# Exam Machine Learning

January 3rd, 2025

13h00 – 16h00

## Write your name on every paper!

Use a text processor to type in your answers and convert the file to pdf format. Use the Blanco paper for figures and equations; Please sanitize your handwriting and clearly enumerate all your answers.

Use illustrations, formulas, and examples to clarify your answers! Use scratch paper for this purpose and scan it using your phone at the end of the exam.

1) Explain the Bias-Variance trade-off. Derive the equation and use figures to describe the concept. How does the bias-variance trade-off relate to the reducible and irreducible error? What is the ideal situation regarding the bias and the variance? Why is this trade-off important in machine learning? (**10pnts**)

2) Explain the K-means clustering! Use figures for the explanation. Which hyperparameters and metrics can be tuned for the k-means clustering technique? Is K-means clustering a stochastic or deterministic procedure? (**4pnts**)

3) Illustrate that applying the principle of the optimal Bayes classifier to the linear discriminant 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 discriminant function below is largest (**5pnts**)

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

4) Give a complete description of classification tree construction and pruning. What is the problem with trees? Explain the concept of bagging. How does it solve the problem of the trees? What is a random forest, and how does it differ from bagging? Clearly Indicate the differences between the three methods and explain the hyper-parameters. (**10pnts**)

5) What is a separable hyper-plane? How can it be used for classification? Is any hyper-plane optimal? How are hyper-plane classifiers improved? Explain the extension to maximum margin classifier, support vector classifier and support vector machine. Explain how the variable C in the optimization problem of a support vector classifier controls the bias-variance trade-off. Use

2 illustrations to explain when $\varepsilon_i$ is larger or smaller than 1. What is the hinge loss in this equation? (**10pnts**)

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,M}{\text{maximize}} \quad M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

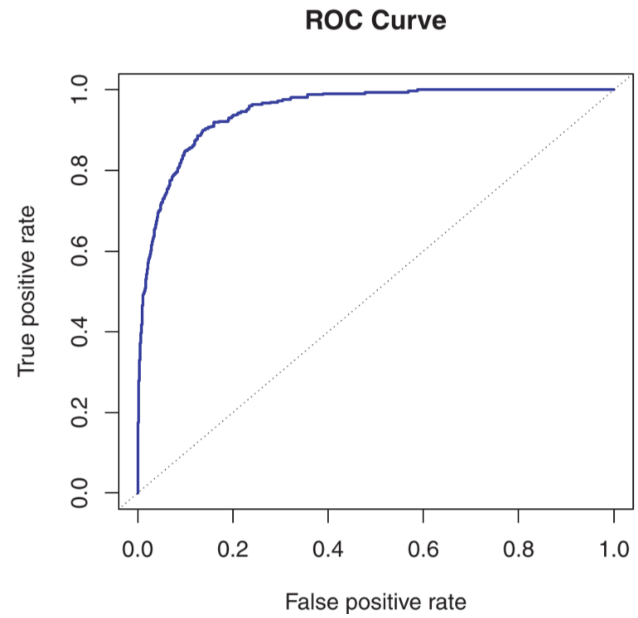
6) Why shouldn't we use linear regression to model binary outcome variables?
Proof that (**5pnts**):

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

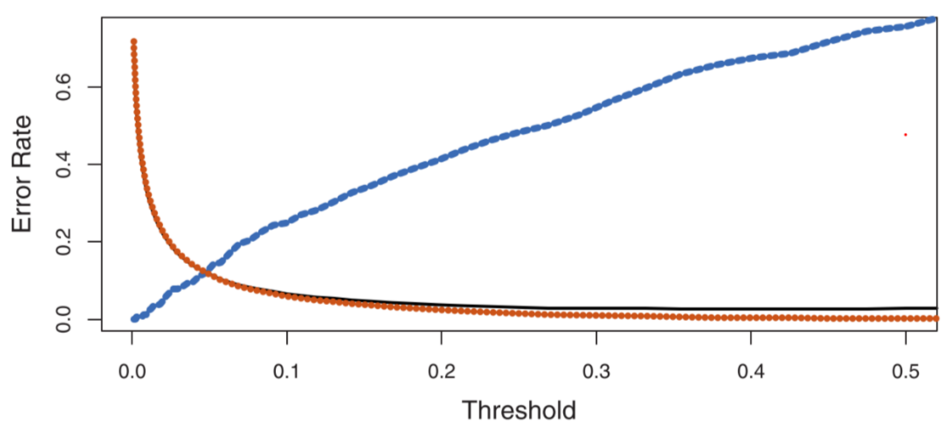
7) Consider the confusion matrix below. This matrix gives an algorithm's predictions that aims to predict whether a client will default on his loan payments. Explain its purpose. What is the overall error rate of this classifier? Compute the sensitivity and specificity and describe these terms in your own words. Which metric is preferred in this case: error rate or sensitivity/specificity? (**3pnts**)

| | | *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 |

8) Consider the ROC curve below. Can you indicate the sensitivity and specificity of question 7 using the optimal Bayes threshold on the plot below?




Consider the additional graph below which displays 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. (**3pnts**)




**Good luck with the exam!!**

**Dirk & Inigo**