Solutions for Exam in Statistical Machine Learning
Statistisk Maskininlärning (1RT700)

Date and time: January 11, 2021, 08.00–13.00

Responsible teacher: David Sumpter

1. (a)

```
import numpy as np
x = np.array([6, 10, 8, 11, 7])
y = np.array([1, -1, 1, -1, 1])
beta0=9
beta1=-1
adj_x= x*beta1 + beta0
l=np.exp(adj_x*y)/(1+np.exp(adj_x*y))
print(l)
likelihood=np.prod(l)
loglikelihood=np.sum(log(l))
print(loglikelihood)

[0.95257413 0.73105858 0.73105858 0.88079708 0.88079708]
-0.9289667486961326
```

The log-likelihood is -0.9289.

(b) Provided the classifier assigns 1 to any $x \leq 8$ and -1 to any $x \geq 10$ then it is perfect. We need to find

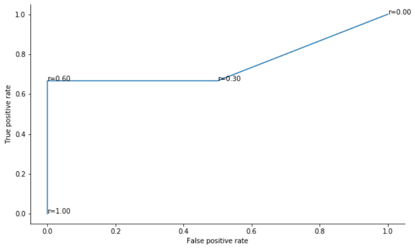
```
upper_r =1/(1+np.exp(10*beta1+beta0))
lower_r =1/(1+np.exp(8*beta1+beta0))
```

which gives $0.2689 \leq r \leq 0.7311$

(c)

```
----------------------------
Confusion matrix for r=0.3:
TP = 2 | FP = 1
FN = 1 | TN = 1
FP rate: 0.50
TP rate: 0.67
Misclassification error: 0.40
----------------------------
----------------------------
Confusion matrix for r=0.6:
TP = 2 | FP = 0
FN = 1 | TN = 2
FP rate: 0.00
TP rate: 0.67
Misclassification error: 0.20
----------------------------
```

(d) The ROC is:



(e) In this one dimensional example the separation is done using a single point, x_* such that $f(x_*) = r$. We solve

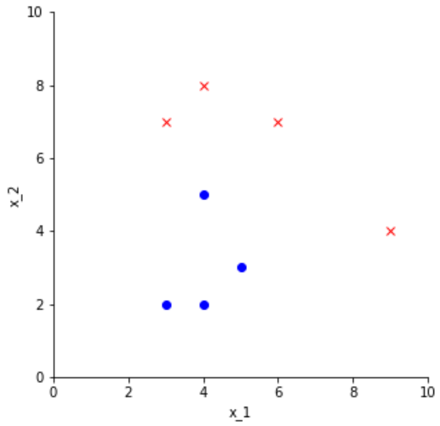
$$\frac{\exp(\beta_0 + \beta_1 x_*)}{1 + \exp(\beta_0 + \beta_1 x_*)} = r$$

which is

$$\log(\frac{1 - r}{r}) = -\beta_0 - \beta_1 x_*$$

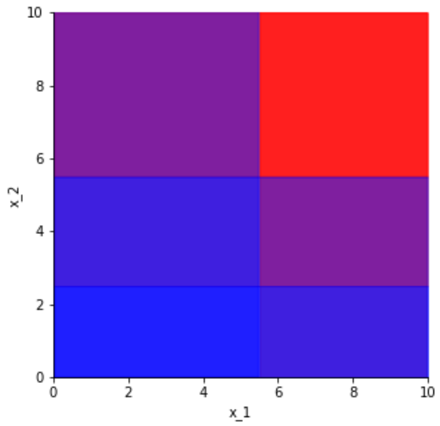
which is a linear function of x_*. So it is a linear classifier.

2. (a) The data looks like this:

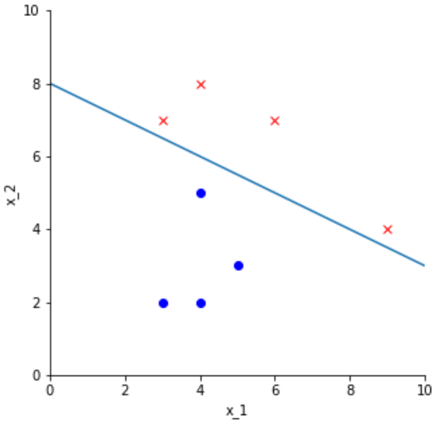


(b) The missclassification of dataset 1 is minimised by a split at $x_1 = 5.5$. The missclassification of dataset 2 is minimised by a split at $x_2 = 6$. For dataset 3, the missclasification is minimized at $x_2 = 2.5$ (or nearby), where it is 1. Looking either side, for example, $x_2 = 4.5$ or $x_2 = 1.5$ gives worse results. In answering the question we don't need to calculate all the entropy values. It is enough to explain that (for example) that we can calculate the entropy values for the cases where one value is incorrectly classified.

(c) In the plot below, red and purple areas are classified as crosses.



(d) Yes. By drawing a line from, for example, (0,8) to (10,3)


10
8
6
4
2
0
x_2
0
2
4
6
8
10
x_1

3. (a) For k-NN algorithm the training error is zero for one nearest neighbour, and typically increases as we add mor neighbours, so the left figure is most appropriate. For k-fold cross validation, a single batch (if all used for training data) will have largest training error because most data points used and training error is summed (not averaged over points); $k = 2$ will have least training error; and then the training error will increase with k. So centre figure is most appropriate. For logistic regression, training error decreases with k, so the right graph is most likely.

   (b) For k-NN test error will decrease with k initially and then might go up. Same for logistic regression. So both of these could be the left figure. The test error for new data set should not depend on k for k-fold cross validation. The validation is instead used to estimate the test error. So the right figure is most appropriate.

   (c) See lecture notes. (reduced correlation)

   (d) See lecture notes.

4. (a) The output initially increases with increasing x_1 and x_2 but then decreases if these variables become large.

   (b) It is larger. The first model minimised the sum of squares. The regularised fitting is a worse fit and has greater sum of squares.

   (c) From lecture notes. k-fold cross validation with different λ values.

   (d) The x_1 variable now takes the role of the squared variables in predicting decreases for larger x_1.

   (e) Colleague A is cheating a bit here! She fit her model on **all** the data, including the 90% which B saved for testing. So colleague B correct.

   (f) Colleague C might be interested in how public transport (and other public health measures) effects spread of disease. This is something that can be controlled later by authorities. Calls about the disease is necessarily correlated with cases.

5. (a) Here are two examples of the types of machine learning problems can be formulated here.

1, Use training data on people who have completed a data science masters, for example, looking at their Education (before masters), age, location, interests key words. In particular data *prior* to taking masters is most useful, since this is what the target group will later have. The output variable should idealy be: did that person take a similar course.

2, Identify people who are particularly active on LinkedIn. Find the properties (level of activity, education, age, location, interests, friends etc.) of people who are currently searching for ?data science masters? vs. those that aren?t. These can be used to predict other people who might be interested too.

(b) Here there many things about how we worked in the mini-project that can be described, in terms of methods and cross-validation. The work should be divided across the data scientists, with each one independently developing their own methods. Data pre-processing is important in this question and understanding features. Is the problem even feasible? This should be checked. Emphasis on collaboration between team members, as well as discussions and comparisons about how methods are working.

(c) Non-Swedes classification accuracy: TP+TN / Total = 400 / 600 = 4/6 = 2/3
Swedes classification accuracy: TP+TN / Total = 800 / 1200 = 8/12 = 2/3

So the accuracy is the same.

Before we put the algorithm into production we also want to check the false negative rate (FNR) and false positive rate (FPR) that can reveal bias.

Non-Swedes:

FNR: FN / FN + TP = 100 / (100+100) = 1/2

FPR: FP / FP + TN = 100 / (100 + 300) = 1/4

Swedes:

FNR: FN / FN + TP = 50 / (50 + 400) = 1/9

FPR: FP / FP + TN = 350 / (350 + 400) = 7/15

We can also calculate TPR = 1- FNR to answer the question. But we are primarily interested in people who didn?t get to see the advert and might be interested in coming to Uppsala (FPR). Here there is clear bias against non-Swedes. The FPR can be mentioned, but is less relevant to the question.

To get full marks, the answer should also include a reasoned answer about whether it should be put in to production. As we saw in the lecture, it is impossible to eliminate all forms of bias. So, the bias in the FNR rate needs to be balanced against the fact it has equal accuracy. An answer that "it is biased" or "it is not biased" is not sufficient here.