STOCKHOLMS UNIVERSITET HT2023

Statistiska institutionen
Dan Hedlin

Machine learning, ST5401

Examination 2022-11-27, 14.00 – 19.00

Approved aids:

1. Course book (supplied as pdf): Lindholm, A., Wahlström, N., Lindsten, F. and Schön, T. Machine Learning - A First Course for Engineers and Scientists.
2. RStudio
3. Computer lab solutions are supplied.
4. Language dictionary

Instructions:

Submit your solutions as a PDF file, including relevant code. All solutions should be in the PDF file, which is the one we grade. It is recommended to also submit your code as a separate markdown (.Rmd) (or .R file if you do not use markdown), to be used as backup. You are not allowed to use the internet during the exam. You are allowed to copy code from your own computer lab solutions (edited if needed). Copying text from the course book is not allowed.

The exam comprises four items, numbered 1 to 4. The maximum number of points is 40. Grades: A: at least 36, B: 32, C: 28, D: 24, E: 20, Fx and F: < 20. To obtain the maximum number of points full and clear motivations are required unless otherwise stated. You may write in English or Swedish.

1.

Provide well-motivated answers for each of items a to d below. You are not required to carry out any computations or simulations on your computer.

a) Discuss the bias-variance trade-off in the context of decision trees.
b) Suppose that the true data generating process is

$$y = \beta_0 + \beta_1 h_1(x_1) + \beta_2 h_2(x_2) + \beta_3 h_3(x_3) + \varepsilon$$

where h_1, h_2 and h_3 are nonlinear functions. Suppose that you fit a linear regression model

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

Let us assume that $\varepsilon \sim N(0,1)$ in both models. Explain the concept of generalisation gap and how you expect it to behave in this situation. As the generalisation gap depends on h_1, h_2 and h_3, you might decide to organise your discussion around some scenarios of the forms of h_1, h_2 and h_3.
c) Discuss advantages (if any) and disadvantages (if any) of doing binary splits (one node into two nodes) in a tree as opposed to multiway splits (one node into more than two nodes).
d) In a dense neural network, discuss how the number of hidden units affects the bias-variance trade-off.

Maximum 10 points.

2.

The breast cancer Wisconsin dataset[1] contains features (variables) computed from digitised images of 569 malign (diagnosis=1) and benign (diagnosis=0) tumours. The features describe the 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 the training of the classifiers below, and `cancer_data_test.RData` containing n = 169 observations to evaluate the performance of the classifiers. You should standardise the features to have zero mean and unit variance. If $\mathbf{X}$ is a matrix of observations and features, this is achieved by `X<-scale(X)` in R. Note that if the model has an intercept, it cannot be included in $\mathbf{X}$ when calling the scale function; include it after the standardisation.

a) Fit a logistic regression with an intercept and with an L1 penalty (lasso) using the `cancer_data_training.RData`. Model the probability that the tumour is malignant. Estimate the optimal regularisation parameter λ by $K = 4$ cross-validation using cv.glmnet
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. Interpret the meaning of precision and recall in the context of the cancer diagnosis example.
c) Fit a classification tree to the breast cancer data using the `tree` package with the default settings. Compute the confusion matrix for the test data, again with threshold 0.5. Compute the precision and recall of the classifier. Considering these measures, and accuracy, would you prefer this classifier or the one in a)?

Maximum 10 points.

3.
Consider again the breast cancer dataset that was used in problem 2. Suppose that we model the data as a dense neural network with several layers,

$$\mathbf{q}^{(1)} = h\left(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right)$$
$$\mathbf{q}^{(2)} = h\left(\mathbf{W}^{(2)}\mathbf{x} + \mathbf{b}^{(2)}\right)$$
$$\mathbf{q}^{(3)} = h\left(\mathbf{W}^{(3)}\mathbf{x} + \mathbf{b}^{(3)}\right)$$

$$z = \mathbf{W}^{(4)}\mathbf{q}^{(3)} + b^{(4)}$$

$$\Pr(y = 1|\mathbf{x}) = g(z)$$

where h is an ReLU activation function, g is the soft-max function and y is diagnosis. Suppose further that the first layer has 16 hidden units, the second layer has 8 hidden units, and the third layer has 4 hidden units. The `keras` package is not available in the exam room, so no code is required to answer the questions.

a) What is the number of parameters for each of the equations above ($\mathbf{x}$, q and z are not parameters)?
b) What are the dimensions of $\mathbf{q}^{(1)}$, $\mathbf{q}^{(2)}$, $\mathbf{q}^{(3)}$, $\mathbf{b}^{(1)}$, $\mathbf{b}^{(2)}$, $\mathbf{b}^{(3)}$, $b^{(4)}$, $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{W}^{(3)}$, $\mathbf{W}^{(4)}$ and z? No motivation is required

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

c) Describe and explain two ways of regularising a neural network.

Maximum 10 points.

4.

The Iris flower dataset contains n = 150 samples (observations) from three species of Iris: Iris setosa, Iris virginica and Iris versicolor. For each species, the length and the width in centimetres of the petals were measured.

$\mathbf{x}_i$ =(length of petal for flower i, width of petal for flower i), $i = 1,2, \ldots 150$.

a) The dataset `iris_supervised_training.RData` contains a fully labelled training dataset with n = 110. Fit a supervised Gaussian mixture model with three components to the data, assuming all components have the same covariance matrix (i.e. linear discriminant analysis, LDA).
b) Fit a supervised Gaussian mixture model with three components to the data in a) assuming now that each component has a separate covariance matrix (i.e. quadratic discriminant analysis, QDA).
c) Evaluate the methods in a) and b) using the test dataset `iris_test.RData` with n = 40. Choose some appropriate performance measures.
d) The dataset `iris_semisupervised_training.RData` contains the same number of observations as the dataset in a), but only partially labelled flowers. Use the EM algorithm to fit a semi-supervised Gaussian mixture model with three components to the data (assume that the components have different covariance matrices). Compare the inferred class labels (predicted species) to the known labels in the training data in a). How many correctly predicted labels do you obtain for each of the three species in the training data?

Maximum 10 points.