

Exact Bagging of k -Nearest Neighbor
Predictors

Brian Steele and Dave Patterson
Dept. of Mathematical Sciences
University of Montana

Overview

- Definitions
- k -nearest neighbor predictors
- Accuracy estimation
- Bagging and exact bagging
- Comparisons

Definitions

- \mathbf{Z}_n denotes a random sample of n observations from a population

$$\mathcal{P} = \{z_1, z_2, \dots, \}$$

where $z_i = (\mathbf{x}_i, \mathbf{y}_i)$

- \mathbf{x}_i is a covariate vector observed for $z_i \in \mathcal{P}$
- \mathbf{y}_i is observed only for $z_i \in \mathbf{Z}_n$

Objective: predict \mathbf{y} or a function of \mathbf{y} using a predictor η constructed from \mathbf{Z}_n

- Distance between z and $z_i \in \mathbf{Z}_n$ is measured on the covariates
- $\{z_{(1)}, \dots, z_{(k)}\}$ denotes the k -nearest neighbors of z among \mathbf{Z}_n
- The elementary k -nearest neighbor predictor of \mathbf{y} is

$$\hat{\mathbf{y}} = k^{-1} \sum_{i=1}^k \mathbf{y}_{(i)}$$

The classification problem

- \mathcal{P} is partitioned as c classes
- y is a scalar identifying the group to which z belongs
- Suppose that the groups are labeled $1, \dots, g$
- Ψ is the indicator function of an event.
Hence

$$\Psi(y = g) = \begin{cases} 1 & \text{if } z \in G_g \\ 0 & \text{if } z \notin G_g, \end{cases} \quad g = 1, \dots, c.$$

- Suppose that z is drawn at random from \mathcal{P} and that x is observed but not y

- $\Psi(y = g)$ is Bernoulli given x , and

$$E[\Psi(y = g) \mid \mathbf{x}] = P(y = g \mid \mathbf{x})$$

- Let $P_g(\mathbf{x}) = P(y = g \mid \mathbf{x})$

- We prefer to estimate $P_g(\mathbf{x})$ using those training observations with covariates identical to \mathbf{x}
- There may not be any such \mathbf{x}_i so we use the k nearest neighbors
- The k -NN estimator of $P_g(\mathbf{x})$ is

$$\hat{P}_g(\mathbf{x}) = k^{-1} \sum_{i=1}^k \Psi[y_{(i)} = g]$$

- The k -NN estimator predictor of class membership is

$$\eta(\mathbf{x}) = \arg \max_g \hat{P}_g(\mathbf{x})$$

- Supposing that there are no ties, then the realization of $\hat{P}_g(\mathbf{x})$ is one of $0, 1/k, \dots, 1$
- The k -NN classifier is relatively imprecise when k is small
- The cost of choosing a small k is a potential increase in the mean square error of $\hat{P}_g(\mathbf{x})$

Bootstrap aggregation

- Bagging is a method of reducing the variance of a classifier prediction
- Bagging is carried out by drawing B bootstrap samples from \mathbf{Z}_n
- The b th sample produces an estimator of $P_g(\mathbf{x})$ denoted by $P_g^{*b}(\mathbf{x})$ using the k -nearest neighbors among the bootstrap sample

- The bagging estimator of $P_g(\mathbf{x})$ is

$$P_g^{\text{BA}}(\mathbf{x}) = B^{-1} \sum_{b=1}^B P_g^{*b}(\mathbf{x}),$$

- The bagging classifier is

$$\eta_g^{\text{BA}}(\mathbf{x}) = \arg \max_g P_g^{\text{BA}}(\mathbf{x}) \quad (1)$$

- $P_g^{\text{BA}}(\mathbf{x})$ is a Monte Carlo approximation to the exact bagging estimator

$$E_{\hat{F}} P_g^*(\mathbf{x})$$

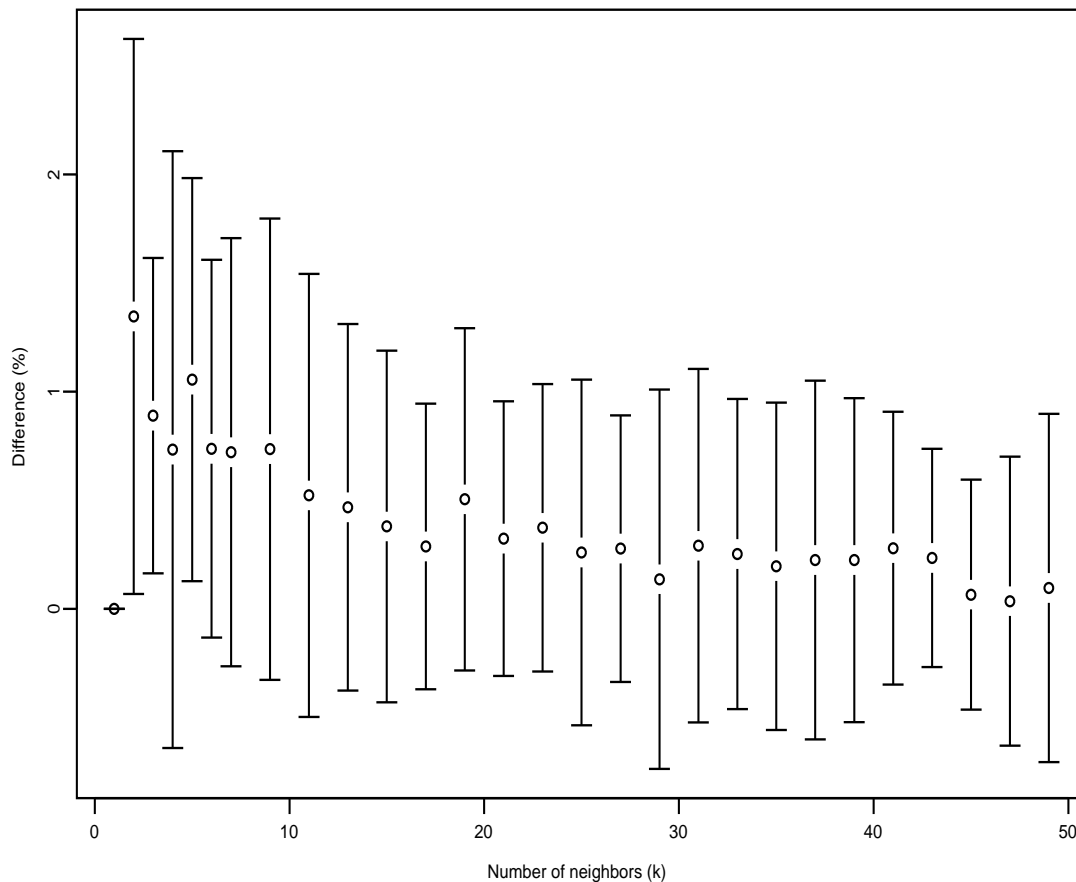
- \hat{F} is the empirical distribution function placing probability mass n^{-1} at each $z_i \in \mathcal{Z}_n$
- $P_g^*(\mathbf{x})$ is the k -NN estimator of $P_g(\mathbf{x})$ obtained from a random sample from \hat{F}

- The exact bagging classifier $E_{\hat{F}}P_g^*(\mathbf{z})$ is usually computationally intractable
- Hence, bagging uses Monte Carlo approximation to approximate $E_{\hat{F}}P_g^*(\mathbf{z})$
- With the k -NN classifier, it is possible to compute $E_{\hat{F}}P_g^*(\mathbf{x})$ analytically

A comparison

- Data were provided by the USDA Forest Service, Northern Region
- 3709 observations on a $g = 15$ class population used for land cover mapping via Landsat imagery (10 covariates)

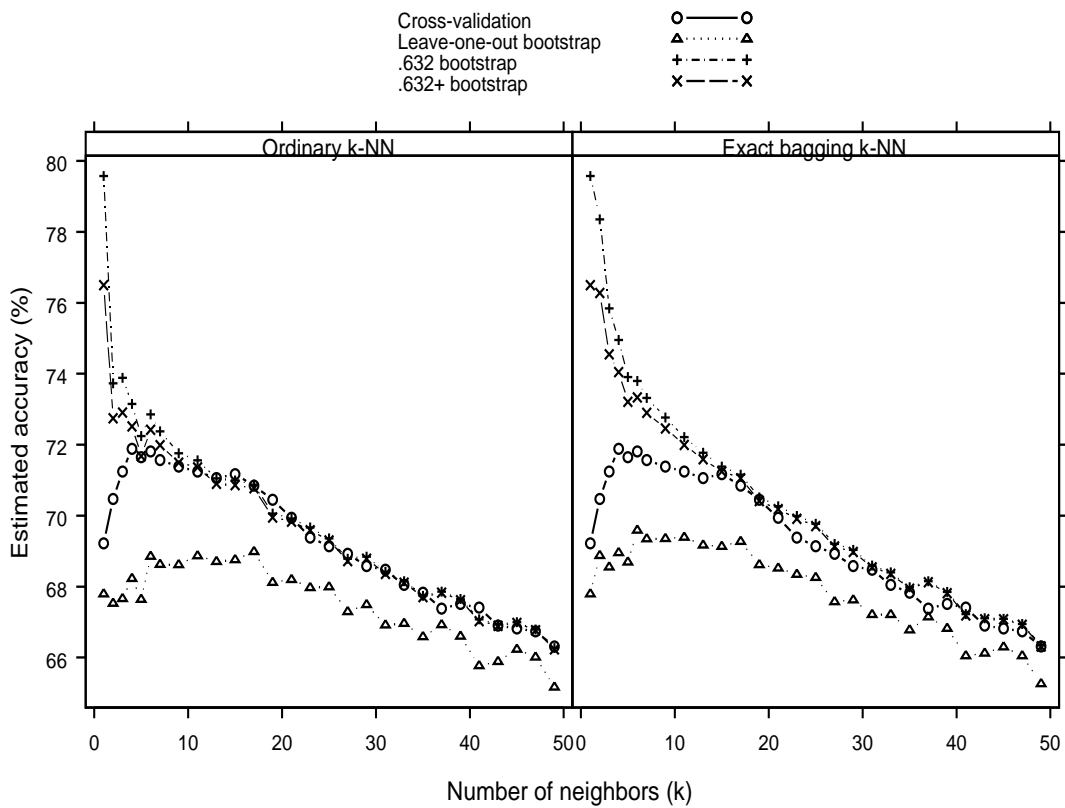
Difference in estimated percent accuracy is plotted against k along with approximate 95% confidence intervals



- This is typical: exact bagging produces small but consistent improvements

A digression

- Accuracy estimation for k -nearest neighbor classifiers requires care
- Cross-validation estimators have small biases but relatively large variances



Estimated accuracy as a function of neighborhood size k

- The .632 and .632+ bootstrap estimators are optimistically biased for small k

- The leave-one-out bootstrap estimator is pessimistically biased
- n -fold cross-validation is most appropriate for assessing the conditional accuracy of η given \mathbf{Z}_n
- Bootstrap estimators are most appropriate for assessing the unconditional accuracy of η

k -NN regression

- y is a continuous random variable
- Use a linear combination of the k nearest covariate vectors $\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(k)}$ to predict y
- $N_k(\mathbf{z})$ denotes the set of the k nearest neighbors of \mathbf{z} among \mathbf{Z}_n

- Assume that in $N_k(\mathbf{z})$,

$$E(y) = \mathbf{x}^T \boldsymbol{\beta}(\mathbf{z})$$

- $\boldsymbol{\beta}(\mathbf{z})$ is specific to $N_k(\mathbf{z})$
- This model is a compromise between a continuously varying $\boldsymbol{\beta}$ and the global parameter of ordinary least squares regression

- Let

$$\psi(z_i) = \Psi[z_i \in N_k(z)]$$

- An estimator for $\beta(z)$ can be obtained by minimizing the prediction error

$$S(z) = \sum_{i=1}^n \psi(z_i) [y_i - \mathbf{x}_i^T \beta(z)]^2$$

- A matrix formulation is

$$S(\mathbf{z}) = [\mathbf{y} - \mathbf{X}\beta(\mathbf{z})]^T \Psi(\mathbf{z}) [\mathbf{y} - \mathbf{X}\beta(\mathbf{z})]$$

where

- $\mathbf{y} = (y_1, \dots, y_n)^T$
- $\mathbf{X} = (x_{ij})$ and
- $\Psi(\mathbf{z})$ is a diagonal with $\psi(\mathbf{z}_1), \dots, \psi(\mathbf{z}_n)$ on the diagonal.

- Differentiation of $S(\mathbf{z})$ with respect to $\beta(\mathbf{z})$ yields the normal equations

$$\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{y} = \mathbf{X}^T \Psi(\mathbf{z}) \mathbf{X} \beta(\mathbf{z})$$

- If $\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{X}$ is full rank, then

$$\hat{\beta}(\mathbf{z}) = [\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{X}]^{-1} \mathbf{X}^T \Psi(\mathbf{z}) \mathbf{y}$$

and

$$\hat{\mathbf{y}} = \mathbf{X} \Psi(\mathbf{z}) \hat{\beta}(\mathbf{z})$$

Problems

- Using only k observations implies the mean square error of \hat{y} will be large
- Presumably, the $\beta(\mathbf{z})$'s vary more smoothly than the k -NN model implies, and information regarding $\beta(\mathbf{z})$ can be extracted from neighbors of \mathbf{z} that are close to \mathbf{z} , but not in the k -neighborhood
- Observations near the boundary of the k -neighborhood may be less informative than those observations that are near the center of the neighborhood.
- $\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{X}$ will not be full rank if $k < p$

Improving on k nearest neighbor regression

- We propose to use the exact bootstrap aggregation approach to smooth the parameter vectors
- Set equal the bootstrap expectations of each side of the normal equations:

$$\begin{aligned} E_{\hat{F}}[\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{y}] &= E_{\hat{F}}[\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{X} \beta(\mathbf{x})] \\ &= E_{\hat{F}}[\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{X}] \beta(\mathbf{x}) \end{aligned}$$

- The solution is

$$\beta^s(\mathbf{z}) = E_{\hat{F}}[\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{X}]^{-1} E_{\hat{F}}[\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{y}].$$

Computing expectations

- Consider

$$E_{\hat{F}}[\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{y}] = \begin{pmatrix} E_{\hat{F}}[\mathbf{x}_1^T \Psi(\mathbf{z}) \mathbf{y}] \\ \vdots \\ E_{\hat{F}}[\mathbf{x}_p^T \Psi(\mathbf{z}) \mathbf{y}] \end{pmatrix}$$

where \mathbf{x}_j is the j th column of \mathbf{X}

- The j th element of this vector is

$$E_{\hat{F}}[\mathbf{x}_j^T \Psi(\mathbf{z}) \mathbf{y}] = \sum_{i=1}^n x_{ij} y_i P_{\hat{F}}[\mathbf{z}_i \in N_k^*(\mathbf{z})]$$

where $N_k^*(\mathbf{z})$ is the set of k closest neighbors of \mathbf{z} among a bootstrap sample drawn from \hat{F}

Another expression

- Suppose that the rows of \mathbf{X} and \mathbf{y} have been arranged according to the distance of x_i from $x, i = 1, \dots, n$
- The ordered matrix \mathbf{X} is \mathbf{X}_o
- The ordered vector \mathbf{y} by \mathbf{y}_o
- The diagonal matrix \mathbf{A}_k contains $P_{\hat{F}}[\mathbf{z}_{(i)} \in N_k^*(\mathbf{z})]$ on the diagonal

Then

$$E_{\hat{F}}[\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{y}] = \mathbf{X}_o \mathbf{A}_k \mathbf{y}_o,$$

$$E_{\hat{F}}[\mathbf{X}^T \Psi(\mathbf{z}) \mathbf{X}] = \mathbf{X}_o^T \mathbf{A}_k \mathbf{X}_o$$

and

$$\beta^s(\mathbf{z}) = (\mathbf{X}_o^T \mathbf{A}_k \mathbf{X}_o)^{-1} \mathbf{X}_o \mathbf{A}_k \mathbf{y}_o$$

Computing $P_{\hat{F}}[z_{(i)} \in N_k^*(z)]$

- This term is the probability that the i th nearest neighbor to z is the r th closest among a bootstrap sample
- Note that

$$P_{\hat{F}}[z_{(i)} \in N_k^*(z)] = \sum_{r=1}^k P_{\hat{F}}[z_{(r)}^* = z_{(i)}]$$

It can be shown that

$$P_{\hat{F}}[z_{(r)}^* = z_{(i)}] = F_{i/n}(r, n - r + 1) - F_{(i-1)/n}(r, n - r + 1)$$

where

$$F_p(r, n) = \frac{n!}{(r-1)!(n-r+1)!} \times \int^p u^{r-1}(1-u)^{n-r} du$$

is the cumulative distribution function of a beta random variable with parameters r and $n-r+1$

Some remarks

- Inverting $\mathbf{X}_o^T \mathbf{A}_k \mathbf{X}_o$ is much less of a problem than inverting $\mathbf{X}^T \Psi(\mathbf{x}) \mathbf{X}$ because
$$\text{rank}(\mathbf{X}_o^T \mathbf{A}_k \mathbf{X}_o) = \min[\text{rank}(\mathbf{X}_o^T \mathbf{X}_o), \text{rank}(\mathbf{A}_k)]$$
- Both k -nearest neighbor and exact bagging k -nearest neighbor are varieties of local linear regression
- Resampling methods are necessary for estimating prediction error and intervals

Some examples

- Time series smoothing: for $z_t \in N_k(z)$ assume the model

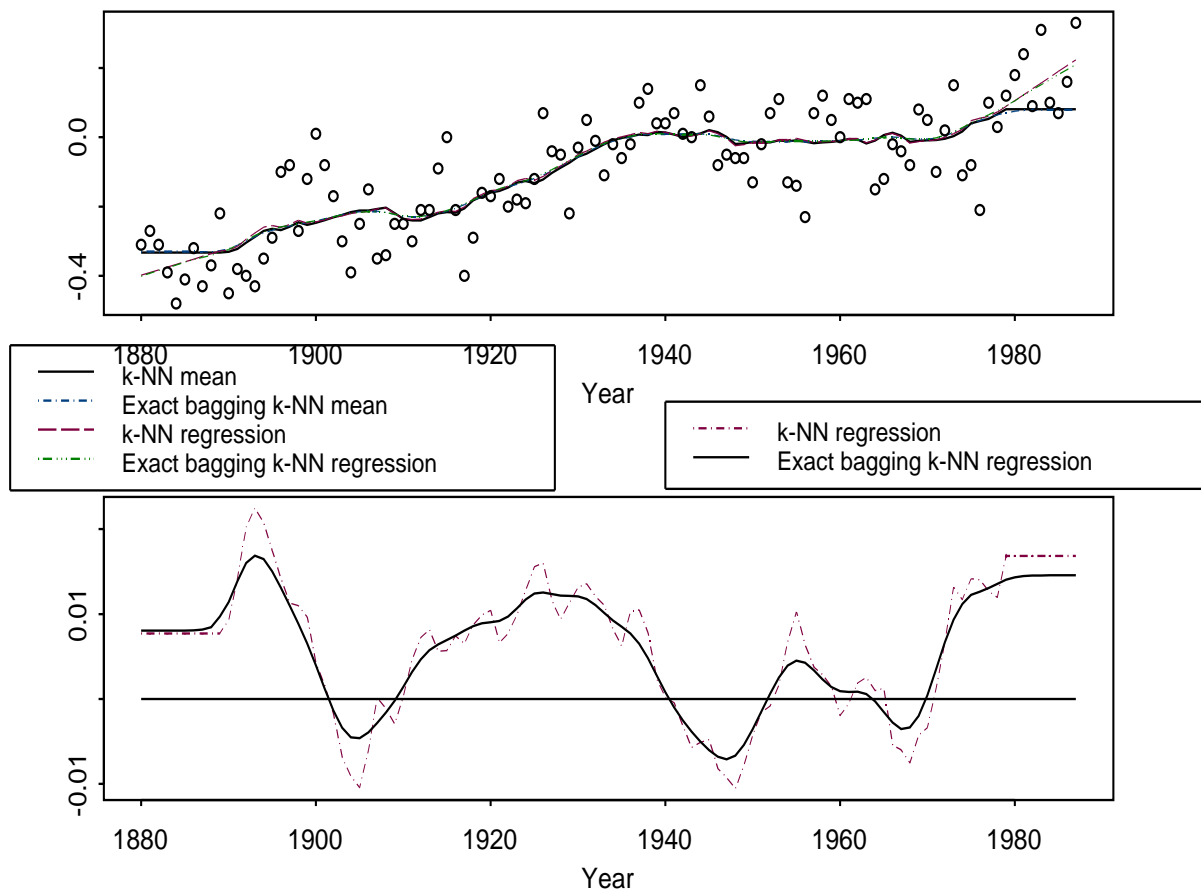
$$E(y_t) = \beta_0(t) + \beta_1(t)t$$

where y_t is the response variable at time t and $\beta_0(t)$ and $\beta_1(t)$ are intercept and slope parameters at time t

- $\hat{\beta}_1(t)$ is the estimated rate of change as a function of time

Temperature data for the Northern hemisphere.

$k = 18$ *



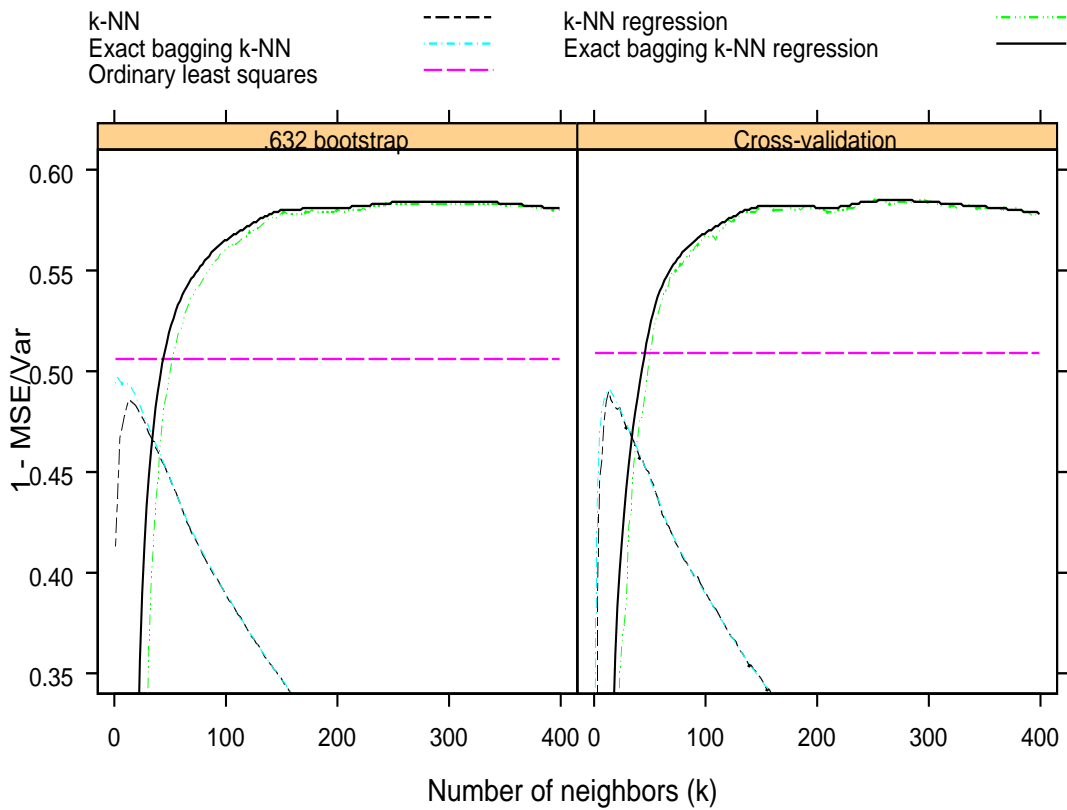
* Jones, P.D. (1988) *Journal of Climatology*, **1**, 654-60

USDA Forest Service Forest and Inventory Analysis (FIA) data from Northern Idaho ($n = 973$)

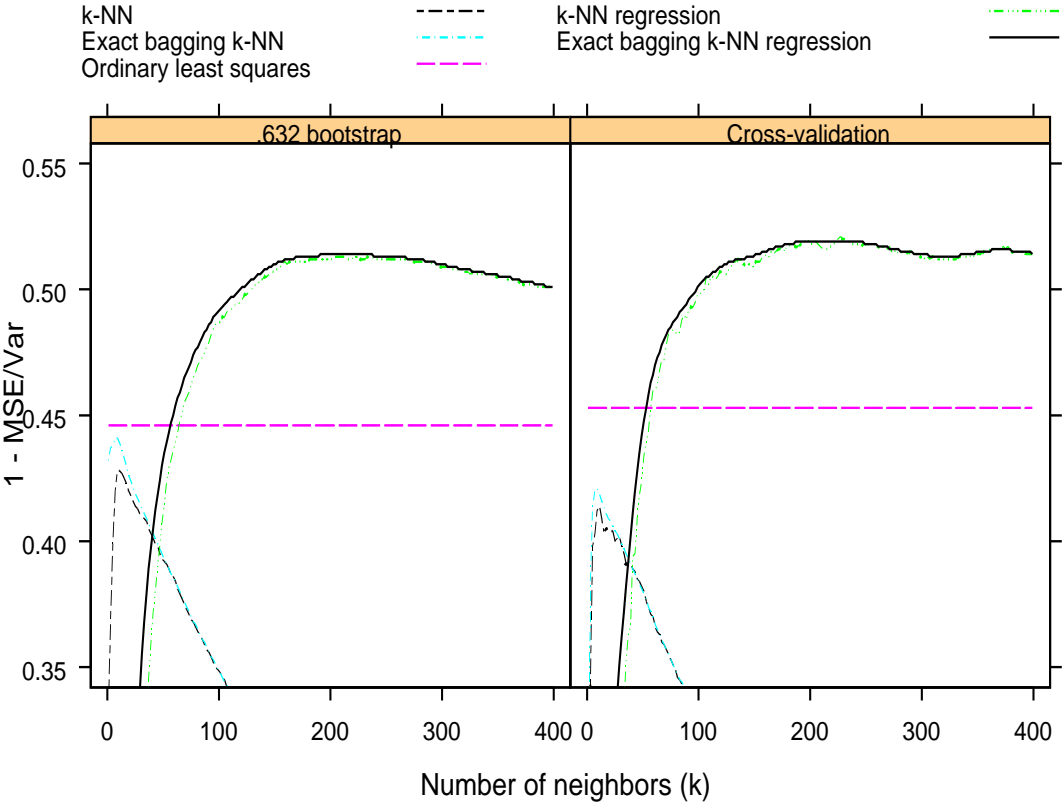
- Monitoring program for forest management with systematically located permanent plot installations
- Objective: predict forest attributes away from plot locations using remotely sensed predictor variables such as Landsat 7 TM imagery and topographic variables
- For comparative purposes, we measure accuracy according to

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Basal area (entire plot, plots with less than 20 sq. ft. were excluded)

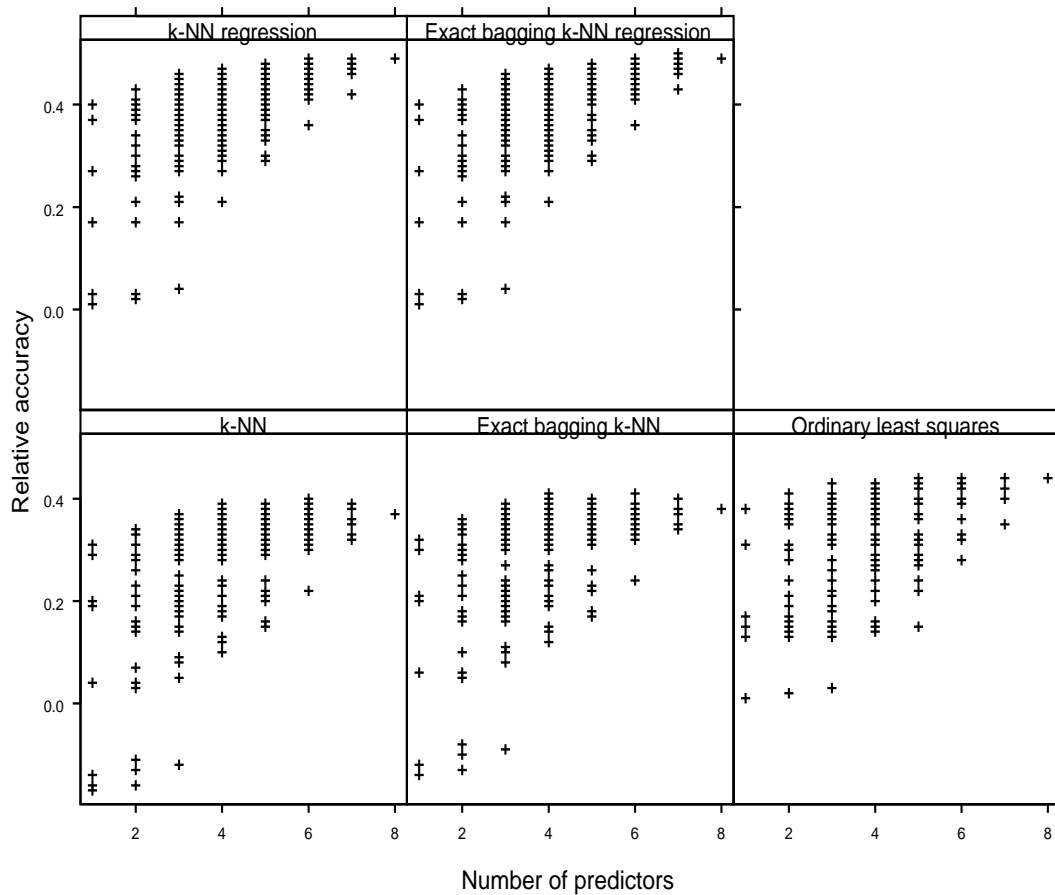


Canopy cover (entire plot)



We fit all possible models (2^p) and use the leave-one-out bootstrap estimates of prediction error ($B = 200$) to identify a best model

Model fit plotted against p



Remarks

- Exact bagging provides modest and consistent improvements in accuracy
- When k is large, there is little difference between ordinary k - and exact bagging k -nearest neighbor predictors
- k -nearest neighbor regression is potentially valuable in remote sensing application

Some references

- Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123-140.
- Hutson, A.D., Ernst, M.D. (2000) The exact bootstrap mean and variance of an L -estimator. *J.R.Statist. Soc. B*, **62**, 89-94.
- Steele, B. M., Patterson, D.A. (2000) Ideal bootstrap estimation of expected prediction error for k -nearest neighbor classifiers: Applications for classification and error assessment. *Statistics and Computing*, **10**, 349-55.

Additional references

- Efron, B., Tibshirani, R. (1997) Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, **92**, 548-560.
- Hall, P., Sarnworth, R.J. (2005) Properties of bagged nearest neighbour classifiers. *J.R.Statist. Soc. B*, **67**, 363-379.
- Hastie, T, Tibshirani, R., Friedman, J. (2001) *The Elements of Statistical Learning*. Springer, New York.
- Steele, B.M., Patterson, D.A., Redmond, R.L. (2003) Toward estimation of map accuracy without a probability sample. *Environmental and Ecological Statistics*, **10**, 333-356.