UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK2100 — Machine Learning and Statistical Methods for Prediction and Classification - Home exam
Day of examination:	June 15 -2021
Examination hours:	13.00 – 17.00.
This problem set consists of 4 pages.	
Appendices:	None
Permitted aids:	Anything available

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

Problem 1

(a) We have

$$\begin{aligned}
&E[(y_t - \hat{y}_t)^2|\boldsymbol{x}_t] \\
&\quad = E[(y_t - Np_t - (\hat{y}_t - Np_t))^2|\boldsymbol{x}_t] \\
&\quad = E[(y_t - Np_t)^2|\boldsymbol{x}_t] + E[(\hat{y}_t - Np_t)^2|\boldsymbol{x}_t] + 2E[(y_t - Np_t)(\hat{y}_t - Np_t)|\boldsymbol{x}_t] \\
&\quad = Np_t(1-p_t) + E[(Np_t - \hat{y}_t)^2|\boldsymbol{x}_t] + 2E[(y_t - Np_t)(\hat{y}_t - Np_t)|\boldsymbol{x}_t]
\end{aligned}$$

where the first term is the variance in the observations themselves, the irreducible part.

The second term can actually be written as

$$\begin{aligned}
&E[(\hat{y}_t - Np_t)^2|\boldsymbol{x}_t] \\
&\quad = E[(\hat{y}_t - E[\hat{y}_t|\boldsymbol{x}] + E[\hat{y}_t|\boldsymbol{x} - Np_t)^2|\boldsymbol{x}_t] \\
&\quad = E[(\hat{y}_t - E[\hat{y}_t)^2|\boldsymbol{x}_t] + E[(E[\hat{y}_t|\boldsymbol{x} - Np_t)^2|\boldsymbol{x}_t] + \\
&\qquad 2E[(\hat{y}_t - E[\hat{y}_t|\boldsymbol{x}])(E[\hat{y}_t|\boldsymbol{x}] - Np_t)|\boldsymbol{x}_t] \\
&\quad = \mathrm{Var}[\hat{y}_t|\boldsymbol{x}_t] + \mathrm{Bias}[\hat{y}_t|\boldsymbol{x}_t]^2 + 0.
\end{aligned}$$

which is the ordinary split into variance and bias.

The last term can be seen as some kind of covariance between y_t and $\hat{y}_t$. For linear regression this covariance is zero, but this do not hold in general. However, for large N, this term will be small.

The maximum value of the variance term is obtained for $p_t = 0.5$.

(b) The main differences between the two methods are that they penalize models somewhat differently, AIC use $2p$ while BIC use $\log(n)p$. BIC will then favour simpler models and thereby corresponds to Model 1.

(Continued on page 2.)

AIC is preferred for prediction while BIC is preferred for model learning.

(c) The variance of y_t is $Np_t(1-p_t)$ which is approximately $Np_t \approx y_t$ so that the variance increases with y_t and thereby the predictions also become more uncertain. To some extent one could also say that his is due to less data for large y_t's.

AIC/BIC:

Model	log-lik	AIC	BIC
Model 1	-718.99	1445.98	1491.71
Model 2	-715.13	1442.26	1510.85

Assuming prediction is the main perpose, preference would be given to model 2. However, an alternative argument is that there is not much difference between the two AIC values while BIC prefer Model 1 which also is simpler.

(d) We have

$$p_t = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{t,j})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{t,j})}$$

and for this to be small we must have $\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{t,j})$ small in which case the denominator ≈ 1, so that

$$p_t \approx \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{t,j}).$$

A simpler explantation is that $\text{logit} p_t = \log(p_t) - \log(1-p_t) \approx \log(p_t)$.

One would expect that y_t/p_t depends linearly on x_t. When we use the log-transform on p_t we then also should do the same on x_t. Also, there are few observations for large v_t's and the log-transform downweight the extreme covariates.

Due to that some values might be zero, adding one avoids getting $-\infty$.

(e) We now get the following table:

Model	log-lik	AIC	BIC
Model 1	-718.99	1445.98	1491.71
Model 2	-715.13	1442.26	1510.85
Model 3	-640.00	1288.00	1333.73
Model 4	-635.57	1285.14	1365.16

Clearly, the use of the log-transformed variables is better. Based on AIC we would prefer Model 4 while BIC would prefer Model 3.

The predictions now seems to be more stable, in that we do not get so extreme high predictions as we did previously.

(Continued on page 3.)

(f) The binomial model assumes independence, which might not be reasonable. In particular, we would expect similar behaviour in days that are close to each other

Further, it assumes that the probability for hospitalization is the same for all individuals, which again might not be reasonable.

When we possibly have dependence between observations, it is not that easy to apply cross-validation.

(g) We should expect differences between regions. Here the model is fit to one region, and then applied to another one, which can be problematic. The model fitted is not generalisable to other regions.

Problem 2

(a) We have that $AIC = -2 * l + 2 * p$ so that $p = l + 0.5AIC = -635.23 + 0.5 * 1295.57+ = 12.55$. For GAMs, we have that $\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ and $p = \text{tr}(\boldsymbol{S})$.

The dependence on v_{t-8} is increasing, which is reasonable given the discussion we have had earlier.

v_{t-9} is probably very correlated with v_{t-8} so that when v_{t-8} is included, we do not gain much in adding the other one.

v_{t-10} is less correlated with v_{t-8} and seems to be somewhat more influential and also with an increasing relationship. The drop in the end does not seem to be significant and be due to little data here.

(b) We get an improvement when evaluated on the same data as used for training, which is reasonable when we apply a more flexible model. However, the AIC value is a bit worse than models 3/4 from Problem 1, indicating a slightly worse fit.

We get worse fit on the Viken data which might indicate that the model is somewhat overfitting towards the Oslo data.

(c) The weak non-linear structure seen earlier can even be less significant when using a log-transform. This we partly also saw in Problem 1 in that the log-transform improved the fit.

The improvement on the Viken data indicates that we now have a more robust model when generalises better.

Problem 3

(a) The first choice corresponds to first making some linear transformations of the input data, similar to principal components with dimension

(Continued on page 4.)

reduction. The α parameters can then be prespecifed through principal components of the covariates.

The second choice correspond to neural network with one hidden layer.

(b) When q is large, we get very many parameters to estimate. In order to avoid overfitting we should use some kind of penalty. The specific type of penalty we use here is L_2 or Ridge-type which has the effect of shrinking the parameters towards zero, and thereby reducing the variability in the parameter estimates.

The batch normalization will stabilize the input variables to the last link to y_i, which we have seen are useful for Ridge/Lasso regression and also makes it more reasonable to have the same penalty on all parameters involved at the same layer.

(c) We have that

$$\begin{aligned}\frac{\partial}{\partial\beta_0}L_{\boldsymbol{\lambda}} =& -2\sum_{i=1}^{n}(y_i-\beta_0-\sum_{k=1}^{q}\beta_k\tilde{z}_{ik})\\ =& -2\sum_{i=1}^{n}(y_i-\beta_0)\end{aligned}$$

showing that

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}y_i$$

Further,

$$\begin{aligned}\frac{\partial}{\partial\beta_l}L_{\boldsymbol{\lambda}} =& -2\sum_{i=1}^{n}(y_i-\bar{y}-\sum_{k=1}^{q}\beta_k\tilde{z}_{ik})\tilde{z}_{il}\\ =& -2\sum_{i=1}^{n}(y_i-\sum_{k=1}^{q}\beta_k\tilde{z}_{ik})\tilde{z}_{il}+2\lambda_2\beta_l\end{aligned}$$

which gives the equations

$$\sum_{k=1}^{q}\beta_k\tilde{z}_{ik}\tilde{z}_{il}+\lambda_2\beta_l = \sum_{k=1}^{q}y_i\tilde{z}_{il} \qquad l=1,...,q$$

One can alternatively use vector/matrix formulations in order to obtain the analytic formulae

$$\hat{\boldsymbol{\beta}} = [\widetilde{\boldsymbol{Z}}^T\widetilde{\boldsymbol{Z}}+\lambda_2\boldsymbol{I}]^{-1}\widetilde{\boldsymbol{Z}}^T\boldsymbol{y}.$$

The penalty term λ_2 has the effect of shrinking the estimate towards zero. Yes, we would expect a similar behaviour on the α's but the equations will be somewhat more difficult.