STOCKHOLM UNIVERSITY
Department of Statistics
Matias Quiroz

FINAL EXAM, MACHINE LEARNING ST5401
2023-01-11

Time for examination: 08.00-13.00

Allowed tools:
Course book (supplied as pdf): Lindholm, A., Wahlström, N., Lindsten, F. and Schön, T. *Machine Learning - A First Course for Engineers and Scientists*.
RStudio: The required packages are loaded in the supplied `R markdown` template `Template_exam_solutions.Rmd`. You may use any additional packages that are already installed.

Instructions: Submit your solutions as a html file generated by `R markdown`, including relevant code, figures, calculations and discussions. Detailed instructions on how to submit your exam are are found on `tenta.stat.su.se/tenta`. Paper sheets are handed out in case you prefer to provide handwritten solutions/discussions. It is recommended to submit your code as a separate markdown (.Rmd) (or .R file if you do not use markdown), to be used as backup. A markdown template is provided (`Template_exam_solutions.Rmd`), but feel free to use your own.

You are not allowed to use the internet during the exam. You are allowed to copy code from your own (possibly polished) computer lab solutions. Copying text from the textbook is strictly forbidden.

Exam format and grades: The exam consists of 4 problems worth a total of 40 marks. *The problems are not sorted by their degree of difficulty.*
Grades (out of 40): A: 36, B: 32, C: 28, D: 24, E: 20, Fx and F: < 20.

`Good luck!`

Problem 1. (10 marks)

Provide a well-motivated answer for each sub-question below. No need to carry out any computations or simulations on your computer. No need for lengthy discussions, I just want you to show that you understand your answer.

(a.) Discuss the bias-variance tradeoff in the context of the k-nearest neighbour algorithm.

(b.) Suppose that the true data generating process is

$$y = \beta_0 + \beta_1 h_1(x_1) + \beta_2 h_2(x_2) + \varepsilon,$$

where h_1 and h_2 are nonlinear functions. Suppose that you fit a simple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Explain the concept of generalisation gap and how you expect it to behave in this particular situation.

(c.) Explain the underlying idea of the trust region Newton method and which shortcoming of Newton's method it addresses.

(d.) What is the purpose of boosting models (as opposed to just fitting a single model on the given dataset)?

(e.) Describe early stopping in an iterative optimisation algorithm. What is the purpose of early stopping?

Problem 2. (10 marks)

The dataset `penguins.RData` contains the dive heart rate (in beats per minute) during a dive and the corresponding duration of the dive (in minutes) for 125 penguins. Assume the dive heart rates y_i follow a Gaussian process (GP) regression model with scaled durations x_i as inputs, i.e.

$$\begin{aligned} y_i &= f(x_i) + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \\ f(x) &\sim \text{GP}(0, k(x, x')), \end{aligned}$$

where

$$k(x, x') = \sigma_f^2 \exp\left(\frac{1}{2\ell^2}(x - x')^2\right), \tag{1}$$

with scale factor $\sigma_f^2 > 0$ and length scale $\ell > 0$. The duration of dive i is scaled into the interval $[0, 1]$ by

$$x_i = \frac{\text{duration of dive } i}{\text{max duration of all dives}}.$$

(a.) Discuss the role of the parameter ℓ and its implication for the bias-variance tradeoff.

(b.) Suppose that $\sigma_f^2 = 10000$, $\sigma_\varepsilon^2 = 150$ and $\ell^2 = 0.6$. Compute the prior correlation between the dive heart rates of two penguins whose diving times are 4 minutes apart.

(c.) Suppose that $\sigma_f^2 = 10000$, $\sigma_\varepsilon^2 = 150$ and $\ell^2 = 0.6$. Predict the Gaussian process on a denser grid of scaled durations (in the interval 0-1) and plot it together with the training data.

(d.) Suppose now that σ_f^2, σ_ε^2 and ℓ^2 are unknown. Suggest how you can estimate them. No need for an implementation, but clearly outline how you would proceed.

Problem 3. (10 marks)

The breast cancer Wisconsin dataset[1] contains features computed from digitised images for 569 malign (`diagnosis=1`) and benign (`diagnosis=0`) tumors. The features describe characteristics of the cell nuclei extracted from the image. The dataset has been split into `cancer_data_training.RData` containing $n = 400$ observations to be used for training the classifier below, and `cancer_data_test.RData` containing $n = 169$ to evaluate its performance. You should standardise the features to have zero mean and unit variance. If `X` is a matrix of features, this is achieved by `X<-scale(X)` in `R`. Note that if the model has an intercept, it cannot be included in `X` when calling the `scale` function; include it after the standardisation.

(a.) Fit a logistic regression (include an intercept in the model) with an L2 penalty (ridge regression) using the `cancer_data_training.RData`. Estimate the optimal λ by $K = 4$ cross-validation using `cv.glmnet` (only using `cancer_data_training.RData`).

(b.) Compute the confusion matrix (for example using the `caret` package) for the test data (`cancer_data_test.RData`) with a threshold of 0.5 for the model estimated in (a.). Compute the precision and recall of the classifier.

(c.) Interpret the meaning of the precision and recall in the context of the cancer diagnosis example. Which one would you prefer to be highest?

[1]Source: `https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data`.

Problem 4. (10 marks)

The Iris flower dataset was introduced by the famous British statistician Ronald Fisher. The dataset contains $n = 150$ samples from three species of Iris (Iris setosa, Iris virginica and Iris versicolor). For each species, the length and the width (in centimeters) of the sepals and petals are measured. In this problem, we model only the length and width of the petals, i.e.

$$\mathbf{x}_i = (\text{length of petal for flower } i, \text{width of petal for flower } i),\ i = 1, \ldots, 150.$$

(a.) The dataset `iris_fully_labelled.RData` contains a fully labelled dataset. Fit a *supervised* Gaussian mixture model with three components to the data, assuming each component has a separate covariance matrix (i.e. quadratic discriminant analysis, QDA).

(b.) Suppose we find a new Iris flower with petal length 1.7 cm and petal width 0.35 cm. Which species does it belong to according to your model in (a.)?

(c.) The dataset `iris_no_labels.RData` contains the same order of the observations as in (a.), but no labels for any of the flowers. Use the EM algorithm to fit an *unsupervised* Gaussian mixture model with three components to the data (assume also here that the components have different covariance matrices). Compare the inferred class labels to the known labels in (a.). How many correctly predicted labels do you get for each of the three classes (species)?