UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

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

Day of examination: Monday, November 25th, 2020

Examination hours: 9.00 – 13.00

This problem set consists of 4 pages.

Appendices: None.

Permitted aids: None.

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

Problem 1 Penalized regression

Consider data simulated with the following setting:

- $\beta_i \sim N(0, 2)$, $i = 1, \dots, p$;
- $X \sim N_p(\underline{0}, \Sigma)$, where: (i) $N_p(\cdot, \cdot)$ denotes a p-dimensional multivariate Gaussian distribution; (ii) $\underline{0}$ is a p-dimensional vector of 0; (iii) Σ is a $p \times p$ matrix with diagonal elements equal to 1 and all other elements equal to 0.9;
- $y = X\beta + \epsilon$, with $\beta = (\beta_1, \dots, \beta_p)^T$ and $\epsilon \sim N(0, 1)$.

a (7 pt.)

If you were forced to choose between ridge regression and lasso, which one would you have used to predict y on a test set generated with the same setting? Why?

b (7 pt.)

Would your choice have been the same if you ignored the first information on β? Why?

c (6 pt.)

Do you think that elastic net could have been a better choice in the situation of point (b)? Why?

(Continued on page 2.)

Problem 2 Hjort-Glad estimator

Consider the Hjort-Glad estimator for density estimation,

$$\hat{f}_{HG}(x) = \frac{1}{N}\sum_{i=1}^{N} K_\lambda(x_i - x)\frac{f_0(x,\hat{\theta})}{f_0(x_i,\hat{\theta})}.$$

a (10 pt.)

Explain the logic behind the construction of such estimator, clarifying the role of the two terms $f_0(x,\hat{\theta})$ and $\sum_{i=1}^{N}\frac{K_\lambda(x_i-x)}{f_0(x_i,\hat{\theta})}$.

b (10 pt.)

Imagine you have to estimate the density of the variable "quantity of wine drunk by a person in a year": Which distribution would you use in $f_0(x,\hat{\theta})$? Explain what kind of problem (that one can face using a classical non-parametric density estimator) your choice can solve.

Problem 3 Cross-validation

Consider the following situation: we want to evaluate the predicting performance of a lasso procedure, but we do not have enough data to split them. Therefore, we decide to use cross-validation and we proceed as follows:

- we implement a K-fold cross-validation procedure to identify the best tuning parameter λ;
- we implement leave-one-out cross-validation to estimate the prediction error: the lasso with the estimated λ is trained on n-1 folds and evaluated on the remaining observation. The procedure is repeated n times, in which each observation plays, in turn, the role of the test observation, and the results averaged.

a (10 pt.)

Explain why the above procedure is wrong and provide a valid alternative.

b (10 pt.)

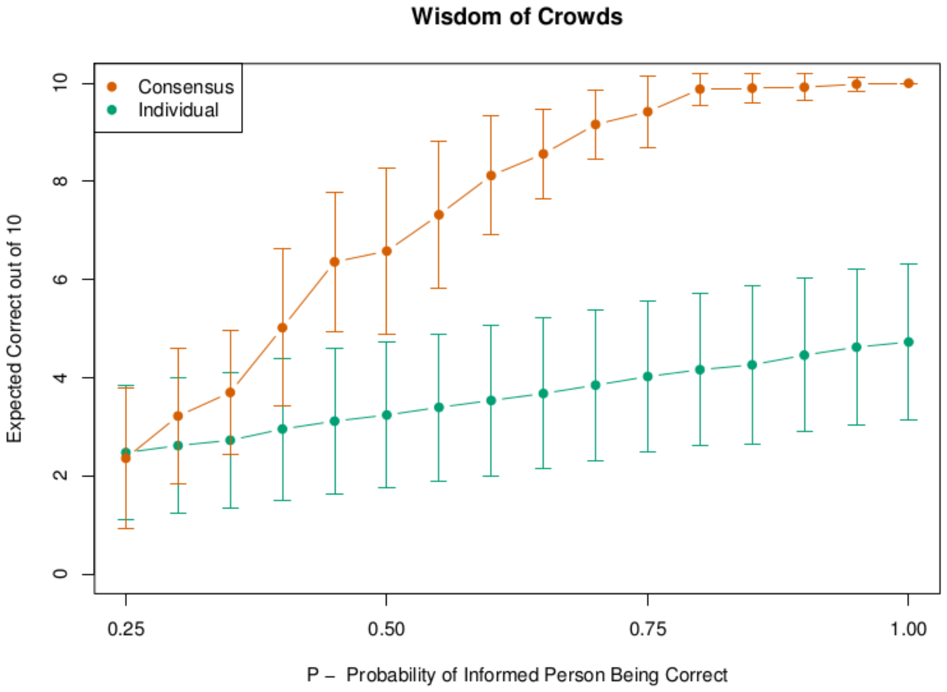
Describe the choice of the number of folds in a cross-validation procedure in terms of bias-variance trade-off.

(Continued on page 3.)

Problem 4 Widsom of the Crowd and Bagging

Consider the following example from “The Elements of Statistical Learning” of Hastie et al (2009, Figure 8.11):

“Simulated academy awards voting. 50 members vote in 10 categories, each with 4 nominations. For any category, only 15 voters have some knowledge, represented by their probability of selecting the “correct” candidate in that category (so $P = 0.25$ means they have no knowledge). For each category, the 15 experts are chosen at random from the 50. Results [reported in the figure below] *show the expected correct (based on 50 simulations) for the consensus, as well as for the individuals. The error bars indicate one standard deviation.”*



a (7 pt.)

How do we expect the two curves behave on the left of the interval considered in the plot (i.e., when $P \in [0; 0.25)$)? What does it mean for bagging?

b (6 pt.)

Compute the value of the green curve at p = 0.

c (7 pt.)

Derive the formula of the variance for B identically distributed random variables,

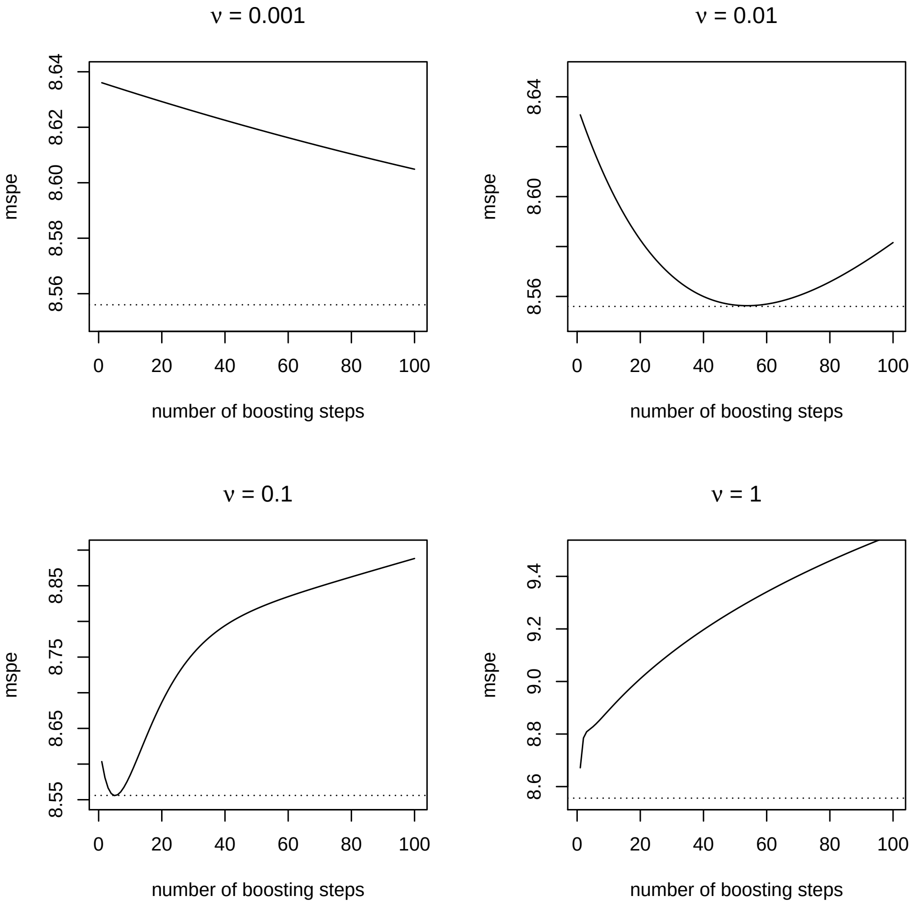
$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2,$$

and explain how bagging acts in order to improve the prediction performance of a single tree, and how this can be further improved.

(Continued on page 4.)

Problem 5 Boosting

Consider the following figure, obtained by using the R package `mboost`. Here the mean square prediction error has been computed as a function of the number of boosting steps, for four component-wise boosting models, each with a different value of the boosting step size ν.



a (10 pt.)

Explain the reasons for which we expected such a behaviour for the four curves, choose the best values for the tuning parameters in this case and justify your choice.

b (10 pt.)

Describe what differentiates the component-wise version of boosting from the standard one and describe two of its advantages.

THE END