

EXAM 2025

Problem 1

a. - Variables used for splits in the tree are considered important:

hr, temp, season and workingday

Some covariates are considered important in Models 1 to 4, but not in Model 5:

humidity, daynight

One covariate is considered important in Model 5, but not in Models 1 to 4:

workingday

- Use the values in the table for Models 1 to 4 and compute the difference. For Model 5 explain how you go through the tree to obtain the predictions.

- We see that the effect of hr is the opposite when workingday = 0 and when workingday = 1: more bikes are rented in the morning on workingdays and bikes are rented in the morning when it is not a working day.

- f. - The training error tends to be too optimistic, and underestimate the true prediction error. Using the training error as a performance measure increases the risk of overfitting.
- Based on test error Model 5 is the best, followed by Model 4. The remaining models have very similar performance, which could be an argument for choosing the simplest one and get the order Model 3, Model 2, Model 1. Alternatively one always chooses the one with lowest test error, and gets the order Model 2, Model 1, Model 3.

Problem 2

a. In the logistic regression model we have:

$$P(Y=1|\underline{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \text{ and } P(Y=0|\underline{x}) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$

The decision border is given by

$$\frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$

$$\longrightarrow e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p} = 1$$

$$\longrightarrow \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \log(1) = 0$$

— No constraints on the parameters, so the number of free parameters is equal to the number of parameters, i.e. $p+1$.

b. In LDA, we have:

$$P(Y=1|x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + (1-\pi_1) f_0(x)}, \text{ where } f_i(x) \text{ is the density of } X \text{ in class } Y=i.$$

$$P(Y=0|x) = \frac{(1-\pi_1) f_0(x)}{\pi_1 f_1(x) + (1-\pi_1) f_0(x)}$$

The decision border is given by:

$$\frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + (1-\pi_1) f_0(x)} = \frac{(1-\pi_1) f_0(x)}{\pi_1 f_1(x) + (1-\pi_1) f_0(x)}$$

$$\rightarrow \pi_1 f_1(x) = (1-\pi_1) f_0(x)$$

$$\rightarrow \log(\pi_1) + \log f_1(x) = \log(1-\pi_1) + \log f_0(x)$$

$$\rightarrow \log(\pi_1) - \frac{1}{2} \log(2\pi) - \log|\Sigma| - \frac{1}{2} (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) = \log(1-\pi_1) - \frac{1}{2} \log(2\pi) - \log|\Sigma| - \frac{1}{2} (x - \mu_0)^T \Sigma^{-1} (x - \mu_0)$$

$$\rightarrow \log(\pi_1) - \frac{1}{2} x^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1} x - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 = \log(1-\pi_1) - \frac{1}{2} x^T \Sigma^{-1} x + \mu_0^T \Sigma^{-1} x - \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0$$

$$\rightarrow \log(\pi_1) - \log(1-\pi_1) - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 + \mu_1^T \Sigma^{-1} x - \mu_0^T \Sigma^{-1} x = 0$$

$$\rightarrow \log \frac{\pi_1}{1-\pi_1} - \frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0) + (\mu_1 - \mu_0)^T \Sigma^{-1} x = 0$$

- No constraints on the parameters $\rightarrow$ the number of free parameters is the same as the number of parameters in the model: $1 + 2p + \frac{p(p+1)}{2}$

C. - Plug in the diagonal Σ_k into the decision border for general QDA (see the solution). We get the decision border:

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

- No constraints $\longrightarrow$ the number of free parameters is the number of model parameters:

$$\text{General QDA: } 1 + 2p + \cancel{2} \cdot \frac{p(p+1)}{\cancel{2}}$$

$$\text{Special QDA: } 1 + 2p + 2p$$

- In terms of number of free parameters, when $p > 3$:

$$\text{Logistic reg} < \text{Special QDA} < \text{LDA} < \text{QDA}.$$

Problem 3

a. - For a given threshold $\tau \in (0, 1)$, we classify as

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

Based on this, we compute

$$tpr = \frac{\sum \mathbb{I}(\hat{y}_i = 1, y_i = 1)}{\sum \mathbb{I}(y_i = 1)}$$

$$fpr = \frac{\sum \mathbb{I}(\hat{y}_i = 1, y_i = 0)}{\sum \mathbb{I}(y_i = 0)}$$

This gives the pair (fpr, tpr) . The ROC curve is a plot of pairs (fpr, tpr) obtained by varying τ along $(0, 1)$.

- The better the classifier, the closer the ROC curve is to the upper left corner.
- Due to crossing curves, it is challenging to rank the models.

b. - The AUC is the area between the ROC curve and the x-axis. A perfect classifier gives an AUC of 1, random guessing gives 0.5. A good classifier should give an AUC as close to 1 as possible.

- Both the misclassification error and the AUC agree on the ranking from best to worst: special QDA, general QDA, Logistic reg, LDA.
- The QDA's with quadratic decision border do better than the models with linear decision border $\rightarrow$ linear decision border is too simplistic.

The Special QDA has a much lower complexity than the general QDA, which is probably why it does better in this case.