# Answers to Exam info 284

## 1. 3-NN – continuous inputs

You are running a k-NN algorithm with k = 3, and use the L2 norm to measure distance between data points. You want to classify a new data point with the following scaled features

(a = 0.9, b = 0.5, c = 0.2, d = 1.0)

The 5 closest data points in the training set has values

(a = 0.9, b = 0.4, c = 0.2, d = 0.8, class = A, y = 4.3)
(a = 1.0, b = 0.5, c = 0.2, d = 1.0, class = B, y = 5.3) *close*
(a = 0.6, b = 0.5, c = 0.2, d = 1.0, class = A, y = 3.6)
(a = 0.8, b = 0.6, c = 0.3, d = 0.9, class = A, y = 4.1) *close*
(a = 0.9, b = 0.5, c = 0.1, d = 1.0, class = B, y = 5.0) *close*

The data is used for classification and for prediction of y. What are the predictions for "class" and "y"?

a) B and 4.8 * (majority among three closest for class, and average for y)
b) A and 4.0
c) B and 5.0
d) A and 4.4

## 2. Accuracy

You have a test set with 1000 data points for measuring how good your machine learning model is. The 454$^{th}$ data point in the test set is classified correctly. How much does this data point contribute to the accuracy score of the model?

a) 0.001 * (every correctly datapoint adds 1/1000 to the accuracy score)
b) 0.454
c) 4.540
d) 0.000

## 3. Selecting a Naïve Bayes approach

You have decided that you would like to use a naïve Bayes approach for your classification model that classifies letters to the editor in a newspaper as one of (Politics, Health, Transport, Other, Rubbish). You using a tool that counts the numbers of each word in each document and choose to represent each document by a bag-of-words vector where each entry in the vector represents the count of a particular word. Which Naïve Bayes-approach fits best to this type of data?

a) Multinomial Naïve Bayes * (This is standard task for multinomial Naïve Bayes)
b) Categorical Naïve Bayes
c) Bernoulli Naïve Bayes
d) Gaussian Naïve Bayes

## 4. Linear regression

You have a data set with 459 samples with 8 input continuous features and one continuous target feature. You want to use this to learn a simple linear regression model. How many parameters does your algorithm learn?

a) 9 * (you need one weight for each input feature (8) + the bias (1) = 9 )
b) 8
c) 459
d) 458

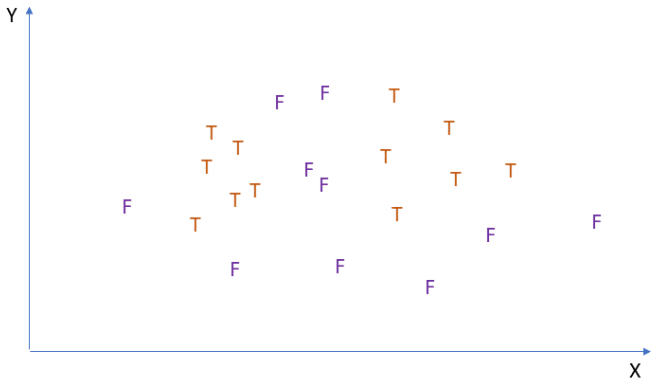
## 5. Logistic function

You are doing binary classification (classes are T or F). In logistic regression one uses the logistic function to estimate the probability $P(c = T)$ of a classification T for a new data point (x,y). You have found that the line 2x – y – 4 = 0 is the line that separates your data best with most T cases being below this line. What is the logistic function based probability for a new data point (x=4,y=1) being in class T?

a) 0.952 * (put 2x4-1-4 into the logistic function)
b) 1.459
c) 0.500
d) 0.002

## 6. Choosing the right kernel

You aim to use a kerneled support vector classifier on your data. You have two input values X and Y and your classification is T or F. You randomly select 22 data points and find that they are placed in the feature space according to this figure. Now it is time to decide what kernel function that fits best. Which fits best?




a) Radial Basis Function Kernel * (this has the ability to detect “islands” in the data, given the right gamma and c parameters)
b) Linear Kernel
c) Quadratic Kernel
d) Logistic Function Kernel

## 7. Decision tree search

Assume that you have the following dataset with classification in column Class

| No | X | Y | Z | W | Class |
|---|---|---|---|---|---|
| 1 | a | g | m | s | T |

| | | | | | |
|---|---|---|---|---|---|
| 2 | b | h | m | s | T |
| 3 | b | g | k | s | T |
| 4 | b | h | k | r | F |
| 5 | a | h | k | s | T |
| 6 | b | g | k | r | T |
| 7 | b | g | m | s | F |
| 8 | a | h | k | r | F |
| 9 | a | h | m | r | F |
| 10 | a | g | m | r | F |

Identify the variable that would be the root of a decision tree, assuming that we select the variable that provides most information gain as the split variable.

a) X
b) Y
c) Z
d) W * (this divides data with s into 4-1 and with r into 1-4, so s gives high likelihood of T and r high likelihood of F. The other variables has 2-3 and 3-2 divisions which does not help so much)

## 8. Gradient boosting decision tree regression

You run a very simple gradient boosting decision tree regression algorithm resulting in only three trees and then a simple additive method for prediction. The trees return the following predictions for a data point:

Tree 1: 4.568
Tree 2: - 1.091
Tree 3: 0.873

What is the final prediction?

a) 4.350 * (A standard gradient boosting algorithm just adds the predictions from each tree, as each tree in a sense predicts the error in the previous one)
b) 1.450
c) 2.873
d) 0.873

## 9. Threshold function

Your friend hates imprecision and insists on using a threshold function as the activation function in a multilayered neural network that is learned through backpropagation and gradient descent. What will happen?

a) The gradient descent algorithm to learn the weights will not be able to converge * (this is really the only meaningful answer)
b) The total loss will become too large in order to compute the weights
c) The weights will themselves become threshold functions
d) The accuracy will be slightly worse than with a sigmoid activation function, but at less computational cost

## 10. Number of parameters

You have a multilayered standard neural network with 144 input values, one hidden layer with 16 units (neurons) and 4 units in the output layer. How many parameters will need to be learned in such a neural network (no regularization)?

a) 2388 *((144+1)*16 + (16+1)*4)
b) 2368
c) 2348
d) 2408

## 11. Convolutional layer

Your image recognition system accepts images with 15 x 15 graytone pixels, and you want to build a convolutional neural network that classify such images. The input layer consists of the graytone values. The first layer you add is a convolutional layer based a receptive field of 3 x 3 and a stride length of 2. You assess that you need 3 feature maps in the layer. How many units (neurons) will be in total be created in the convolutional layer?

a) 147 * (7 receptive fields in both x and y dimension, 3 feature maps give 7 x 7 x 3 )
b) 7
c) 49
d) 135

## 12. Loss function

You have a convolutional neural network that should be able to classify images of children according to their age, with the following classes: 0-1 year, 1-4 year, 4-7 years, 8-12 years, 13-18 years, not child). Your last layer is a so-called softmax layer so the output nodes gives numbers representing the probability of each of the 6 classes. Which loss function for the purpose of backpropagation would you normally use here?

a) Cross-entropy * (cross-entropy is the only meaningful loss function when using softmax, the others are essentially not meaningful)
b) Accuracy
c) Precision
d) L2 distance

## 13. LSTM units

LSTM units are rather complex structures, with a number of weight matrices. In a standard RNN unit we have two weight matrices, one for assigning weights to the current state and one for assigning weights to the input when updating the state. In LSTM units we have so-called input gates, output gates, and forget gates in addition to the input itself. How many weight matrices to we normally need to learn?

a) 8 * (a difficult question, there are two matrices for each gate plus two for the input itself)
b) 5

c) 16
d) 12

## 14. Binning and one-hot encoding

You have a data set with 4 feature values and one target value. The feature $X_2$ may be divided into 5 equally sized intervals where data points in each of the intervals has a relatively small variance with regard to the target value. You decide to test out an approach with binning where the 5 bins of $X_2$ are one-hot encoded. You keep $X_2$ in the data set, and also add a new feature $X_s$-squared which is the square value of the values in $X_2$. How many features do you now have in your data set?

a) 21
b) 11
c) 10 * (one for each original feature (4), one for each bin (5), one for the squared value)
d) 8

## 15. Confusion matrix

A medical team wants to use AI and machine learning to improve their treatment of a particular disease that is difficult to diagnose. The conclusion about the right diagnosis can only be found a week after the suspected onset of the disease. They have one possible treatment for the disease, but it is a bit unpleasant for the patient, and may lead to lifelong, but less serious health problems. Not treating the disease will be fatal for the patient, which is not acceptable. The team has gathered data on the disease in form of symptoms and the final diagnoses for earlier cases.

Machine learning algorithms A, B, C and D has different outcomes with these data, all are good with respect to accuracy. But you would like to base your decision also on the confusion matrices for the approaches with 834 test cases. They are given below.

| Algorithm A | Actually Had Disease | Actually did not have disease |
|---|---|---|
| Predicted Disease | 219 | 42 |
| Predicted No Disease | 17 | 556 |

| Algorithm B | Actually Had Disease | Actually did not have disease |
|---|---|---|
| Predicted Disease | 284 | 7 |
| Predicted No Disease | 52 | 591 |

| Algorithm C | Actually Had Disease | Actually did not have disease |
|---|---|---|
| Predicted Disease | 233 | 86 |
| Predicted No Disease | 3 | 512 |

| Algorithm D | Actually Had Disease | Actually did not have disease |
|---|---|---|

| | | |
|---|---|---|
| Predicted Disease | 152 | 0 |
| Predicted No Disease | 84 | 598 |

Which algorithm is the most promising to continue with given the medical constraint.

a) A
b) B
c) C * (doctors do not want false negatives, death in unfixable, this one has near 0 false negatives)
d) D

## 16. Cross-validation

You have a data set about customers in a csv file that is sorted based on their total monthly purchases in terms of money spent. The data set contains 1012 customers, and you would like to assess your time-consuming machine learning model based on cross-validation. What kind of technique among these would you use?

a) Stratified 5-fold cross-validation * (data is sorted, so you need to shuffle them)
b) Leave one out cross-validation (this takes too long, 1012 runs of a time-consuming model)
c) Non-stratified 10-fold cross-validation (data becomes unsorted, and you will get strange results)
d) Random grid search (not cross-validation)

## 17. Principal Component Analysis

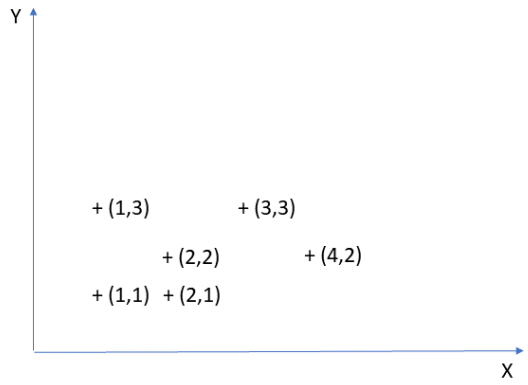
You have a data set with 11 input values, and find that the sum of variances in all of the variables is 19.2 You compute the principal components of the data and find that new principal components $Z_1$, $Z_2$, and $Z_3$ has variances 11.2, 6.7 and 1.1 respectively. What do you think is the sum of the variances of the remaining 8 principal components:

a) 0.2 * (sum of variances is equal in both feature sets, only 0.2 is remaining)
b) 1.6
c) 11.4
d) 6.3

## 18. Agglomerative clustering

You run an agglomerative clustering algorithm on the data points you see in the figure below. You want three clusters. The criterion for selecting groups for merging is the pair with least maximal distance among points in the pair.

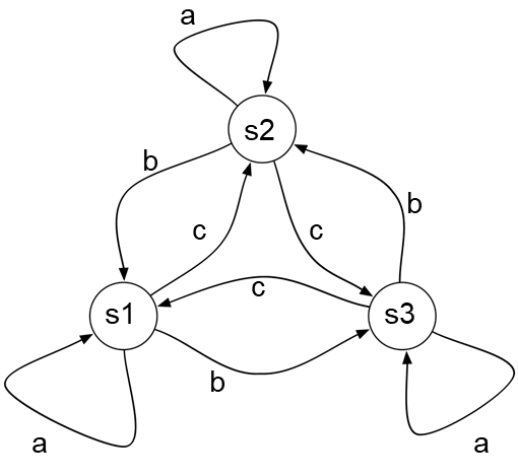
How many groups will there be in each cluster when the algorithm stops?




a) 3, 2, and 1 * (two answers are correct, as a bright student pointed out to me)
b) 4, 1, and 1
c) 2, 2, and 2 * (this is also correct, depending on ordering of datapoints)
d) 4, 2, and 0

## 19.Q-learning

You want to find an optimal strategy for the deterministic decision model below, and run a q-learning approach.




| state | action | Q-value |
|---|---|---|
| s1 | a | -1 |
| s1 | b | 1 |
| s1 | c | 1 |
| s2 | a | -1 |
| s2 | b | 1 |
| s2 | c | 1 |
| s3 | a | -1 |
| s3 | b | 1 |
| s3 | c | 1 |

Initially you have heard that it may not be so smart to stay in the same state, so there may be a small punishment (negative reward) if you do action a. For the other actions you think these will lead to positive rewards. This means that you start with q-value -1.0 for all state action pairs $Q(s_i,a_j)$ where $a_j$ is equal to action a. You also start with q-value 1.0 for all other states and actions. See table above. You now are in state s3 and randomly select action b giving you a reward of 5. You have a learning rate of 0.1 and and a discount factor of 0.5. What will the Q-value of s3 and b be afterwards?

a) 1.45 * (Application of standard q-learning rule)
b) 3.05
c) 1.9
d) 1.5

## 20. Exploration vs exploitation

IKEA's robot runs a q-learning algorithm to learn moving around in the IKEA warehouse to perform its various tasks. It quickly learned to find its way through, but seems to find way too long and inefficient routes through the warehouse. You believe that this has to do with exploration and exploitation strategy used in the learning process. What could be done to improve performance?

a) Adjust so it uses an explore strategy more often in the beginning of learning * (best strategy to behave randomly early, and then focus more and more on best choices, the other options will just behave worse or not improve at all)
b) Adjust so it uses an exploit strategy more often in the beginning of learning
c) Adjust so it uses an explore strategy increasingly often as the learning proceeds
d) Adjust so it mainly uses an explore strategy