

# UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK2100 — Machine Learning and Statistical  
Methods for Prediction and Classification  
SKETCH OF SOLUTION

Day of examination: Thursday June 5th, 2025

This problem set consists of 7 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

**a**

- The P-values from the Wald tests on the individual covariates indicate that the most important covariates in the linear regression model are **season**, **hr**, **daynight**, **weather**, **temp** and **hum**, whereas **holiday** and **windspeed** seem to be less important, and **workingday** does not have a significant effect.
- The estimated coefficient for **temp** is 4.3, which means that the expected number of rented bikes increases by a bit more than 4 when the temperature increases by 1°C and all other covariates are the same.
- What is less straightforward is that **daynight** depends on the value of **hr**, and cannot necessarily be kept fixed when **hr** is varied, and vice versa.

**b**

- Let  $l(\hat{\theta})$  be the log-likelihood of the model that is considered, with  $\hat{\theta}$  the maximum likelihood estimate of the model parameters  $\theta$ . Then

$$AIC = -2l(\hat{\theta}) + 2|\hat{\theta}| \quad \text{and} \quad BIC = -2l(\hat{\theta}) + \log(N)|\hat{\theta}|,$$

where  $|\hat{\theta}|$  is the number of free parameters and  $N$  is the size of training set. The BIC is consistently higher than the AIC because it puts a higher penalty on the number of free parameters.

- The covariates that are left out correspond to the ones that were either not significant, or were the least significant in the full model (see **a**).
- Lower values of either AIC or BIC indicates a better model. These are however not lowest for the same model. If we choose to use the AIC as criterion, which is natural if the main aim is to do prediction, we would choose Model 2. If we on the other hand are more interested in interpretation and choosing the correct model, we may choose according to the BIC, and hence select Model 3.

(Continued on page 2.)

c

- The procedure for doing stepwise forward selection with  $K$ -fold cross-validation could be as follows.

- Divide the data randomly into  $K$  folds.
- For each size  $q$  of the subset of covariates from 1 to  $p$ , where  $p$  is the maximum number of covariates in the model, for all folds from 1 to  $K$ , one must fit all models including one more of the covariates that have not yet been included in the model to all the data except the observations in the current fold.
- Each fitted model is then used to predict the observations that were not used in the fitting, and the corresponding loss (typically the squared error loss) is then computed, and averaged over all the folds.
- For each subset size, one adds the covariate that gives the lowest average error over the folds.

- The advantage of using cross-validation instead of AIC and BIC is that it relies less on the model assumptions. On the other hand, cross-validation is in this setting very computationally demanding due the large number of models that must be fitted.

d

- The expected number of rented bikes increases with the hour of day from the morning, but especially in the afternoon, up to a peak around 20:00, and then decreases during the night. The wind speed does not seem to affect the number of rented bikes until it becomes quite high ( $\sim 30$  km/h), and then has an increasingly negative effect.

- According to the tests of the contribution of the non-linear parts of the effects modelled with splines, these are all significant compared to a simple linear term (this is seen in the table "Anova for Nonparametric Effects"). Further, both the AIC and BIC are significantly lower in the additive model, compared to the linear model from **a** with the same covariates, despite a higher number of degrees of freedom. This all suggests that there are non-linear effects of these covariates.

e

- The most important covariates in Model 5 are **season**, **hr** and **temp**, which are shared with Models 1-4, but interestingly also **workingday**, which did not have a significant effect in any of the other models. Further, the other covariates describing the weather conditions were important in Models 1-4, but not in Model 5, and the same goes for **daynight**, which is replaced by many splits in **hr**.

- We know that the prediction from Model 1 for Case 2 is 193, and the only difference in Case 4 is that **hr**=11, where the estimated effect of **hr** is 5.6. Hence, the prediction for Case 4 must be  $193 + 3 \cdot 5.6 \approx 210$ . For Models 2 and 3, there is no difference between Cases 1 and 2, and between 3 and 4, since **workingday** is not part of these models, which helps us filling out the

(Continued on page 3.)

missing predictions for the missing cases without any need for computations (see the table below). Finally for model 5 we have:

- Case 2: same path as for Case 1 including "hr < 16.5" go left, then "workingday: 0" go right, at "hr < 9.5" go left, giving the prediction 342.
- Case 3: same path as for Case 1 including workingday: 0" go left, then at "hr < 10.5" go right, giving the prediction 333.
- Case 4: same path as for Case 1 including "hr < 16.5" go left, then "workingday: 0" go right, at "hr < 9.5" go right, giving the prediction 163.

The table with all the predictions is then

| Setting | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---------|---------|---------|---------|---------|---------|
| Case 1  | 193     | 193     | 192     | 172     | 107     |
| Case 2  | 193     | 193     | 192     | 172     | 342     |
| Case 3  | 210     | 210     | 209     | 183     | 333     |
| Case 4  | 210     | 210     | 209     | 183     | 163     |

- We see that when it is a working day, more bikes are rented in the early morning compared to the middle of the day, whereas when it is not a working day, more bikes tend to be rented in the middle of the day compared to early morning. This means that hr has the opposite effect in that time span depending on whether it is a working day or not, indicating an interaction effect between these two covariates.

**f**

- The training error typically underestimates the true prediction error, from prediction on new values of the covariate vector, and will keep decreasing with increasing model complexity. Hence, if we used the training error to evaluate the fitted models, we would always favour the most complex models, and would not detect potential overfit.

- According to the mean squared errors on the test set, the tree model, Model 5, is by far the best, probably because it is able to capture both non-linearities and interactions. The second best is the additive model, Model 4, again due to the ability to model non-linear effects. The remaining three models have very similar performance, and it is therefore natural to choose one of the simplest models, i.e. Model 2 or 3.

## Problem 2

**a**

In the logistic regression model, we have

$$\Pr(Y = 1|\mathbf{x}) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ij}}},$$

(Continued on page 4.)

so the decision border is given by

$$\begin{aligned} \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ij}}} &= \frac{1}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ij}}} \\ \longrightarrow e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ij}} &= 1 \\ \longrightarrow \beta_0 + \sum_{j=1}^p \beta_j x_{ij} &= 0 \end{aligned}$$

- There are no constraints on the parameters, and so the number of degrees of freedom is the same as the number of parameters, i.e.  $p + 1$ .

**b**

In the LDA model, we use Bayes' law to find  $\Pr(Y = k|\mathbf{x})$ ,  $k = 0, 1$ , letting  $f_k(\mathbf{x})$  denote the density of  $\mathbf{X}$ , given that  $Y = k$ :

$$\Pr(Y = k|\mathbf{x}) = \frac{\Pr(Y = k)f_k(\mathbf{x})}{\Pr(Y = k)f_k(\mathbf{x}) + \Pr(Y = 1 - k)f_{1-k}(\mathbf{x})}$$

Hence, the decision border is given by

$$\begin{aligned} \frac{\pi_1 f_1(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + (1 - \pi_1) f_0(\mathbf{x})} &= \frac{(1 - \pi_1) f_0(\mathbf{x})}{(1 - \pi_1) f_0(\mathbf{x}) + \pi_1 f_1(\mathbf{x})} \\ \pi_1 f_1(\mathbf{x}) &= (1 - \pi_1) f_0(\mathbf{x}) \\ \log(\pi_1) + \log(f_1(\mathbf{x})) &= \log((1 - \pi_1)) + \log(f_0(\mathbf{x})), \end{aligned}$$

with  $f_k(\mathbf{x}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp(-1/2(\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_k))$ , so that

$$\begin{aligned} \log(f_k(\mathbf{x})) &= -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \\ &= -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x} + \boldsymbol{\mu}_k^T \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k \end{aligned}$$

Hence, we get

$$\begin{aligned} \log(\pi_1) - \frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x} + \boldsymbol{\mu}_1^T \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 &= \\ \log((1 - \pi_1)) - \frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x} + \boldsymbol{\mu}_0^T \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_0 &= \\ \longrightarrow \log(\pi_1) - \log(1 - \pi_1) - \frac{1}{2} \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1^T \Sigma^{-1} \mathbf{x} - \boldsymbol{\mu}_0^T \Sigma^{-1} \mathbf{x} &= 0 \\ \longrightarrow \log\left(\frac{\pi_1}{1 - \pi_1}\right) - \frac{1}{2} (\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \Sigma^{-1} \mathbf{x} &= 0. \end{aligned}$$

- Again, the number of degrees of freedom is the number of parameters in the model, which is 1 for the marginal probability of  $Y = 1$ ,  $p$  for each of the two mean vectors and  $p(p + 1)/2$  for the common covariance matrix, i.e.  $1 + 2p + p(p + 1)/2 = 1/2p^2 + 5/2p + 1$ .

(Continued on page 5.)

c

To find the decision border of the special case, we simply need to plug in the diagonal covariance matrices into the general expression for a QDA model.

We have:

$$\log \frac{|\Sigma_1|}{|\Sigma_0|} = \log \frac{\sigma_{11}^2 \cdots \sigma_{1p}^2}{\sigma_{01}^2 \cdots \sigma_{0p}^2} = 2 \sum_{j=1}^p \log \frac{\sigma_{1j}}{\sigma_{0j}}$$

Further,

$$\begin{aligned} \mu_k^T \Sigma_k^{-1} \mu_k &= (\mu_{k1} \quad \dots \quad \mu_{kp}) \begin{pmatrix} 1/\sigma_{k1}^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_{k2}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1/\sigma_{kp}^2 \end{pmatrix} \begin{pmatrix} \mu_{k1} \\ \vdots \\ \mu_{kp} \end{pmatrix} \\ &= (\mu_{k1} \quad \dots \quad \mu_{kp}) \begin{pmatrix} \mu_{k1}/\sigma_{k1}^2 \\ \vdots \\ \mu_{kp}/\sigma_{kp}^2 \end{pmatrix} \\ &= \sum_{j=1}^p \frac{\mu_{kj}^2}{\sigma_{kj}^2}. \end{aligned}$$

Moreover,

$$\begin{aligned} (\mu_1 \Sigma_1^{-1} - \mu_0 \Sigma_0^{-1})^T \mathbf{x} &= ((\mu_{11}/\sigma_{11}^2 \quad \dots \quad \mu_{1p}/\sigma_{1p}^2) - (\mu_{01}/\sigma_{01}^2 \quad \dots \quad \mu_{0p}/\sigma_{0p}^2)) \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \\ &= \sum_{j=1}^p \left( \frac{\mu_{1j}}{\sigma_{1j}^2} - \frac{\mu_{0j}}{\sigma_{0j}^2} \right) x_j. \end{aligned}$$

Finally,

$$\begin{aligned} \mathbf{x}^T (\Sigma_1^{-1} - \Sigma_0^{-1}) \mathbf{x} &= (x_1 \quad \dots \quad x_p) \begin{pmatrix} 1/\sigma_{11}^2 - 1/\sigma_{01}^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_{12}^2 - 1/\sigma_{02}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1/\sigma_{1p}^2 - 1/\sigma_{0p}^2 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \\ &= (x_1 \quad \dots \quad x_p) \begin{pmatrix} (1/\sigma_{11}^2 - 1/\sigma_{01}^2)x_1 \\ \vdots \\ (1/\sigma_{1p}^2 - 1/\sigma_{0p}^2)x_p \end{pmatrix} \\ &= \sum_{j=1}^p \left( \frac{1}{\sigma_{1j}^2} - \frac{1}{\sigma_{0j}^2} \right) x_j^2. \end{aligned}$$

Putting it all together, we get the decision boundary

$$\log \left( \frac{\pi_1}{1 - \pi_1} \right) - \sum_{j=1}^p \log \frac{\sigma_{1j}}{\sigma_{0j}} - \frac{1}{2} \sum_{j=1}^p \left( \frac{\mu_{1j}^2}{\sigma_{1j}^2} - \frac{\mu_{0j}^2}{\sigma_{0j}^2} \right) + \sum_{j=1}^p \left( \frac{\mu_{1j}}{\sigma_{1j}^2} - \frac{\mu_{0j}}{\sigma_{0j}^2} \right) x_j - \frac{1}{2} \sum_{j=1}^p \left( \frac{1}{\sigma_{1j}^2} - \frac{1}{\sigma_{0j}^2} \right) x_j^2 = 0,$$

which is indeed quadratic in  $\mathbf{x}$ .

- The number of degrees of freedom, which is the number of model

(Continued on page 6.)

parameters, is 1 for the marginal probability of  $Y = 1$ ,  $p$  for each of the two mean vectors and  $p$  for each of the two sets of variances, i.e.  $1 + 2p + 2p = 4p + 1$  (which is smaller than the degrees of freedom of the LDA when  $p > 3$ ).

- For the standard QDA, the number of degrees of freedom is  $p(p + 1)/2$  more than for the LDA due to the extra covariance matrix, i.e.  $1 + 2p + 2p(p + 1)/2 = p^2 + 3p + 1$ . Hence, the simplest model is the logistic regression model from **a**, for  $p > 3$ , the QDA with conditional independence is second, the LDA third and the standard QDA fourth.

### Problem 3

**a**

- For a given threshold  $\tau \in (0, 1)$ , and the predicted probabilities  $\hat{p}_i$ ,  $i = 1, \dots, N_{te}$ , for the test set, assume that we classify as

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

Then, the true positive rate (tpr) is the fraction of the true positives  $y_i = 1$  that are classified as positive, i.e.  $\hat{y}_i = 1$  and the false positive rate (fpr) is the fraction of the true negatives  $y_i = 0$  that are classified as positive, i.e.  $\hat{y}_i = 1$ . Hence,

$$\begin{aligned} \text{tpr} &= \frac{\sum_{i=1}^{N_{te}} I(\hat{y}_i = 1, y_i = 1)}{\sum_{i=1}^{N_{te}} I(y_i = 1)} \\ \text{fpr} &= \frac{\sum_{i=1}^{N_{te}} I(\hat{y}_i = 1, y_i = 0)}{\sum_{i=1}^{N_{te}} I(y_i = 0)}. \end{aligned}$$

For the given  $\tau$ , we get the pair (tpr, fpr). The receiver operating characteristic (ROC) curve is obtained by computing this pair for the full range of values of  $\tau$ , and then plotting the tprs against the fprs.

- A perfect classifier has a ROC curve lying along the upper left corner of the plot. On the other hand, random guessing gives a ROC curve along the diagonal. Hence the closer the ROC curve of a classifier is to the top left corner, the better it is, and the closer to the diagonal, the worse it is. The area under the curve could also have been mentioned here, but it is the topic of Problem **b**.

- Since the ROC curves are crossing each other, it is difficult to rank the models merely based on the curves. However, the black and the red curves are quite close, indicating that the performance of the logistic regression model and the LDA is quite similar. The blue curve seems to be overall closer to the top left corner, indicating that the QDA with conditional independence might perform better. The ROC curve of the standard QDA is difficult to evaluate.

**b**

- The area under the curve (AUC), is the area defined by the ROC curve, and the lower right corner of the plot. A perfect classifier has an AUC of 1

(Continued on page 7.)

and random guessing gives an AUC of 0.5.

- As suggested in **a**, the AUCs of the logistic regression model and the LDA are quite close. Considering the fact that the logistic regression model has much fewer parameters and also has a lower misclassification error on the test set, we rank the logistic regression model above the LDA. The standard QDA has both a quite higher AUC and lower misclassification error than the other two, and is therefore ranked above these. The by far best performing model is however the QDA with conditional independence, with substantially higher AUC and lower misclassification error, and is therefore our preferred model.

- Since both the standard QDA and the QDA with conditional independence give much better results than the logistic regression model and the LDA, a linear decision boundary does not seem to be a sufficiently good approximation in this case. The number of degrees of freedom of the different models is  $32 + 1 = 33$  for the logistic regression model,  $1/2 \cdot 32^2 + 5/2 \cdot 32 + 1 = 593$  for the LDA,  $32^2 + 3 \cdot 32 + 1 = 1121$  and  $4 \cdot 32 + 1 = 129$  for the QDA with conditional independence. Hence, the combination of a quadratic decision boundary and not too much complexity is probably the reason why the QDA with conditional independence has the best performance.