EXAMINATION QUESTIONS

Faculty: **Science and Technology**

Examination in: **DAT200** **Applied Machine Learning**
Course code *Course name*

Time for exams: Monday, 28.05.2018 9:00 – 12:30 (3.5 hours)
Day and date *As from – to and duration of examinations (hours)*

Course responsible: Kristian Hovde Liland (6723 1624) and Oliver Tomic
Name

Permissible aids:

A1: no calculator, no other aids

The exams papers includes: 9
Number of pages incl. attachment

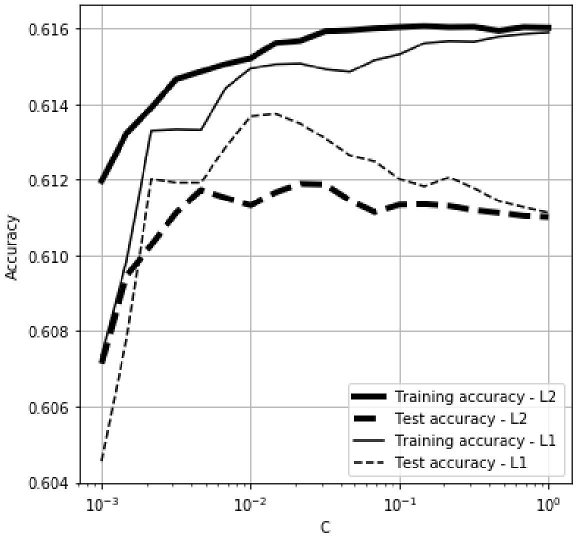
If the examination consists of several parts, information must be given as to how much each part will count toward the grade

Course responsible: Kristian Hovde Liland and Oliver Tomic

External examiner: Bjørn-Helge Mevik

Exercise 1 (10 points in total)

When analysing a data set, two Logistic Regression models have been fitted. L2 and L1 norm regularization have been applied. Mean accuracy results for repeated training-test splits are reported in the following figure plotted against the inverse regularization parameter.



a) (5 points)
Which regularization would you choose for future predictions on new data? Approximately, which corresponding C-value would you choose?

b) (5 points)
What are the L2 and L1 regularizing? Which of these can lead to variable selection during regularization?

Exercise 2 (10 points in total)

A classification problem with three classes leads to the following one-versus-all confusion matrices.

True	21	4
	2	13
	Predicted	

True	29	6
	1	4
	Predicted	

True	15	5
	6	14
	Predicted	

a) (2 points)
Sketch a 2 x 2 confusion matrix and fill it with TP, TN, FP and FN, assuming the first (top) class is negative and the second (bottom) class is positive.

b) (4 points)
Given the formulas for precision = TP/(TP+FP) and recall = TP/(FN+TP), how would you explain the interpretation of precision and recall based on the confusion matrix you sketched to someone not familiar with machine learning? (Only explain the type of performance that is measured.)

c) (4 points)
How would you compute precision and recall for the combined three-class classification? Show the elements of the calculations and explain your choice.

Exercise 3 (15 points in total)

The following algorithm is the AdaBoost in random order.

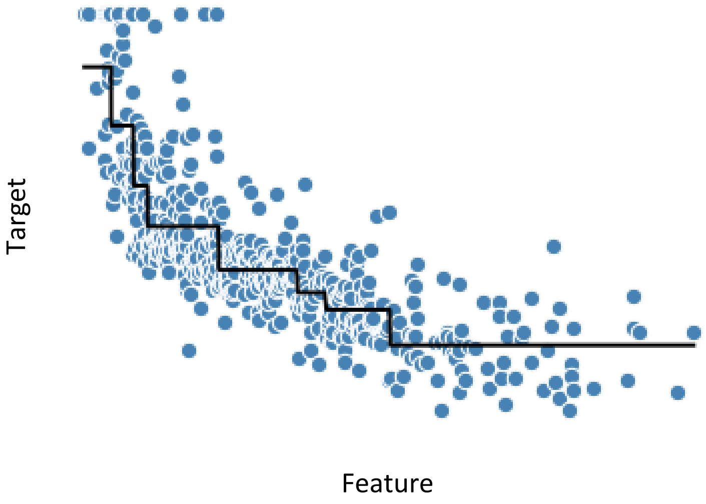
A	For j in m boosting rounds, do the following:
B	Predict class labels: $\hat{y} = \text{predict}\left(C_j, \boldsymbol{X}\right)$.
C	Set the weight vector **w** to uniform weights, where $\sum_i w_i = 1$.
D	Compute coefficient: $\alpha_j = 0.5 \log \frac{1-\varepsilon}{\varepsilon}$.
E	Compute the final prediction: $\hat{\boldsymbol{y}} = \left(\sum_{j=1}^{m}\left(\boldsymbol{\alpha}_j \times \text{predict}\left(C_j, \boldsymbol{X}\right)\right) > 0\right)$.
F	Normalize weights to sum to 1: $\boldsymbol{w} := \boldsymbol{w} / \sum_i w_i$.
G	Update weights: $\boldsymbol{w} := \boldsymbol{w} \times \exp\left(-\alpha_j \times \hat{\boldsymbol{y}} \times \boldsymbol{y}\right)$.
H	Train a weighted weak learner: $C_j = \text{train}\left(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}\right)$.
I	Compute weighted error rate: $\varepsilon = \boldsymbol{w} \cdot \left(\hat{\boldsymbol{y}} \neq \boldsymbol{y}\right)$.

a) (10 points)
Reorder the lines correctly. (The correct sequence of line letters is enough).

b) (5 points)
Comment briefly on what each line does (short and concise, preferably one or two sentences per line).

Exercise 4 (12 points in total)

A decision tree has been used to fit the jagged curve going through the points in the following plot.



a) (6 points)
Describe briefly how the curve is made. Focus on the decision process of splitting the feature. What is optimised in the steps of this process?

b) (6 points)
Explain how Random Forests generalizes this decision tree. What makes Random Forests a more likely candidate for robust and general predictions in this type of problem?

(Please limit descriptions to less than a page in total for Exercise 4.)

Exercise 5 (8 points in total)

After a clustering algorithm has been applied to six data points, the following two clusters are found.

	Feature 1	Feature2
Obj. 1	1	1
Obj. 2	1	2
Obj. 3	2	2

	Feature 1	Feature2
Obj. 4	2	3
Obj. 5	3	3
Obj. 6	3	4

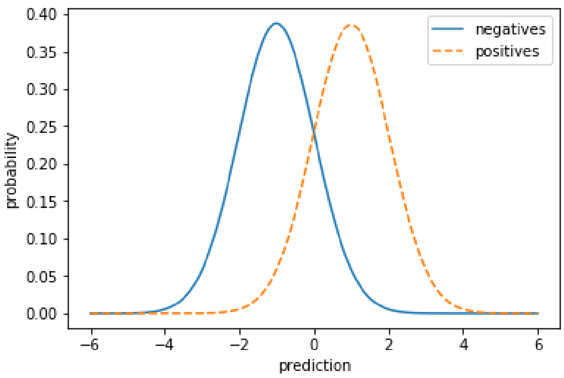
Manhattan/Cityblock/L1/"sum absolute" distance matrix:

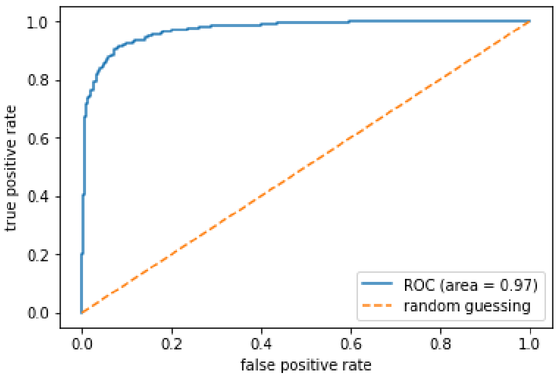
Objects	1	2	3	4	5	6
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

Compute the silhouette values of all observations and sketch the resulting silhouette plot.

Exercise 6 (15 points in total)

Assume that a classification has been performed where the class decision is made for predictions < 0 (class A) or ≥ 0 (class B). The corresponding class probabilities are shown in the left hand plot and the Receiver Operating Characteristic curve in the right hand plot.





a) (10 points)
Explain the basics of the Receiver Operating Characteristic. What happens as you trace the ROC curve from lower left to upper right? (You can use the prediction densities to aid your interpretation.) What does the ROC look like for a perfect classifier? What does it look like for a classifier that is worse than random guessing?

b) (5 points)
Imagine a ROC curve only slightly higher than random guessing. What kind of technique could be used to try to leverage such a low performing model?

Exercise 7 (15 points in total)

We want to compare three basic classifiers: the Perceptron, Adaline and Logistic Regression, but the teacher has been sloppy, mixing his cost functions and activation functions.

$$\boxed{\phi\ (z) = z}$$

$$\boxed{\phi\ (z) = \begin{cases} 1\ if\ z \geq 0 \\ -1\ otherwise \end{cases}}$$

$$\boxed{J(\boldsymbol{w}) = \sum_{i=1}^{n} \left[-y^{(i)} \log\left(\phi\left(z^{(i)}\right)\right) - (1 - y^{(i)}) \log\left(1 - \phi\left(z^{(i)}\right)\right)\right]}$$

$$\boxed{J(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \phi(z^{(i)}))^2}$$

$$\boxed{\phi(z) = \frac{1}{1 + e^{-z}}}$$

a) (5 points)
Help the teacher associate the correct cost functions and activation functions to the three classifiers.

b) (5 points)
Sketch the activation functions and report their threshold functions (the regions leading to different $\hat{y}$ values).

c) (5 points)
Sketch a hypothetical cost function curve and explain briefly how gradient decent searches for the minimum.

Exercise 8 (15 points in total)

Using the code below, describe briefly the role of each step of the analysis being performed and precisely what the choice of parameters in each line means (1 point for step 1, 2 points for each of the other steps).

```
# Step 1:
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.model_selection import GridSearchCV

# Step 2:
df = pd.read_csv('https://espionage.no/top_secret/surveilance'
                 '/foreign_affairs/spy_list.data')

# Step 3:
X = pd.concat([df.iloc[:, 2:10], \
               pd.get_dummies(df['spy_network'],\
               drop_first=True)], axis=1).values
y = df.iloc[:, 0].values

# Step 4:
X_train, X_test, y_train, y_test = \
    train_test_split(X, y,
                     test_size=0.30,
                     stratify=y,
                     random_state=1)

# Step 5:
pipe_svc = make_pipeline(StandardScaler(),
                         SVC(random_state=1))

# Step 6:
param_range_C = \
    [0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0]
param_range_g = [0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0]
param_grid = [{'svc__C': param_range_C,
               'svc__gamma': param_range_g,
               'svc__kernel': ['rbf']}]

# Step 7:
gs = GridSearchCV(estimator=pipe_svc,
                  param_grid=param_grid,
                  cv=10,
                  n_jobs=1)
gs = gs.fit(X_train, y_train)

#Step 8:
result = gs.best_estimator_.score(X_test, y_test)
```