

UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK-IN4300/STK-IN9300 — Statistical learning methods in Data Science

Day of examination: Friday, December 3rd, 2021

Examination hours: 9.00–13.00

This problem set consists of 6 pages.

Appendices: None.

Permitted aids: Any.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1 Elastic Net

Consider the following formulation of Elastic Net,

$$\hat{\beta}_{\text{EL-Net}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\}$$

where $y_1, \dots, y_N$ is a continuous response; x_{ij} is the i -th observation ($i = 1, \dots, N$) of the j -th variable ($j = 1, \dots, p$); $\beta_0, \beta_1, \dots, \beta_p$ are regression coefficients; and λ_1, λ_2 are hyperparameters.

a (8 pt.)

Using this formulation, briefly describe the method, including: (i) when it is equivalent to lasso and when to ridge regression; (ii) main advantage(s) and disadvantage(s).

b (8 pt.)

Describe the relationship between Elastic Net and Bridge regression,

$$\hat{\beta}_{\text{Bridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|^q \right\}$$

highlighting their similarities and differences in the case $1 \leq q \leq 2$. In addition, provide two specific cases in which the two methods are completely equivalent.

(Continued on page 2.)

Problem 2 Random Forests

Consider $Z = (y, X)$, where y is an N -dimensional vector of response and X an $N \times p$ input matrix, and B a large number of iterations. An early version of random forest (Ho, 1998) can be described by the following algorithm:

- For $b = 1, \dots, B$,
 1. generate a bootstrap sample $Z^* = (y^*, X^*)$;
 2. select $m < p$ variables;
 3. fit a statistical tree on (y^*, X_m^*) , where X_m^* only contains the m variable selected in the point (2);
- aggregate the results to get the final estimate.

a (8 pt.)

Highlight the difference with Leo Breiman's version of random forest that became "the" random forest algorithm (and was presented in the course). In addition, consider the simulation example described in the caption of Figure 1 (edited version of Figure 15.7 from Hastie et al, 2009). Do you expect that the algorithm above would have performed better, worse or equally good than Breiman's version? Why?

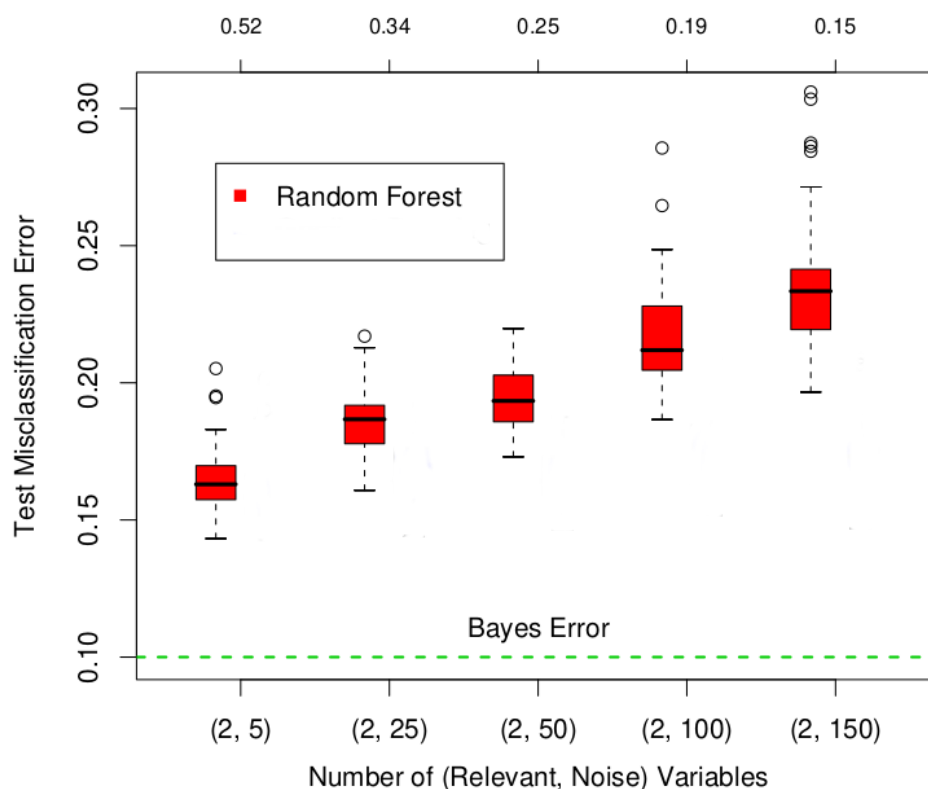


Figure 1: Results over 50 simulations (each time a training set of 300 observations has been generated, with a test set of 500 observations). Here the response depends on two variables, and an increasing number of noise variables are added (see the x-axis). The default $m = \sqrt{p}$ has been used, where p is the total number of variables and m the number of selected ones.

(Continued on page 3.)

b (8 pt.)

Consider the bias of a (Breiman's) random forest (regression) estimator. Hastie et al (2009) argue that it

“is the same as the bias of any of the individual sample trees”

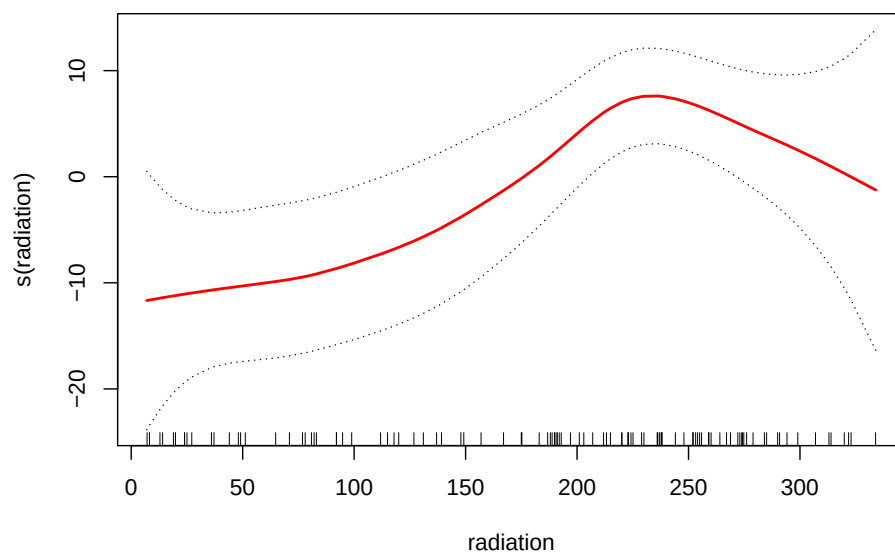
and it

“is also typically greater (in absolute terms) than the bias of an unpruned tree grown to Z , since the randomization and reduced sample space impose restrictions”.

Explain why a random forest bias is the same of any individual sample tree and why randomization and reduced sample space cause larger bias.

Problem 3 Splines

The effect of the solar radiation (in Ly) on the ozone concentration (in ppb) in the city of New York has been modelled with a spline in an additive model, obtaining the following result:



(data from Hastie et al, 2009)

a (8 pt.)

Describe the effect of the solar radiation on the ozone concentration, highlighting the differences with what one would have obtained in the case of a classical linear regression model.

(Continued on page 4.)

b (8 pt.)

Argue that the natural cubic spline $g(x)$ with knots at each of the x_i , with $a < x_1 < \dots < x_N < b$, $N \geq 2$, is the solution of

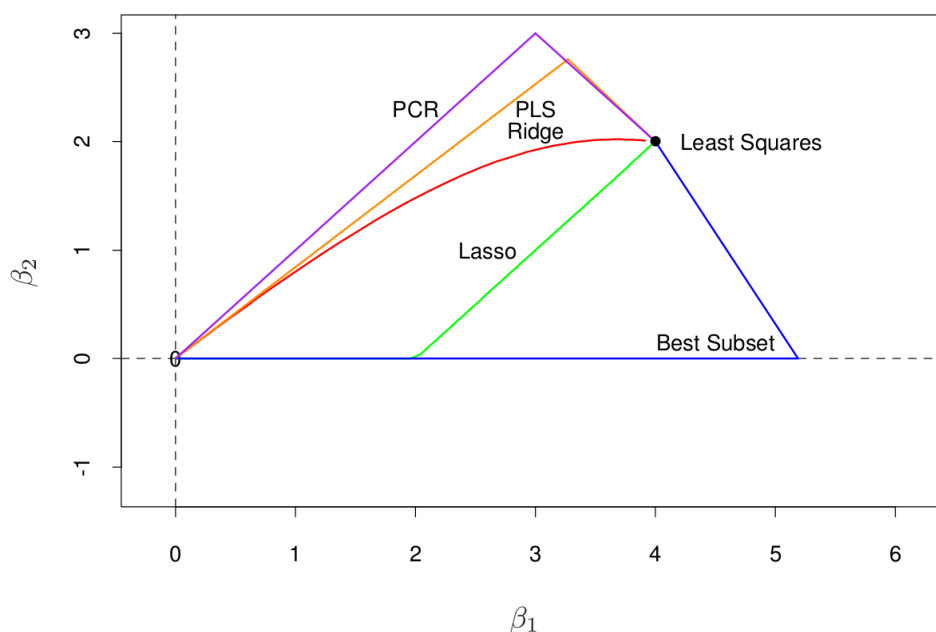
$$\operatorname{argmin}_f \left\{ \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda \int_a^b f''(t)^2 dt \right\}.$$

To do so, first show that $\int_a^b g''(t)^2 dt \leq \int_a^b \tilde{g}''(t)^2 dt$ for any other differentiable function $\tilde{g}(x)$ that interpolates the x_i . You may want to use the fact that $\int_a^b g''(t)(\tilde{g}(t)'' - g(t)'') dt = 0$.

Problem 4 Comparison among methods

Consider the following figure from Hastie et al. (2009),

$$\rho = 0.5$$



where “PCR” stands for “principal component regression”, “PLS” for “partial least square” and ρ is the correlation between X_1 and X_2 , whose regression coefficients are β_1 and β_2 , respectively. Here the point “Least Squares” denotes $\beta^{LS} = (\beta_1^{LS}, \beta_2^{LS})$, i.e., the point of the parametric space obtained by ordinary least squares.

a (12 pt.)

Describe the figure above, focusing, in particular on:

- the relationship between PCR and Ridge regression, explaining what makes their path similar and what makes it different;

(Continued on page 5.)

- the behaviour of the Lasso's path, both explaining why it changes direction in $(\beta_1, \beta_2) = (2, 0)$ and pointing out the values (if they can be identified) when $\lambda = 0$ and $\lambda = \infty$;
- why Best Subset reaches the point $(\beta_1, \beta_2) = (5.2, 0)$ even if the first component of $\beta^{LS} = (\beta_1^{LS}, \beta_2^{LS})$ is smaller than 5.2 (specifically, 4).

b (8 pt.)

Add the path of a component-wise boosting algorithm with linear base learners to the figure. Describe the reasons behind your choice.

Problem 5 Boosting

Consider a gradient boosting algorithm with a linear base learner, in a regression setting, $y = f(x) + \epsilon$, where $E[\epsilon] = 0$, and a boosting algorithm in the form

-
1. initialize the estimate, e.g., $\hat{f}^{[0]}(x) \equiv 0$;
 2. for $m = 1, \dots, m_{\text{stop}}$,
 - compute the negative gradient vector, $u(y, f(x)) = -\frac{\partial L(y, f(x))}{\partial f(x)}$;
 - fit the base learner to the negative gradient vector, $\hat{h}(u(y, \hat{f}^{[m-1]}), x)$;
 - update the estimate, $\hat{f}^{[m]}(x) = \hat{f}^{[m-1]} + \nu \hat{h}(u, x)$;
 3. return the final estimate $\hat{f}^{[m_{\text{stop}}]}(x)$.
-

where $L(y, f(x))$ is the loss function and $h(\cdot)$ a base-learner.

a (8 pt.)

Briefly describe the concept of bias-variance trade-off and relate the choice of the tuning parameters m_{stop} and ν to it.

b (8 pt.)

Consider the case in which we have a count as the response and we want to use a Poisson negative log-likelihood as a loss function. Given that the density of a Poisson random variable is

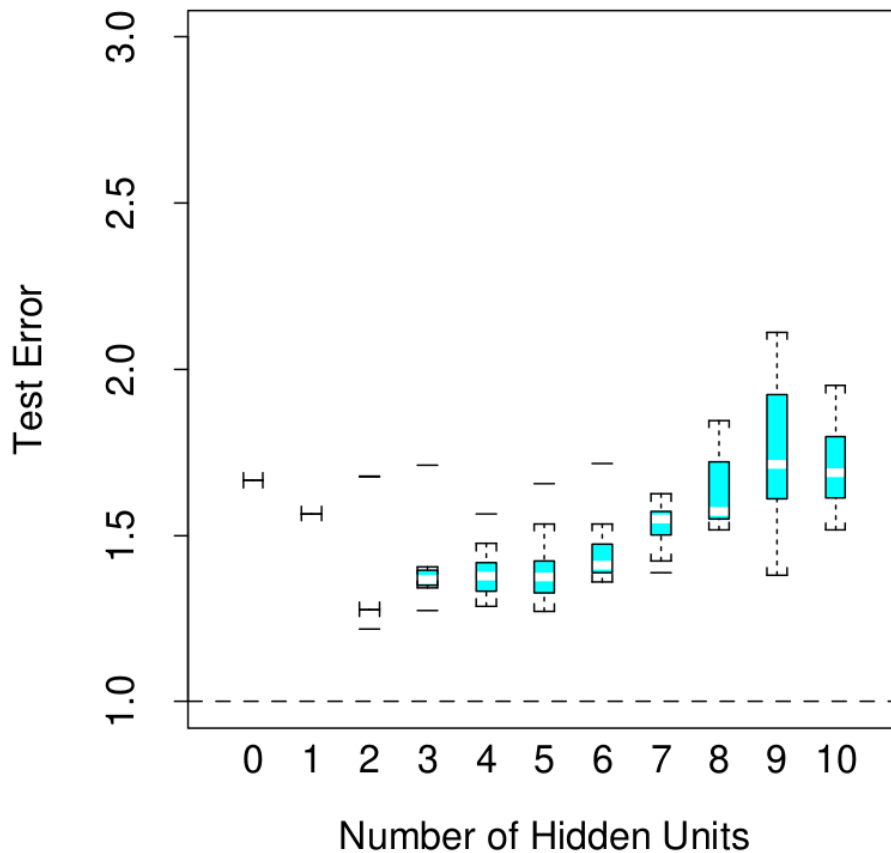
$$p(y) = \frac{e^{-\lambda(x)} \lambda(x)^y}{y!}$$

and that we may want to use a canonical link function, i.e., $\lambda(x) = e^{f(x)}$, derive the negative gradient $u(y, f(x))$.

(Continued on page 6.)

Problem 6 Neural Networks

Consider the following figure (Hastie et al, 2009, left panel of Figure 11.6),



where the results of a simulation study are shown. The test errors are computed for Neural Networks with different numbers of Hidden Units (x-axis), each for 10 different choices of starting weights. The Neural Networks have only one hidden layer and a fixed weight decay parameter of 5×10^{-4} .

a (8 pt.)

Despite the general suggestion of using a large number of Hidden Units, here the performance clearly worsen when increasing this parameter. Explain why this happens and what can be done to solve the issue.

b (8 pt.)

Relate the size of the Neural Network weights to the (model) complexity.

Bibliography

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20: 832-844.

THE END