UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK2100 — Machine Learning and Statistical Methods for Prediction and Classification
Day of examination:	May 31 - 2023
Examination hours:	15.00 – 19.00.

This problem set consists of 4 pages.

Appendices:	List of formulas for STK1100/STK1110 and STK2100
Permitted aids:	Approved calculator

Please make sure that your copy of the problem set is
complete before you attempt to answer anything.
All subquestions are counted equally!

Suggested solutions

Problem 1

(a) If we have a separate validation set we can get an unbiased estimate of the prediction error. If we had used the training data to also estimate the error, we would have got a too optimistic measure.

However, when dividing the data into two, we reduce the training set, making estimation of the model/parameters less reliable.

The training and validation set should come from the same population. If the database is structured in some systematic way, taking for instance the first instances as training could lead to systematic differences between these datasets. Random division avoids this problem.

(b) We have the classical bias-variance tradeoff in that for low q, the model underfit while for large q the model overfit, which in both cases lead to worse performance.

Due to that BIC is penalizing model complexity more, it is reasonable that BIC gives a smaller model.

(c) Since the variables are selected among all possible subsets, not through a sequential procedure, it can happen that the BIC model is not a subset of the AIC model. This in particular can happen when there is correlation between the variables.

The correlation is probably also the reason for why the P-value changes so much for some of the variables.

(Continued on page 2.)

(d) The GAM procedure is based on splines, which is a special case of basis functions. Given the basis functions, we have a linear model in the parameters involved, which makes the machinery for linear regression directly applicable. In particular degrees of freedom is calculated through trace($\boldsymbol{S}$) where $\boldsymbol{S}$ is the matrix defining $\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$.

From the plot it seems like **Room.Board** has a linear structure, which also makes it reasonable that the BIC is lower and the AIC is almost similar to the model with all variables non-linear. Even though the **Expend** variable seems to give a very non-linear structure, the BIC value is actually lower when only considering a linear term here. This is probably due to that the non-linear part is in a region with very few observations and the uncertainty is high here.

(e) Due to that neural networks usually are fitted through an optimization routine with random starting values, you can obtain different results when repeating the procedure. We clearly see the variability in results for the 20 repetitions.

There seems to be an improvement in increasing the number of hidden nodes to about 140, after which the results seems to be more stable.

It is tempting to choose the model that gave the best validation error. This is quite reasonable. However, due to that you now have made a model selection on the validation set, the error rate will be too optimistic. In order to really evaluate the performance, one should have a separate test set. Further, since more nodes probably will mean more local modes, the variability in the performance will also increase with increasing number of nodes and overfitting on the test set can occur. It might therefore be more robust to choose a somewhat lower number of nodes.

Problem 2

(a) For each applicant we have that they can be accepted (success) or not, a binary outcome. Assuming the outcomes are independent and all have the same probability for success (within each college), we then end up with a binomial distribution.

Both the independence and the same probability assumption can be questionable. In particular, there might be a fixed number of available places, making the probability of acceptance depend on the number of applicants.

(b) We have

$$y_i - \hat{y}_i = N_i^{10} + N_i^{11} - (N_i^{01} + N_i^{11}) = N_i^{10} - N_i^{01}.$$

What this essentially means is that although we can make individual misstakes either way, these are in some sense averaged out when only comparing y_i and $\hat{y}_i$. The error made on the aggregated level will therefore be smaller than when considered at an individual level.

(c) The main assumption now is independence. In principle this likelihood allows for different success probabilities at individual level.

(Continued on page 3.)

Due to that we do not have any individual based covariates, we have $\hat{p}_{ij} = h(\boldsymbol{x}_i)$ where $\boldsymbol{x}_i$ are the covariates at college level, but then we end up with $\hat{p}_{ij} = \hat{p}_i$. In that case,

$$\widehat{L} = \prod_i \prod_j \hat{p}_i^{z_{ij}}(1-\hat{p}_i)^{1-z_{ij}} = \prod_i \hat{p}_i^{\sum_j z_{ij}}(1-\hat{p}_i)^{n_i - \sum_j z_{ij}} = \prod_i \hat{p}_i^{y_i}(1-\hat{p}_i)^{n_i-y_i}.$$

(d) Assuming the College data is given as a matrix. We can then repeat each row n_i times. We then make a new response variable z which is equal to 1 for the first y_i repetitions and then equal to 0 for the last $n_i - y_i$ repetitions.

Since we only have $\hat{p}_{ij} = \hat{p}_i$, all z_{ij} within the same college will have the same decision, either 0 or 1. If all are 0, then also $\hat{y}_j = 0$. If all are 1, then $\hat{y}_j = n_j$.

When we use $\widehat{y}_i = \sum_j \hat{p}_{ij} = n_i \hat{p}_i$, we are taking the uncertainty about the decision into account. So we are estimating more directly the expectation of y_i, taking uncertainty in the individual z_{ij} into account.

(e) For the given set of covariates, we get that $\hat{p}_i = 0.7428$. In order to get a prediction on the number of accepted applicants, we have to multiply this number by the number of applicants, resulting in $\hat{y}_i = 1233.05$, which is very close to the actual value!

Tree classifiers are very easy to interprete. In this case the whole tree only depends on 3 covariates in a very simple way.

Problem 3

(a) We have that the log-likelihood in this case is given by

$$\ell = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^N \log(w_i\sigma^2) - \sum_{i=1}^N \frac{1}{w_i\sigma^2}(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2$$

and we see that maximizing ℓ is equivalent to minimizing $\sum_{i=1}^N \frac{1}{w_i}(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2$

Since $\boldsymbol{W}$ is a diagonal matrix, we get the vector/matrix formulation.

(b) We have

$$\begin{aligned} RSS =& \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ \frac{\partial}{\partial\boldsymbol{\beta}^T}RSS =& -2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

which when put to zero gives the result.

We have

$$\begin{aligned} E[\widehat{\beta}] =& E[[\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{Y}] \\ =& [\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}. \end{aligned}$$

(Continued on page 4.)

Further, we have

$$\begin{aligned}\text{Var}[\widehat{\beta}] =&\text{Var}[[\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{Y}]\\ =&[\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\sigma^2\boldsymbol{W}\boldsymbol{W}^{-1}\boldsymbol{X}[\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}]^{-1}\\ =&[\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}]^{-1}\sigma^2.\end{aligned}$$

(c) By dividing by $\sqrt{w_i}$ and defining $\tilde{y}_i = y_i/\sqrt{w_i}$ and $\tilde{x}_{ij} = x_{ij}/\sqrt{w_i}$, we obtain

$$\widetilde{Y}_i = \beta_0\frac{1}{w_i} + \sum_{j=1}^{p}\beta_j\tilde{x}_{ij} + \tilde{\varepsilon}_i, \quad \tilde{\varepsilon}_i \overset{iid}{\sim} N(0,\sigma^2)$$

which then becomes an ordinary linear regression model. The results above the follow from the general results from linear regression. In particular, if $\widetilde{\boldsymbol{X}}$ is the design matrix for the $\tilde{x}_{ij}$'s and $\widetilde{\boldsymbol{Y}}$ is the vector of $\tilde{y}_i$'s, we have $\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{Y}} = \boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{Y}$.