

EXAM 2022

Problem 3

c. The number of free parameters is still 3 because there are no constraints on the parameter estimates, we are simply replacing the covariates by derived covariates. The log-likelihood is lower than for Model 2, so the AIC and BIC must be higher. For values see the Solution.

$$d. \text{ tpr: True positive rate} = \hat{P}(\hat{y}_i = 1 | y_i = 1) = \frac{\sum_{i=1}^N \mathbb{I}(\hat{y}_i = 1, y_i = 1)}{\sum_{i=1}^N \mathbb{I}(y_i = 1)}$$

= Sensitivity

$$\text{fpr: False positive rate} = \hat{P}(\hat{y}_i = 1 | y_i = 0) = \frac{\sum_{i=1}^N \mathbb{I}(\hat{y}_i = 1, y_i = 0)}{\sum_{i=1}^N \mathbb{I}(y_i = 0)}$$

= $1 - \underbrace{\text{True negative rate}}_{\text{Specificity}}$

To get $\hat{y}_i$, we need a threshold $\tau \in (0, 1)$:

$$\hat{y}_i = \begin{cases} 1, & \hat{p}_i > \tau \\ 0, & \hat{p}_i \leq \tau \end{cases}$$

The ROC curve is tpr plotted against fpr for different values of τ .

The blue curve has the largest AUC $\rightarrow$

Model 1, the red curve has the second largest

AUC $\rightarrow$ Model 2 and the green curve must

the correspond to Model 3. The order of

the models according to the AUC is still

Model 1 > Model 2 > Model 3.

e. Model 6 : very local decision boundary $\rightarrow$ KNN

Model 5 : linear decision boundary $\rightarrow$ LDA

Model 7 : decision boundary consists of rectangles $\rightarrow$ tree

Model 4 : non-linear decision boundary $\rightarrow$ ANN

According to the AUC, from best to worst:

Model 1 $>$ Model 2 $>$ Model 4 $>$ Model 5 $>$ Model 3 $>$ Model 6
 $>$ Model 7

- Higher AUC with GAM than with linear effects in Models 1 and 2
- The curves in the plots look non-linear
- The tests for the significance of the non-linear parts of the effects in the second table indicate significant non-linearity in all 4 covariates

All these three point indicate that there are important non-linear effects.

EXAM 2023

Problem 1

a - Strengths : - More realistic estimate of prediction error (not too optimistic)

Weaknesses : - More uncertain model estimates because the models are trained on fewer data points.

The training and the test set should come from the same distribution - Therefore they are selected randomly.

b. Both the log-likelihood and the penalties of the AIC and BIC increase with the number of covariates in the model. At some point the log-likelihood increase is slower and then the curves increase. This is related to the bias-variance tradeoff, where the decrease in the bias for higher number of covariates at some point is lower than the increase of the variance. The BIC penalises the number of parameters more, and therefore chooses a lower q .