Exam questions:

Chapter 4:

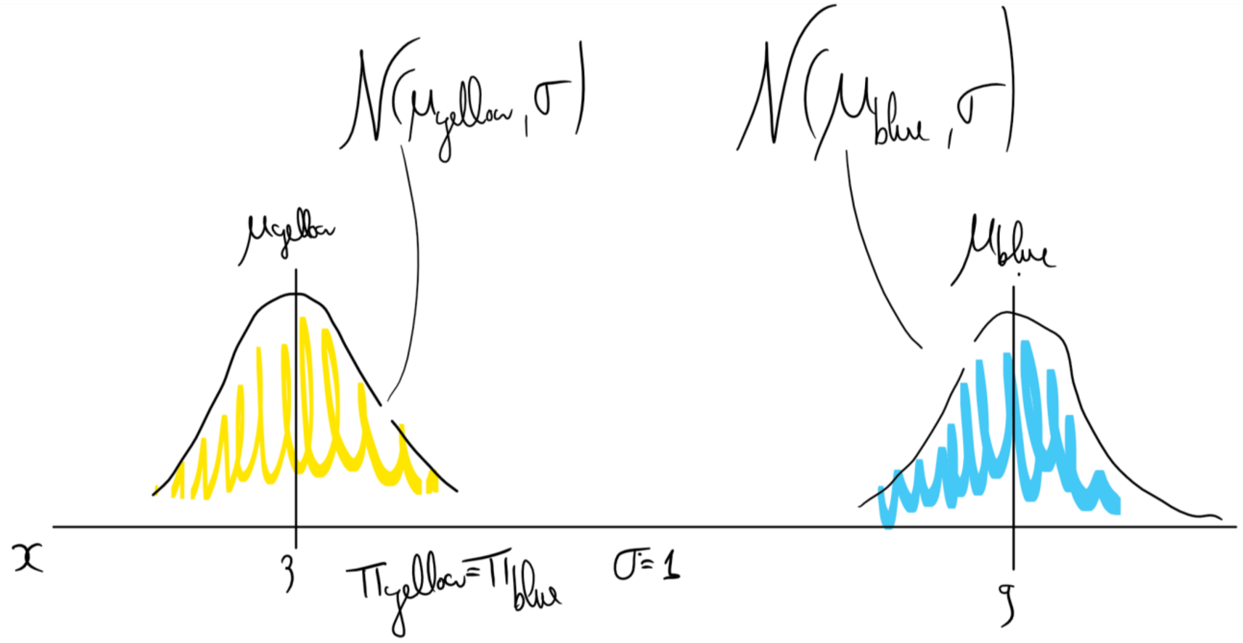
- Why shouldn’t we use linear regression to model binary outcome variables?
- Proof that:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \text{ follows from } p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

- Can you sketch the function of y = f(x) = x/(1+x) when the range of x is in [0,inf]]?
- What are the components in the likelihood function associated to the logistic regression?

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

- What is the difference in approach between KNN, logistic regression and linear discriminant analysis when estimating the conditional probability that Pr(Y=k|X)?
- How would you estimate the prior probability?
- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem. Why is that?
- Consider following distributions below. Can you compute the posterior probability of the LDA classifier for $P_{yellow}(x=6)$? Write down the equations and compute probability value.



- Illustrate that applying the principle of the Bayes classifier to the linear discriminate analysis approach:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}.$$

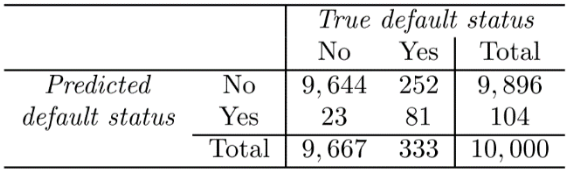
  Is equivalent as assigning the class label to the observation for which the expression below is largest:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

- Given the discriminant function below. Derive the decision boundary for x in the case of 2 classes and π = 0.5.

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

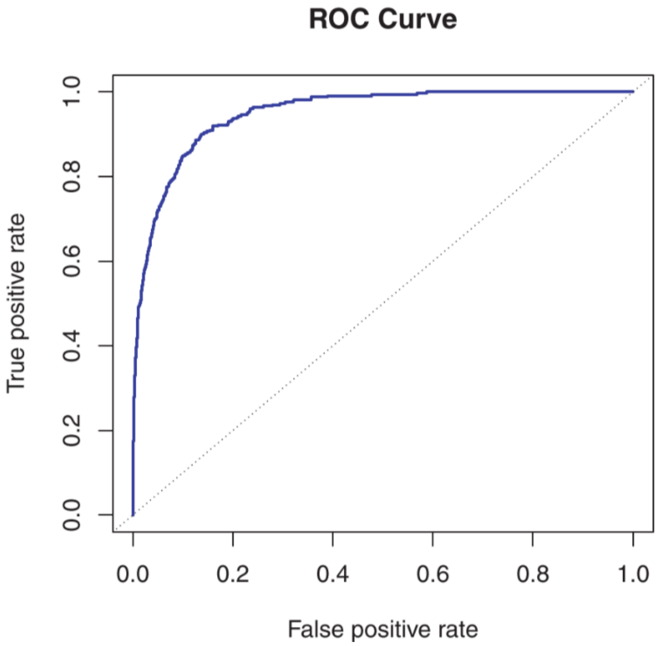
- The book mentions that logistic regression is unstable with well separable classes. Should I prefer an LDA approach when my predictor variables are categorical? Explain!
- What is the difference between LDA and QDA?
- In the multivariate case with p>1 – what is assumed about variance-covariance structure?
- What is a null classifier?
- Consider the confusion matrix below. This matrix gives the predictions of an algorithm that aims to predict whether a client will default on his loan payments. Explain its purpose. What is the overall error rate of this classifier? Compute the sensitivity and specificity and describe the term in your own words. Why are these metrics important and which metric is preferred in this case?



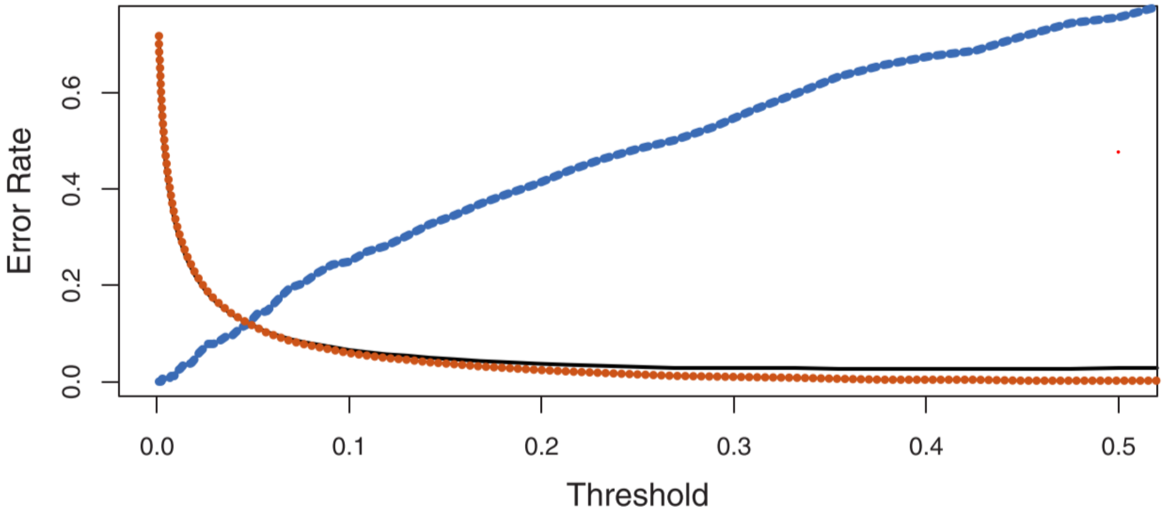
		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

- The Bayes classifier assign the class label to the observation with the largest posterior probability. For two classes this boils down to a probability larger than 0.5. In some case it is favorable to deviate from this rule. When is this the case? What happens to the overall error rate? What is the optimal threshold in this case?
- What is an ROC curve? How do I construct such a curve? What is considered the optimal operational point of the classifier? What does it mean if the AUC of the ROC is 0.5? Draw this! Explain what is plotted on the axis? How does it relate to sensitivity and specificity? Explain by using a confusion matrix.
- Exercise 4.7.2, 4.7.3, 4.7.8

- Consider the ROC curve below:



Given the type 2 error in blue and the type 1 error in orange in function of the classification threshold. Which threshold is optimal for this classifier? Indicate this point in both plots. Indicate in the ROC curve where the value for threshold 0 should lie.



Chapter 5:

- What are the two objectives of validation?
- How does the validation set approach work? What are the disadvantages?

- How does leave-one-out-cross-validation work? Why has it less bias than the validation set approach?
- Explain k-fold cross-validation? What happens when k = n?
- Explain the bias-variance trade-off in validation? Why has the LOOCV the highest variance and the lowest bias on the prediction of the test MSE? Can you illustrate this with the variance of the sum of two correlated stochastic variables.
- When using LOOCV in the context of least squares linear or polynomial regression the estimated test MSE can be computed as follows

  $$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

  Where the model is fitted on the entire training dataset and no validation dataset is hold-out from the training. Can you argue on the interpretation of this equation? Compare the short-cut formula with LOOCV in the case an observation with leverage $h_i = 0$ is hold-out for validation. What is the function of denominator?
- What is bootstrapping and how does it differ from validation?
- Exercise 5.4.1