STOCKHOLMS UNIVERSITET VT2024

Statistiska institutionen
Dan Hedlin

Machine learning, ST5401

Examination 2024-01-04, 08.00 – 13.00

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
5. Pocket calculator

Instructions:

Submit your solutions as a PDF or HTML file, including relevant code. We recommend that you 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 required to carry out any computations or simulations on your computer. However, any non-trivial code that you rely on must be included in your reported solutions. 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 three items, numbered 1 to 3. 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 f below.

a. Explain what makes a convolutional neural network different from a dense neural network.

b. Are unbiased estimators always desirable? Discuss a situation in which you would prefer an unbiased estimator and another in which you would prefer a biased estimator.

c. Why is a decaying learning rate needed for a stochastic gradient descent algorithm to converge?

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

e. Why are features randomly selected in a random forest (as opposed to using all features in all splits throughout)?

f. Explain how a good value of k is chosen in k-nearest neighbour.

Maximum 18 points.

2.

a) What is the problem with this matrix, if you need the inverse of **X**?

$$\mathbf{X} = \begin{pmatrix} 1 & 4 & 3 \\ 2 & 2 & 4 \\ 3 & 3 & 6 \end{pmatrix}$$

b) Suppose you need the inverse of $\mathbf{X}^T\mathbf{X}$, that is, $(\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}$. Suggest a way of regularising the problem.

c) Use your idea to write a matrix **T** that is approximately equal to $\mathbf{X}^T\mathbf{X}$. Compute $\mathbf{T}^{-1}\mathbf{X}^T\mathbf{X}$. (It may not be close to the identity matrix.)

Maximum 9 points.

3.

A function is going to be fitted to data: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

The loss function is the sum of squared errors. Use stochastic gradient descent to estimate the parameters. The data consists of only six observations, which are displayed in the table below.

Unit	y	x_1	x_2
1	2	4	1
2	-14	2	8
3	1	1	0
4	-1	3	2
5	-7	1	4
6	8	6	7

Let the step size multiplier be $\gamma = 0.05$ and each mini-batch consist one observation. Starting values are selected randomly; in this case they turned out to be:

$\beta_0^{(1)} = -0.066$

$\beta_1^{(1)} = -0.044$

$\beta_2^{(1)} = 0.042$

The next step, step 2, of the algorithm of stochastic gradient descent will produce the values $\beta_0^{(2)}$, $\beta_1^{(2)}$ and $\beta_2^{(2)}$.

A minibatch is selected randomly. It turned out to be unit 1.

If there are any further random elements or choices needed, just make some "random" choice(s) as though a truly random process were in use.

a) What is the gradient in this problem? Express it in a formula as close as possible to something that can be used for computation.

b) What is the general purpose of using the gradient in gradient decent (as opposed to for example the vector $(1,1,1)^T$)?

c) What is the prediction of y in unit 1 in step 1: $\hat{y}^{(1)}$? Express it in both a formula and a number.

d) How are $\beta_0^{(2)}$, $\beta_1^{(2)}$ and $\beta_2^{(2)}$ produced? Express it in formulas, not numbers.

e) What are $\beta_0^{(2)}$, $\beta_1^{(2)}$ and $\beta_2^{(2)}$? Express them in numbers.

Maximum 13 points.