



KANDIDAT

1019

PRØVE

# DAT550 Practice exam

Emnekode	--
Vurderingsform	--
Starttid	08.05.2020 09:00
Sluttid	08.05.2020 11:00
Sensurfrist	--
PDF opprettet	08.05.2020 11:01

## Information

Oppgave	Tittel	Oppgavetype
i	Information	Dokument

## Dimensionality Reduction

Oppgave	Tittel	Oppgavetype
1	PCA	Sammensatt

## Classification

Oppgave	Tittel	Oppgavetype
2	Decision Tree	Sammensatt
3	Naive bayes properties	Sammensatt
4	Naive bayes	Sammensatt

## Deep Learning

Oppgave	Tittel	Oppgavetype
5	Neural network	Sammensatt
6	Learning curves	Paring
7	Activation functions and neural networks	Sammensatt
8	RNNs	Sammensatt

## Clustering

Oppgave	Tittel	Oppgavetype
---------	--------	-------------

9	Clustering 1	Sammensatt
10	Clustering 2	Sammensatt

### Scratch area

Oppgave	Tittel	Oppgavetype
11	Provide your explanation for numerical computations here!	Tekstfelt

# 1 PCA

The goal of PCA is to interpret the underlying structure of the data in terms of the principal components that are best at predicting the output variable.

**Select an alternative**

- ☐ True
- ☒ False

The output of PCA is a subset of the original features in lower dimensions

**Select an alternative**

- ☒ True
- ☐ False

The output of PCA is a new representation of the data that is always of lower dimensionality than the original feature representation.

**Select an alternative**

- ☒ True
- ☐ False

Subsequent principal components are always orthogonal to each other

**Select an alternative**

- ☒ True
- ☐ False

Assume we apply PCA to a matrix  $X \in \mathbb{R}^{n \times m}$  and obtain a set of PCA features,  $Z \in \mathbb{R}^{m \times n}$ . We divide this set into two parts,  $Z_1$  and  $Z_2$ . The first part,  $Z_1$ , corresponds to the top principal components. The second set,  $Z_2$ , corresponds to the remaining principal components. Is it common to expect a point with large feature values in  $Z_2$  and small feature values in  $Z_1$ ?

**Select an alternative**

- ☐ True
- ☒ False

Maks poeng: 10

## 2 Decision Tree

It is late Norwegian summer and you decide to go for a mushroom picking with your friends. There are lots of types of mushrooms in the forest, but some of them can be deadly because they may be poisonous. Fortunately, one of your friends has collected some attributes of the poisonous and non-poisonous mushrooms from the past trip as below:

P.S: Don't use this data to classify mushrooms in real life! :D

Mushroom

Sample					
a					
b					
c					
d					
e					
f					
g					

Sample	Heavy	Smelly	Spotted	Scales	Poisonous
a	0	0	0	0	0
b	0	0	1	0	0
c	1	1	0	1	0
d	1	0	0	1	1
e	0	1	1	0	1
f	0	0	1	1	1
g	0	0	0	1	1
h	1	1	0	0	1

(a) What is the entropy of Poisonous label?

(b) Which of the attributes should you choose as root of the decision tree? Hint: You may have to compute information gain.  (Scales, Spotted, Heavy, Smelly)

(c) What is the information gain of the you got for the attribute you chose the previous questions?

(c) Build a decision tree to classify a mushroom which has all of the attributes as 1

**Select one alternative**

- ☒ Poisonous (1)
- ☐ Not Poisonous (0)

Maks poeng: 16

### 3 Naive bayes properties

Which of the following are FALSE about Naive bayes?

**Select one or more alternatives:**

- ☐ Naive bayes is an unsupervised classification algorithm
- ☒ Naiveness in naive bayes comes from the conditional independence assumption.
- ☒ Naive bayes can deal with missing attributes
- ☐ The Naive Bayes algorithm is sensitive to irrelevant attributes.
- ☐ If there are dependencies between the features Naive bayes is preferred.

Maks poeng: 5

## 4 Naive bayes

(16%) Consider the training and testing data in the Table below. Classify the test records in Table b using the Naive Bayes classifier trained on the training data in Table a. You only need to compute the probabilities you will need for the classification. For your answer, you need to tell which class has the highest posterior probability – you do not have to compute the final posteriors as long as it is clear which one is bigger.

X	Y	X	Class
1	1	1	+
1	0	0	-
1	1	1	+
1	1	0	+
1	1	1	-
1	1	0	-
0	1	1	-
0	1	1	+
0	0	1	+
0	0	0	+

Using the naive bayes for training from above training data We are interested in classifying two test data entries:

Test1: X=0, Y=1, Z=0

Test2: X=1, Y=0, Z=1

Compute the likelihood of test cases belonging to certain class.

Note: You can omit the final posterior (the denominator) for computing the  $P(X,Y,Z)$  in the bayes theorem formula to fill in the following probabilities.

$$P[C = + | X = 0, Y = 1, Z = 0] = 0.055$$

$$P[C = - | X = 0, Y = 1, Z = 0] = 0.093$$

$$P[C = + | X = 1, Y = 0, Z = 1] = 0.11$$

$$P[C = - | X = 1, Y = 0, Z = 1] = 0.093$$

Class for X=0, Y=1, Z=0?  (+, -)

Class for X=1, Y=0, Z=1? placeholder

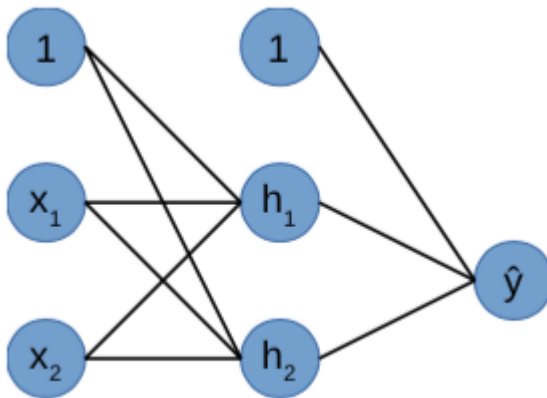


## 5 Neural network

(5%) Suppose you are given the following neural network (flowing left to right) and parameters, which uses ReLU activation in every layer and the loss is the squared error  $L = (y - y')^2$ .

Weights for the first layer (start from bias)  $W_1 = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \end{bmatrix}$

Weights for second layer (start from bias)  $W_2 = \begin{bmatrix} 1 & 1 & 2 \end{bmatrix}$



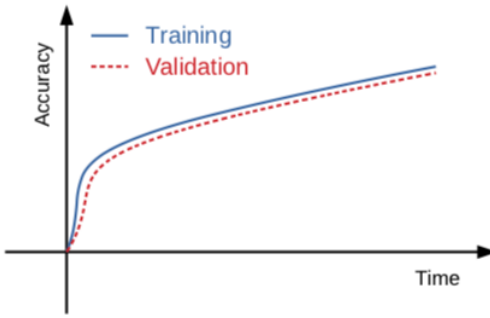
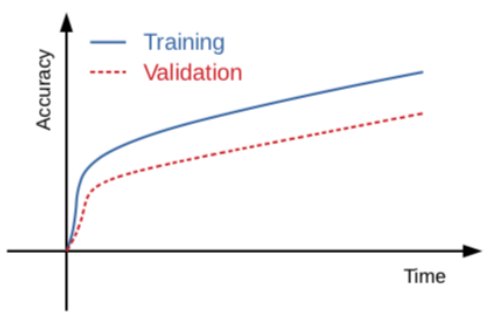
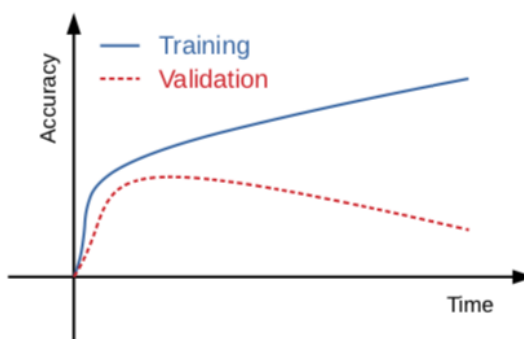
What is the loss value (squared loss) for the input  $x = (2 \ 4)$  and  $y = 26$

Maks poeng: 6

## 6 Learning curves

For each of the following pairs of training and validation curves, select the appropriate reason for this behaviour.

Please match the values:

	Too small training data	Underfitting	Overfitting
<p>Plot 3</p> 	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
<p>Plot 1</p> 	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
<p>Plot 2</p> 	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

## 7 Activation functions and neural networks

Select all statements which are true.

### Select one or more alternatives

- ☒ The ReLU activation usually works better than sigmoid activation function for hidden units because the sigmoid activations are sparse.
- ☐ Increasing the training set size generally does not hurt an algorithm's performance, and it may help significantly.
- ☒ In logistic regression, the weights  $w$  should be initialized randomly rather than to all zeros, because if you initialize to all zeros, logistic regression will fail to learn a useful decision boundary as it will fail to 'break symmetry'.
- ☒ The tanh activation usually works better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data better for the next layer.
- ☐ Increasing the size of a neural network generally does not hurt an algorithm's performance, and it may help significantly.

Maks poeng: 9

## 8 RNNs

a. The network learns where to “pay attention” by learning the values  $e < t, t' >$ , which are computed using a small neural network: We can't replace  $s < t - 1 >$  with  $s$  as an input to this neural network. This is because  $s$  depends on  $\alpha < t, t' >$  which in turn depends on  $e < t, t' >$  so at the time we need to evaluate this network, we haven't computed  $s$  yet.

### Select an alternative

- ☒ True
- ☐ False

b. You have to fill a blank in a sentence of a long passage: “.....Sam liked teddy as a leader.....”. There are many options for the same such as “bear” or “roosevelt” etc. You'd like to build a model to do it for you. Use of bi-RNN is not necessary for this purpose.

### Select an alternative

- ☐ False
- ☒ True

Maks poeng: 6

## 9 Clustering 1

Answer the following questions appropriately.

What is the minimum numbers of variables or features required to perform clustering? 1

K means and K-mediods are example of which type of clustering method?

Partition (Spectral, Hierarchical, Probabilistic, Partition)

Which of the following are considered as unsupervised learning?

**Select one or more alternatives**

- ☒ Clustering
- ☐ Dimensionality reduction
- ☐ Decision trees
- ☐ Logistic regression
- ☐ None of these

Maks poeng: 12

## 10 Clustering 2

Consider the following statements pertaining to K-Means algorithm

a. Once an example has been assigned to a particular centroid, it will never be reassigned to another centroid.

**Select an alternative**

- ☐ False
- ☒ True

A good way to initialize K-means is to select K (distinct) examples from the training set and set them as cluster centroids.

**Select an alternative**

- ☒ True
- ☐ False

On every iteration of K-means, the cost function  $J(c(1), \dots, c(m), \mu_1, \dots, \mu_k)$  (the distortion function) either stays the same or decreases; in particular, it should never increase.

**Select an alternative**

- ☐ True
- ☒ False

K-Means will always give the same results regardless of the initialization of the centroids.

**Select an alternative**

- ☐ True
- ☒ False

For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide.

**Select an alternative**

- ☐ False
- ☒ True

If we are worried about K-means getting stuck in bad local optima, one way to reduce this problem is if we try using multiple random initializations.

**Select an alternative**

- ☒ True
- ☐ False

Maks poeng: 12

**11 Provide your explanation for numerical computations here!**

You can use this space to provide explanation for your answers!

**Fill in your answer here**

Numerical computations are a way to automate real life problems by implementing computer algorithms and mathematics such as linear algebra and calculus

Maks poeng: 0