

Problem 1a and 1b

1)Sentiment Analysis

Given an input in text- or audio-format, the goal is to identify the sentiment of the author, e.g. whether he's angry or happy.

The inputs are content in written text or spoken audio format, possibly together with meta-information such as author data (age, profession) or information concerning the retrieval of the data itself like the platform / channel it was obtained from (Twitter, Blogs, Survey) or date/time it was recorded.

The label is one of many (typically 2: good / bad) predefined states the subject may be in, such as angry, scared, happy, adventurous.

Intuitively, humans are good at judging content like choice of words, body-language or intonation to a person's sentiment, but hard-coded rules are much more difficult for computers to generate and adapt. The aspect of typical human interaction and inferring sentiment from a given situation is hard for computers as they do not natively understand "typical" human behavior.

2)Author Influence

Given (social media) data of multiple sources structured by author / publisher of the information, the goal is to identify key figures of influence in a given network of information.

The inputs are data metrics to published articles, research, appearing and connections to an identities public appearance. This may also take into account meta-information like covariance to other identities and their metrics or association with a predefined group of typically influential identities.

The labels are typically a numeric value (hence appliance of regression rather than classification) indicating overall influence. This number may also be broken down to influence on a specified sub-group divided by age / interests / marketing potential /

Author influence is a rather complex topic for us humans, too. We generally understand which figure has a major influence on which area of interest, like which musician will reach most people may be easily deduced e.g. by how frequently said musician appears on local radio. But braking it down to an accurate ranking amongst many individuals concerning a multitude of fields to estimate influence upon is a process of highest-order number crunching and statistics, at which a computer naturally excels.

3) Malicious Network Traffic Detection

Machine Learning is also used for security purposes. As such, it was e.g. used for learning what kind of network traffic was to be expected by a given to-be-supervised network. In the case of an initiated attack on the network or immediate leaking of data from within, this could be then detected as an abnormal behavior to proclaim that "the network's security was compromised".

For input, the data traffic (size, frequency, table of recipients, ...) is monitored and metrics learned and foreshadowed.

The labels would be a numeric estimation of "normal behavior", to which a threshold should be applied to create some buffer for "different, yet still ordinary" behavior. As the numeric estimation regresses to unexpected values, alarm is being thrown.

This all may also incorporate associate rule learning to draw direct connections to which sequence of connected actions should result in alarms being thrown and which ones should not.

Detecting malicious behavior of a network involves keeping track of a multitude of logs, network traffic, health-checks and more. A system-administrator knows a plethora of tools and software to help keeping track of all of this information, but an automated system (if it is able to properly map all dependent systems) can more easily draw non-intuitive connections and is needed to cover the ever-growing complexity and scope of systems at hand.

Problem 1c

1) Insiders

It was founded in 1998, has around 100 employees and its headquarter in Kaiserslautern (with another seat in Berlin).

[Homepage](<https://www.insiders-technologies.de/home.html>)

Insiders Technology specializes on AI for understanding documents. Intelligent software solutions are applied to business processes to optimize business insights and streamline reporting. Its overall goal is to orchestrate a business's digital transformation with respect to automation, customer relation and general information management.

Its offerings include *Input Management* (digital mailroom, billing service, order management), *Customer Communication* (Customer engagement platform, response management, process transparency), *Mobile* (chat / capture / info / contracts / services applications), *Cloud* (IT-Infrastructure offerings, scalable (Web-)services) and *Business Intelligence* (engaging business processes, service quality assurance, forecasting; connected to the *digital mailroom*-offering).

2) Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Seated at Kaiserslautern, Saarbrücken (Headquarter), Bremen, Berlin (project bureau) and with an associated branch in Osnabrück, the DFKI presents itself as a "Center of Excellence" and one of the biggest research centers surrounding the field of Artificial Intelligence. It was founded in 1988 and is a non-profit with around 300 employees.

The DFKI focuses on AI research and appliance of related technology. To name some specific focus points, it structures its areas of research, research groups and living labs as follows: Smart Data, Cyber-Physical systems, multilingual technology, plan-based robot control, educational tech., interactive fabric, innovative retail, business informatics, embedded intelligence, smart service engineering, intelligent large-scale analytics, innovative industrial systems, intelligent networks, augmented reality, speech technology and intelligent interfaces.

It also focuses on a tight incorporation of the respective location's focus in research to draw local expertise into the equation.

Problem 1d

Though not having directly visited one of the described companies, there was the opportunity to talk with an employee of DFKI as well as listen to a talk of an DFKI worker. Both gave the general impression of the institute promoting personally motivated efforts for focus and work in research.

Many projects are inspired by real-world problems and specific application of ML and Deep Learning techniques. The direction the institute is heading into is a modern and well-defined one of leveraging given techniques and making them utilizable.

Problem 2

We have used SVM module from the sklearn library to train the classification module, and we achieved an accuracy of .939 on Kaggle's test data. Below is the source code of our solution:

```
digits_recognizer.py
1 import csv
2 from sklearn import svm
3
4 #global lists
5 trainLabels = []
6 trainImages = []
7 testImages = []
8
9 #read train data csv file
10 with open('train.csv','r') as trainDataFile:
11     trainDataReader = csv.reader(trainDataFile)
12     for row in trainDataReader:
13         trainLabels.append(row[0])
14         # convert images to binary format (0,1) if the pixel has value != 0 it will be just 1 to simplify classification
15         #idea taken from this source : https://www.kaggle.com/archaeocharlie/a-beginner-s-approach-to-classification
16         trainImages.append([1 if pixelValue != '0' else 0 for pixelValue in row[1:]])
17
18
19 #read test data, the same way as train data
20 with open('test.csv','r') as testDataFile:
21     testDataReader = csv.reader(testDataFile)
22     for row in testDataReader:
23         testImages.append([1 if pixelValue != '0' else 0 for pixelValue in row[0:] ])
24
25 #delete the first element as its the file header not the data
26 del trainLabels[0]
27 del trainImages[0]
28 del testImages[0]
29
30 # train SVM module with the extracted train (labels, images)
31 svmClassifier = svm.SVC()
32 svmClassifier.fit([trainImages,trainLabels])
33
34 #run the classifier against the test data and get predicted labels
35 predictedLabels = svmClassifier.predict(testImages)
36
37 #write results to csv file
38 with open('output.csv', 'w') as outputDataFile:
39     csvWriter = csv.DictWriter(outputDataFile, fieldnames = ['ImageId','Label'])
40     csvWriter.writeheader()
41     imageId = 1
42     for label in predictedLabels:
43         csvWriter.writerow({'ImageId' : imageId,'Label' : label})
44         imageId += 1
45
46
```