UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK-INF3000/4000 — Selected topics in data science
Day of examination:	June 13, 2016
Examination hours:	14.30 – 18.30
This problem set consists of 6 pages.	
Appendices:	None
Permitted aids:	Calculator

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Contents

Note: All of the questions below are multiple choice. One or more answers are correct. Each question is worth one point, which is gained if and only if all correct answers are checked *and* none of the wrong answers are checked.

Problem 1 Python

a Lists

What will be the output of the python code listed below?

```
print range(10)[1:5]
```

- ❑ `[1, 2, 3, 4, 5]`
- ❑ `[0, 1, 2, 3, 4, 5]`
- ❑ `[1, 2, 3, 4]`
- ❑ `[0, 1, 2, 3, 4]`

b Lists

What will be the output of the python code listed below?

```
print ['a', 'b', 'c'].index('b')
```

- ❑ `b`
- ❑ `1`
- ❑ `0`

(Continued on page 2.)

c Dictionaries

What will be the output of the python code listed below?

```
print {i: i**2 for i in range(10) if i % 2}
```

- ❏ `[1, 9, 25, 49, 81]`
- ❏ `{1: 1, 3: 9, 5: 25, 7: 49, 9: 81}`
- ❏ `{0: 0, 8: 64, 2: 4, 4: 16, 6: 36}`
- ❏ `{1: 1, 3: 9, 5: 25, 7: 49}`

Problem 2 Data Processing

a Data Frame Transformation

Given a `pandas` data frame `a`, given by

```
  category label  value
0     high     a      0
1      low     b      1
2     high     c      2
3      low     a      3
4     high     b      4
5      low     c      5
```

what command can be used to transform it into the following form?

```
category  high  low
label
a            0    3
b            4    1
c            2    5
```

- ❏ `a.stack`
- ❏ `a.groupby`
- ❏ `a.pivot`

b Data Frame Transformation

Given a `pandas` data frame `a`, given by

```
  category label  value
0     high     a      0
1      low     b      1
2     high     c      2
3      low     a      3
4     high     b      4
5      low     c      5
```

what command can be used to transform it into the following form?

(Continued on page 3.)

```
          value
category
high          6
low           9
```

- ❑ `a.stack`
- ❑ `a.groupby`
- ❑ `a.pivot`

Problem 3 Apache Spark

a Map and reduce

What output will the following Apache Spark code produce (a Spark Context is assumed to exist as `sc`).

```
print (sc
  .parallelize(range(10))
  .map(lambda x: [x % 2, x])
  .reduceByKey(lambda x, y: x + y)
  .collect())
```

- ❑ `[(0, 20), (1, 20)]`
- ❑ `[(0, 20), (1, 25)]`
- ❑ `[(0, 10), (1, 15), (3, 10)]`

b Filter

What output will the following Apache Spark code produce (a Spark Context is assumed to exist as `sc`).

```
print (sc
  .parallelize(range(10))
  .filter(lambda x: x % 2 == 0)
  .map(lambda x:  x**2)
  .reduce(lambda x, y: x + y))
```

- ❑ `120`
- ❑ `128`
- ❑ `[0, 4, 16, 36, 64]`
- ❑ `100`

Problem 4 Computers, Storage, and Communication

a Latencies

What is the correct ordering of data storage elements in a computer by latency for an access by the CPU, from fastest to slowest?

- ❑ Cache, RAM, Hard Disk
- ❑ RAM, Cache, Hard Disk
- ❑ Hard Disk, Cache Ram

(Continued on page 4.)

b MongoDB

What data format does the MongoDB database use internally?

- ❏ BSON, a binary version of JSON.
- ❏ ASCII strings.
- ❏ Python dictionaries.

c REST APIs

Which of the following code fragments can be used to interact with a REST API in Python?

- ❏ `open('https://api.company.com/clients.json', 'https').readlines()`
- ❏
```
import requests
data = requests.get('https://api.company.com/clients.json').json()
```
- ❏ `open('https://api.company.com/clients.json').get()`

Problem 5 Machine learning

a Decision trees

Which of the following statements about decision trees are correct?

- ❏ Trees of very low depth tend to show high bias.
- ❏ Variable importance can be estimated from the splitting points chosen by the algorithm and resulting information gain.
- ❏ Trees of excessive depth tend to show high variance.

b Decision trees

Which methods are available to help towards overcoming the tendency of trees to over-fit the data?

- ❏ Splitting the available data in train and test sets.
- ❏ Train the tree on less data.
- ❏ Pruning.
- ❏ Using ensembles of tress in random forests.

c Gradient boosting

Which loss functions can be used with the gradient boosting algorithm?

- ❏ Huber Loss.
- ❏ Exponential loss.
- ❏ Any sensible loss function with well defined gradient.

d Cross-validation

Which of the following statements about k-fold cross-validation are correct?

- ❏ k-fold cross-validation helps controlling correlations between variables.
- ❏ k-fold cross-validation is an effective method to identify issues related to over-fitting.
- ❏ The choice of k must be made with the available data quantity in mind.

(Continued on page 5.)

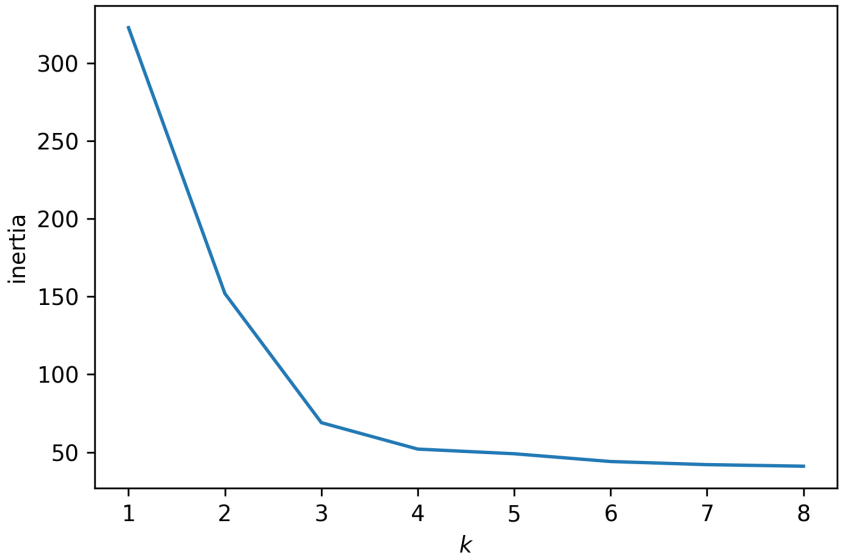
e Variable Selection

Which of the methods listed below can be used for eliminating variables in a machine learning model?

- ❏ The lasso method (for regression models).
- ❏ Ridge regression (for regression models).
- ❏ Train-test split.
- ❏ Forward step-wise variable selection.

f K-Means

The elbow method is used to determine the optimal numbers of clusters k in a data set. The resulting plot looks as follows.



What will the optimal number of clusters be?

- ❏ k cannot safely be determined from the plot.
- ❏ 3
- ❏ 4
- ❏ 1

g Logistic Regression

You have fitted a logistic regression model that you use to understand the factors contributing to a person smoking or not. You have an indicator variable

$$X_{\text{male}} = \begin{cases} 1 & \text{if person is male} \\ 0 & \text{if person is female.} \end{cases}$$

The corresponding regression coefficient is $\beta_{\text{male}} = 0.21$. Which of the following statements holds true provided your model is valid and the coefficient has a low P-value?

(Continued on page 6.)

- ❏ The odds of a female smoking are twice as high as for a male.
- ❏ The odds of male smoking are 23% higher than for a female.
- ❏ The odds of male smoking are 21% higher than for a female.
- ❏ The odds of female smoking are 21% higher than for a male.

h Classification

Which of the following problems should be solved using a classification algorithm?

- ❏ Grouping similar candidates for a job together.
- ❏ Predicting if an email is spam or not.
- ❏ Predicting the price of an item.
- ❏ Predicting the species of a bird.

i Classification

Which of the following algorithms can be safely used for classification problems?

- ❏ Linear regression.
- ❏ Decision trees.
- ❏ Agglomerative clustering.
- ❏ Linear discriminant analysis.

j Anomaly Detection

Can classification be used for anomaly detection?

- ❏ Yes, a classification algorithm should be used to classify an observation as anomalous.
- ❏ No, due to the skew in data towards the non-anomalous labels one cannot use classification algorithms.
- ❏ It is often not advisable to use classification algorithms to detect anomalous behavior but can, if enough data is available and correct class weights are used, be a useful method.