set.

The empirical error of the classifier, $\tilde{g}_{49,7,\lambda}$ is

$$\tilde{L}_{49}(7, 50, \lambda) = \frac{1}{50} \sum_{t=1}^{50} I(\tilde{g}_{49,7,\lambda}(X_{2t}) \neq Y_{2t}) = 0.18.$$

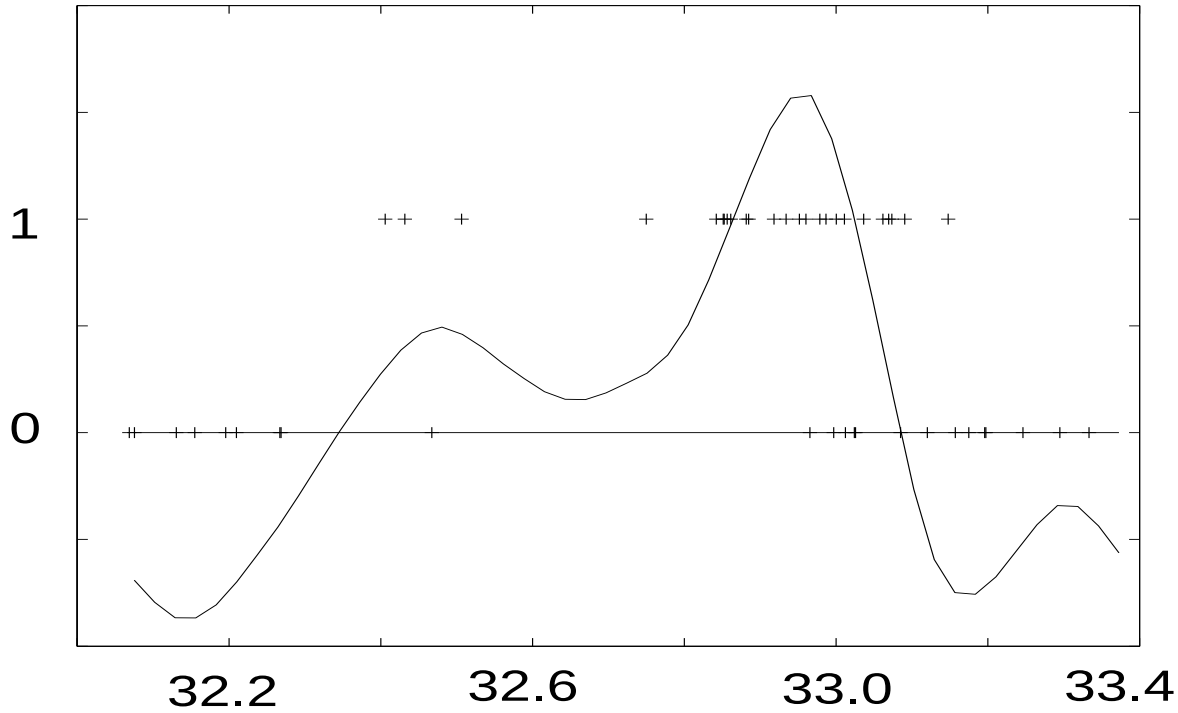


Figure 3: Wavelet Classifier in Paper Production Data

Conclusions

■ Empirical Wavelet-Based Classifier is Consistent [its risk approaches the Bayes risk]

■ Classification is a Trendy Topic: Data Mining, Pattern Recognition.

■ Strong Consistency, Rates of Convergence, Multivariate.

■ Double Bayes: Regularization achieved in Bayesian Fashion. Bayesian Wavelet Shrinkage [Ruggeri, Müller, Vannucci, Clyde, George, ...].

■ Robustness: WRT Wavelet Basis, Choice of J, J_0.

■ Combining classifiers, possibly Triple Bayes.

■ Software Available [MATLAB].