Optimal weighted nearest neighbour classifiers

Richard Samworth

University of Cambridge

r.samworth@statslab.cam.ac.uk

The basic binary classification problem

Let $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n)$ **be i.i.d. pairs in** $\mathbb{R}^d \times \{1,2\}$**, with** $\mathbb{P}(Y=1)=\pi=1-\mathbb{P}(Y=2)$ **and** $(X|Y=r)\sim P_r$**, for** $r=1,2$**.**

A *classifier* is a function $C:\mathbb{R}^d \rightarrow \{1,2\}$**.**

We aim to minimise the *misclassification error rate* or *risk* over a (measurable) set $\mathcal{R}\subseteq\mathbb{R}^d$**:**

$$R_{\mathcal{R}}(C)=\mathbb{P}\Bigl\{(C(X)\neq Y)\mathbb{1}_{\{X\in\mathcal{R}\}}\Bigr\}.$$

Bayes classifier

Let $\bar{P} = \pi P_1 + (1-\pi)P_2$ denote the marginal distribution of X, and $\eta(x) = \mathbb{P}(Y=1|X=x)$ denote the regression function. The Bayes classifier is

$$C^{\mathrm{Bayes}}(x) = \begin{cases} 1 & \textbf{if } \eta(x) \geq 1/2 \\ 2 & \textbf{otherwise.} \end{cases}$$

Its risk is optimal, and is given by

$$R_{\mathcal{R}}(C^{\mathrm{Bayes}}) = \int_{\mathcal{R}} \min\{\eta(x), 1-\eta(x)\}\, d\bar{P}(x)$$

. . .

Bayes classifier

Let $\bar{P} = \pi P_1 + (1-\pi)P_2$ **denote the marginal distribution of** X**, and** $\eta(x) = \mathbb{P}(Y = 1 | X = x)$ **denote the regression function. The Bayes classifier is**

$$C^{\text{Bayes}}(x) = \begin{cases} 1 & \textbf{if } \eta(x) \geq 1/2 \\ 2 & \textbf{otherwise.} \end{cases}$$

Its risk is optimal, and is given by

$$R_{\mathcal{R}}(C^{\text{Bayes}}) = \int_{\mathcal{R}} \min\{\eta(x), 1-\eta(x)\}\, d\bar{P}(x)$$

. . . ***but it can't be used in practice*!**

Nearest neighbour classifiers

Fix and Hodges (1951), Cover and Hart (1967)

Fix $x \in \mathcal{R}$ **and let** $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ **be such that** $\|X_{(1)} - x\| \leq \ldots \leq \|X_{(n)} - x\|$**. The** k**-nn classifier is**

$$\hat{C}_n^{\mathrm{knn}}(x) = \begin{cases} 1 & \textbf{if } k^{-1} \sum_{i=1}^{k} \mathbb{1}_{\{Y_{(i)}=1\}} \geq 1/2 \\ 2 & \textbf{otherwise.} \end{cases}$$

Nearest neighbour classifiers

Fix and Hodges (1951), Cover and Hart (1967)

Fix $x \in \mathcal{R}$ **and let** $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ **be such that** $\|X_{(1)} - x\| \leq \ldots \leq \|X_{(n)} - x\|$. **The** k**-nn classifier is**

$$\hat{C}_n^{\mathrm{knn}}(x) = \begin{cases} 1 & \text{if } k^{-1} \sum_{i=1}^k \mathbb{1}_{\{Y_{(i)}=1\}} \geq 1/2 \\ 2 & \text{otherwise.} \end{cases}$$

Let $\mathbf{w}_n = (w_{ni})_{i=1}^n$ **denote a set of weights normalised so that** $\sum_{i=1}^n w_{ni} = 1$. **The weighted nearest neighbour classifier** (Royall, 1966) **is**

$$\hat{C}_n^{\mathrm{wnn}}(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n w_{ni} \mathbb{1}_{\{Y_{(i)}=1\}} \geq 1/2 \\ 2 & \text{otherwise.} \end{cases}$$

Assumptions

(A.1) **The set $\mathcal{R} \subseteq \mathbb{R}^d$ is a compact d-dimensional manifold with boundary $\partial\mathcal{R}$.**

(A.2) **The set $\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = 1/2\}$ is non-empty. There is an open subset U_0 of $\mathbb{R}^d$ containing $\mathcal{S}$ s.t.: (i) $|\eta(x) - 1/2|$ is bounded away from zero for $x \in U \setminus U_0$, where $U \supseteq \mathcal{R}$ is open; (ii) the restrictions of P_1 and P_2 to U_0 are absolutely continuous w.r.t. Lebesgue measure, with twice continuously differentiable Radon–Nikodym derivatives f_1 and f_2.**

Assumptions

(A.3) **There exists $\rho > 0$ such that $\int_{\mathbb{R}^d} \|x\|^\rho \, d\bar{P}(x) < \infty$. For small $\delta > 0$, the ratio $\bar{P}(B_\delta(x))/(a_d \delta^d)$ is bounded away from zero, uniformly for $x \in \mathcal{R}$.**

(A.4) **For all $x \in \mathcal{S}$, we have $\dot{\eta}(x) \neq 0$, and for all $x \in \mathcal{S} \cap \partial\mathcal{R}$, we have $\dot{\partial\eta}(x) \neq 0$, where $\partial\eta$ denotes the restriction of η to $\partial\mathcal{R}$.**

Allowable weight vectors

Let $s_n^2 = \sum_{i=1}^n w_{ni}^2$ **and** $t_n = n^{-2/d} \sum_{i=1}^n \alpha_i w_{ni}$**, where** $\alpha_i = i^{1+2/d} - (i-1)^{1+2/d}$**. For** $\beta > 0$**, let** $W_{n,\beta}$ **denote the set of sequences of non-negative weight vectors with**

- $s_n^2 \leq n^{-\beta}$**;**
- $t_n^2 \leq n^{-\beta}$**;**
- $\sum_{i=k_2+1}^n w_{ni}/t_n \leq 1/\log n$**, where** $k_2 = \lfloor n^{1-\beta} \rfloor$**;**
- $\sum_{i=k_2+1}^n w_{ni}^2/s_n^2 \leq 1/\log n$**;**
- $\sum_{i=1}^n w_{ni}^3/s_n^3 \leq 1/\log n$**.**

The unweighted k**-nearest neighbour classifier weights belong to** $W_{n,\beta}$ **for small** $\beta > 0$ **provided that** $\max(n^\beta, \log^2 n) \leq k \leq \min(n^{(1-\beta d/4)}, n^{1-\beta})$**.**

Excess risk expansion (S., 2012)

Assume (A.1)–(A.4). For each $\beta > 0$,

$$R_{\mathcal{R}}(\hat{C}_n^{\mathrm{wnn}}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}}) = \gamma_n(\mathbf{w}_n)\{1 + o(1)\}$$

as $n \to \infty$, uniformly for $\mathbf{w}_n \in W_{n,\beta}$, where

$$\gamma_n(\mathbf{w}_n) = B_1 s_n^2 + B_2 t_n^2.$$

Defining the constants

Let $\bar{f} = \pi f_1 + (1-\pi) f_2$, let a_d be the volume of the unit ball in the norm $\|\cdot\|$ in $\mathbb{R}^d$ and let

$$a(x) = \frac{\sum_{j=1}^d c_{j,d}\{\eta_j(x)\bar{f}_j(x) + \frac{1}{2}\eta_{jj}(x)\bar{f}(x)\}}{a_d^{1+2/d}\bar{f}(x)^{1+2/d}},$$

where $c_{j,d} = \int_{v:\|v\|\leq 1} v_j^2\,dv$. Define

$$B_1 = \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|}\,dx_0 \quad \textbf{and} \quad B_2 = \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{\|\dot{\eta}(x_0)\|} a(x_0)^2\,dx_0,$$

where dx_0 denotes the natural volume element of $\mathcal{S}$. Note that $B_1 > 0$, and $B_2 \geq 0$, with equality if and only if a is identically zero on $\mathcal{S}$.

Optimal weights

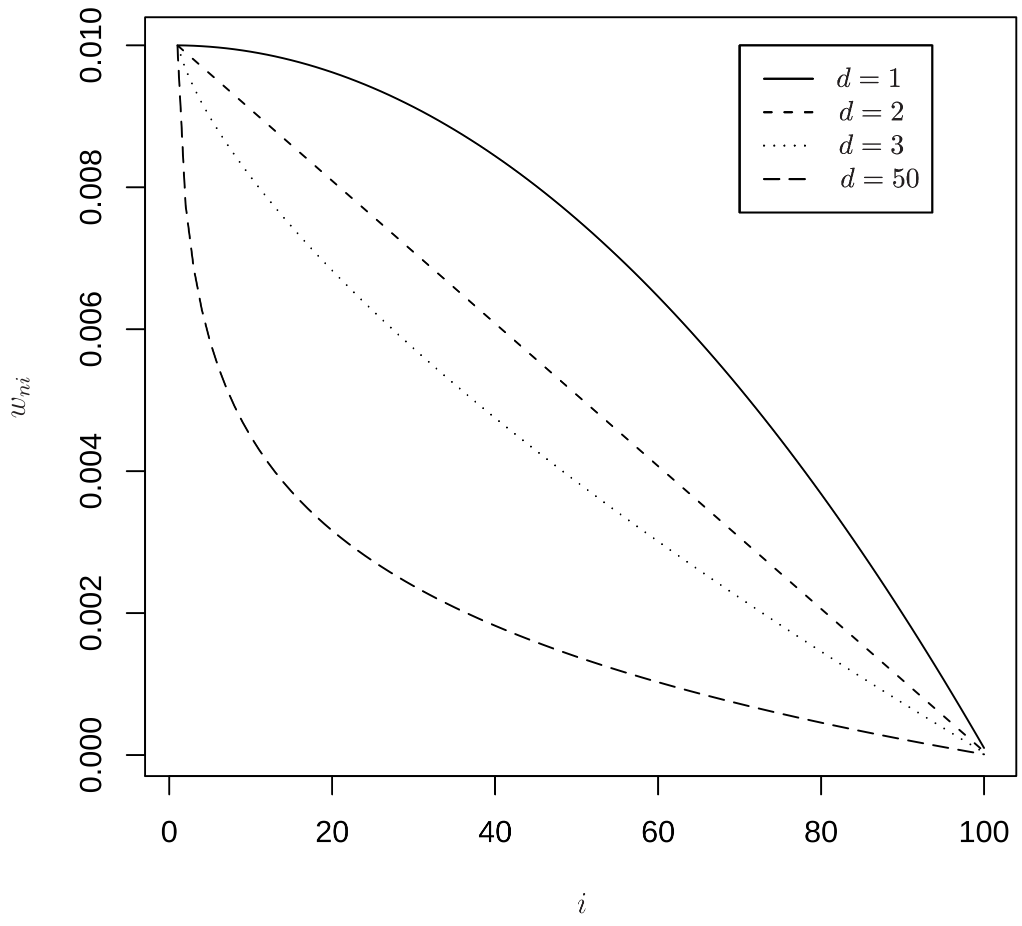
Let

$$k^* = \left\lfloor \left\{ \frac{d(d+4)}{2(d+2)} \right\}^{d/(d+4)} \left(\frac{B_1}{B_2} \right)^{d/(d+4)} n^{4/(d+4)} \right\rfloor,$$

and set

$$w_{ni}^* = \begin{cases} \frac{1}{k^*}\left(1 + \frac{d}{2} - \frac{d\alpha_i}{2(k^*)^{2/d}}\right) & \textbf{for } i = 1, \ldots, k^* \\ 0 & \textbf{for } i = k^* + 1, \ldots, n\textbf{.} \end{cases}$$

Diagram of optimal weights



Optimality statement

For any $\beta > 0$ **and any** $\mathbf{w}_n = (w_{ni})_{i=1}^n \in W_{n,\beta}$**, we have**

$$\liminf_{n\to\infty} \frac{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^*}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})} \geq 1.$$

Moreover, the ratio converges to 1 if and only if both $\sum_{i=1}^n w_{ni}^2 / \sum_{i=1}^n (w_{ni}^*)^2 \to 1$ **and** $\sum_{i=1}^n \alpha_i w_{ni} / \sum_{i=1}^n \alpha_i w_{ni}^* \to 1$**.**

Asymptotic improvement over k-nn

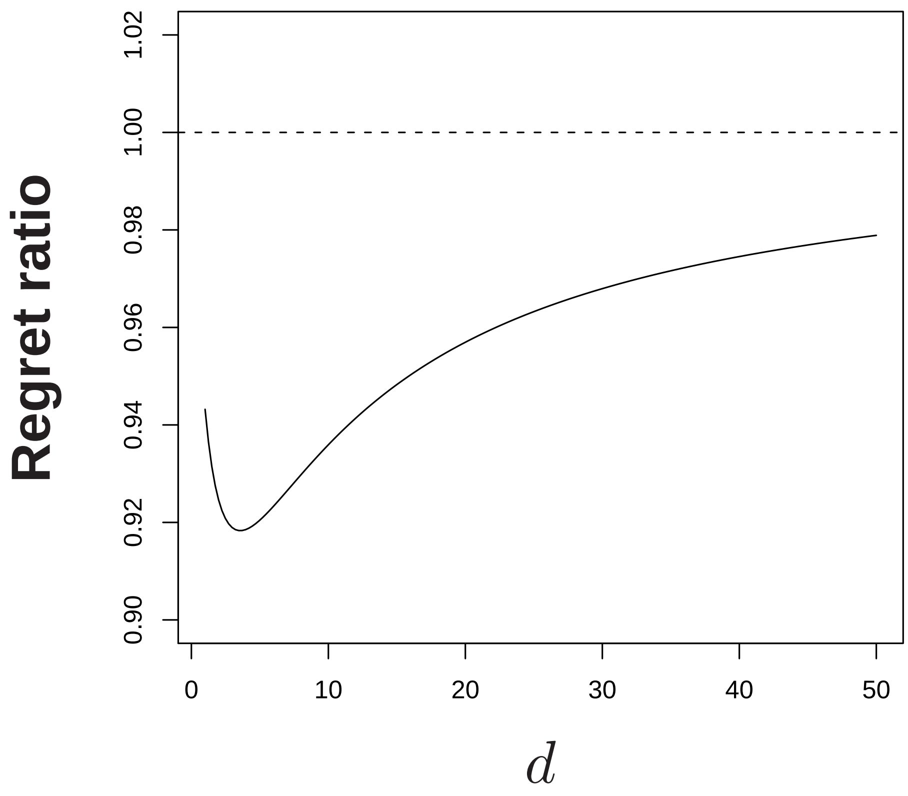
Let $\hat{C}^{\mathrm{knn}}_{n,k}$ denote the unweighted k-nearest neighbour classifier with optimal k (Hall, Park and S., 2008). Then

$$k^* = \left(\frac{2(d+4)}{d+2}\right)^{d/(d+4)} k,$$

and

$$\frac{R_{\mathcal{R}}(\hat{C}^{\mathrm{wnn}}_{n,\mathbf{w}_n^*}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\mathrm{knn}}_{n,k}) - R_{\mathcal{R}}(C^{\mathrm{Bayes}})} \to \frac{1}{2^{2d/(d+4)}}\left(\frac{2d+4}{d+4}\right)^{(2d+4)/(d+4)}.$$

Comparison between asymptotic regrets



The bagged nearest neighbour classifier

(Hall and S.,2005, Biau, Cérou and Guyader, 2010)

The bagged nearest neighbour classifier applies a majority vote to the classifications of a 1-nearest neighbour classifier applied to resamples of the data.

In the infinite-simulation case, for a resample size m,

$$w_{ni}^{\mathrm{b,with}} = \left(1 - \frac{i-1}{n}\right)^m - \left(1 - \frac{i}{n}\right)^m, \quad i = 1, \ldots, n$$

and

$$w_{ni}^{\mathrm{b,w/o}} = \begin{cases} \binom{n-i}{m-1} / \binom{n}{m} & \textbf{for } i = 1, \ldots, n-m+1 \\ 0 & \textbf{for } i = n-m+2, \ldots, n. \end{cases}$$

Intuition for the bnn classifier

Let $q = m/n$ denote the sampling fraction. For large n, both types of bnn classifier are similar to the classifier that places a geometric distribution on the observations:

$$w_{ni}^{\mathrm{Geo}} = \frac{q(1-q)^{i-1}}{1-(1-q)^n}, \quad i = 1, \ldots, n.$$

Write $\hat{C}_{n,q}^{\mathrm{bnn}}$ for any of these three classifiers.

Bagged nearest-neighbour classifier risk

Assume (A.1)–(A.4). If $n^{-(1-\epsilon)} \leq q \leq n^{-\epsilon}$, then

$$R_{\mathcal{R}}(\hat{C}^{\text{bnn}}_{n,q}) - R_{\mathcal{R}}(C^{\text{Bayes}}) = \tilde{\gamma}_n(q)\{1 + o(1)\},$$

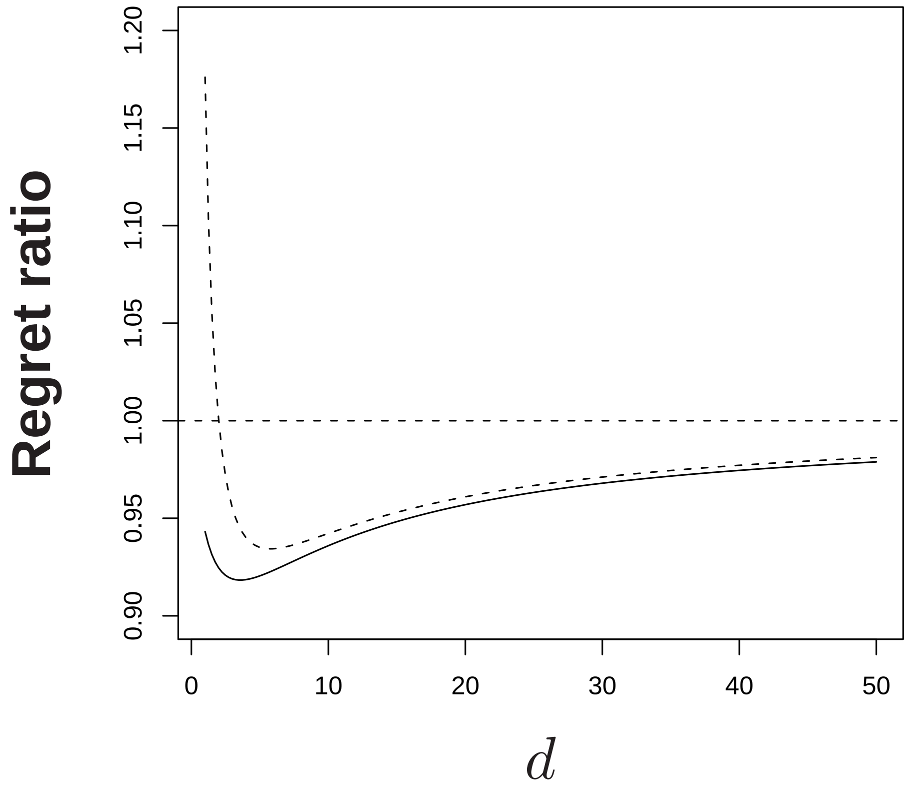
uniformly in q, where

$$\tilde{\gamma}_n(q) = \frac{B_1}{2} q + \frac{B_2\,\Gamma\big(2 + \frac{2}{d}\big)^2}{n^{4/d} q^{4/d}}.$$

If $n^{\beta} \leq k \leq n^{1-\beta}$ and $\hat{C}^{\text{bnn}}_{n,q}$ denotes the bnn classifier with $q = \frac{\Gamma(2+\frac{2}{d})^2}{2^{d/(d+4)}} \frac{1}{k}$, then

$$\frac{R_{\mathcal{R}}(\hat{C}^{\text{bnn}}_{n,q}) - R_{\mathcal{R}}(C^{\text{Bayes}})}{R_{\mathcal{R}}(\hat{C}^{\text{knn}}_{n,k}) - R_{\mathcal{R}}(C^{\text{Bayes}})} \rightarrow \frac{\Gamma\big(2 + \frac{2}{d}\big)^{2d/(d+4)}}{2^{4/(d+4)}}.$$

Further regret comparisons



Summary

- The optimal (non-negative) weights have a relatively simple form
- The improvement over the unweighted k-nearest neighbour classifier can be quantified
- The bagged nearest neighbour classifer is somewhat suboptimal for small d, but close to optimal when d is large
- Improvements in the rate of convergence are possible under stronger smoothness assumptions, provided we allow negative weights.

References

- Biau, G., Cérou, F. and Guyader, A. (2010) On the rate of convergence of the bagged nearest neighbor estimate. *J. Mach. Learn. Res.*, 11, 687–712.
- Cover, T. M. and Hart, P. E. (1967) Nearest neighbor pattern classification, *IEEE Trans. Inf. Th.*, 13, 21–27.
- Fix, E. and Hodges, J. L. (1951) Discriminatory analysis – nonparametric discrimination: Consistency properties. Tech. Rep. 4, Project no. 21-29-004, USAF School of Aviation Medicine, Randolph Field, Texas.
- Hall, P., Park, B. U. and Samworth, R. J. (2008) Choice of neighbor order in nearest-neighbor classification. *Ann. Statist.*, 36, 2135–2152.
- Hall, P. and Samworth, R. J. (2005), Properties of bagged nearest-neighbour classifiers, *J. Roy. Statist. Soc., Ser. B*, 67, 363–379.

- **Royall, R. (1966) *A Class of Nonparametric Estimators of a Smooth Regression Function*. PhD Thesis, Stanford University, Stanford, CA.**

- **Samworth, R. J. (2012), Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, to appear.**