1. Selecting machine learning approach

Netflix recently reported that they had lost a large number of customers. Churn is a name for cancelling a subscription. Netflix wants to find out the causes for churn and decides to use machine learning models to find out what it is that characterize the churn customers. What machine learning approach would be most natural to use for this problem?

a) An artificial neural network for classification
b) A support vector machine for regression
c) A principal component algorithm
d) A k-means clustering algorithm

2. 3-NN – categorical inputs

You are running a 3-NN algorithm with categorical features, one class and one numerical target. You use the Hamming distance to measure distance between data points. You want to classify a new data point with the following features

(f1 = a, f2 = b, f3 = a, f4 = c)

The 5 data points in the training set has values

(f1 = a, f2 = a, f3 = a, f4 = c, class = G, y = 4.3)
(f1 = a, f2 = b, f3 = c, f4 = a, class = H, y = 4.6)
(f1 = c, f2 = a, f3 = b, f4 = a, class = G, y = 5.2)
(f1 = a, f2 = c, f3 = a, f4 = c, class = H, y = 4.9)
(f1 = a, f2 = c, f3 = b, f4 = a, class = G, y = 4.0)

The data is used for classification and for prediction of y. What are the predictions for “class” and “y”?

a) H and 4.6
b) G and 4.9
c) H and 4.5
d) G and 4.6

3. Accuracy

You have a test set with 1000 data points classified as Cats or Dogs. Your Convolutional Neural Network model classifies example 611 as Dog with a probability of 0.875, but it should have been Cat. How much does data point 611 contribute to the accuracy score of the model?

a) 0.001
b) 0.125
c) 0.875
d) 0.000 *

4. Running a Naïve Bayes model

You have learned a model for categorical Naïve Bayes prediction. The target value is in the set {K,L,M,N} with learned a priori probabilities p(K) = 0.3, p(L) = 0.5, p(M) = 0.1, p(N) = 0.1. A and B are binary feature values with value in {T,F}.

You have learned the following probability tables from the training data:

	Class = K	Class = L	Class = M	Class = N

A = T	0.6	0.3	0.2	0.4
A = F	0.4	0.7	0.8	0.6

	Class = K	Class = L	Class = M	Class = N
B = T	0.4	0.2	0.9	0.3
B = F	0.6	0.8	0.1	0.7

Now you want to classify the data points (A = T, B= F). What is its classification?

a) K
b) L
c) M
d) N

5. Overfitting

You have run 4 machine learning models, a random forest model, a two-layered neural network, a support vector machine and a logistic regression model. The training and test accuracy scores are

Random Forest: training: 0.863 test: 0.902
Neural network: training: 0.901 test: 0.911
Support Vector Machine: training: 0.842 test: 0.730
Logistic regression: training: 0.765 test:0.744

Which of these models is the most overfitted?

a) Random Forest
b) Neural Network
c) Support Vector Machine
d) Logistic regression

6. Linear regression

You have a data set with 1275 samples with 12 input continuous features and one continuous target feature. You want to use this to learn a simple linear regression model. How many parameters does your algorithm learn?

a) 13
b) 12
c) 1275
d) 1276

7. Reducing input data size

You have a data set with 364 input features. For explainability purposes you would like to reduce this number of features when building a linear regression model. What approach would support explainability?

a) Use LASSO to learn the regression model and remove those features that have zero weight
b) Use Random Forest to identify which features should be excluded

c) Use Principal Component Analysis as dimension reduction strategy to extract meaningful new features

d) Use t-SNE to visualize the regression model for the user

8. Selecting a classifier

The data set you have consists of 2 continuous features, and you want to do binary classification (T,F). When you assess the distribution for each feature, you see that the instances that are classified as T are such that for each feature F_i, (i=1 or i=2) the values of the feature has a tendency to be in an interval [low_i, $high_i$] with higher density in the middle of the interval. The instances classified as F for more than 90% have feature values outside the interval [low_i, $high_i$]. Which of these methods are likely to provide a model with highest accuracy?

a) Kerneled Support Vector Classifier

b) Logistic Regression

c) Naïve Bayes classifier

d) Linear Support Vector Classifier

9. Random Forest regression

You aim at running a random forest regression algorithm with 1000 decision trees. Each of them returns a value prediction. How would you compute the final output from the model based on these values?

a) Voting

b) Averaging

c) Computing R^2

d) Computing accuracy

10. The ReLU activation function

A neural network unit has two inputs x[1] = 0.52, x[2] = 0.74, weights w[1] = -0.33, w[2] = 0.55 and b = 0.21. We run the standard linear computation and use a ReLU activation function. What will be the output from the unit

a) 0.445

b) -0.445

c) 0.000

d) 0.235

11. Number of parameters

You have a multilayered standard neural network with 256 input values, one hidden layer with 10 units (neurons) and 3 units in the output layer. How many parameters will need to be learned in such a neural network (no regularization)?

a) 2603

b) 2570

c) 2590

d) 2593

12. Neural network learning

Your co-student Monald Frump is quite confident in himself and has a data set with 36,000 instances and 256 input features, and a binary output feature. He thinks a standard neural network with two hidden layers and sigmoid activation functions will do the job, so he initially splits the data set into a training set and a test set. He starts with a network with 100 and 15 units in the two hidden layers respectively, but notices that the test sets scores are not so good, (training set accuracy: 0.84, test set accuracy: 0.61). He reduces the number of units in each layer to 90 and 12, gets some improvement, and repeats the process until he has found a good test score at 60 and 6 units respectively (training set accuracy: 0.85 test set accuracy: 0.83). His worst enemy, Zoe Liden, comes up with thousands of new data for the same task. Monald runs his model on this data set, but has to realise that on this new set his model scores only with an accuracy of 0.71. What should Frump do now as his next step?

a) Redo the learning process using a validation set.
b) Redo the learning process testing out alternative activation functions like ReLU and tanh.
c) Redo the learning process but include an extra hidden layer in the model.
d) Redo the learning process but including all the new data in the training set.

13. Convolutional neural networks and feature maps

You have made a constructed a convolutional neural network, and in a convolutional layer you want to have 10 feature maps on receptive fields of size 5 x 5. How many parameters will you need to learn for this layer?

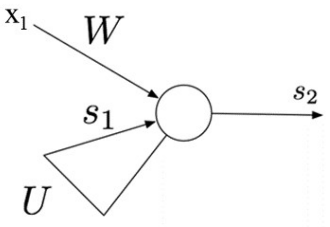
a) 260
b) 25
c) 500
d) 35

14. Choosing a good ML model for smart cities

You have data from about automatic 100 traffic counters around Bergen, which are reported each minute. You also have weather reports for each hour: wind, temperature, precipitation, pressure. In addition, you have hourly measures of NOx pollution on several sites in the city. You know that pollution levels are really dependent on longer trends in traffic and weather conditions rather than just spikes in traffic. You also have hourly weather forecasts for the next 24 hours. You decide on how to preprocess the data and select a model for predicting the pollution levels for the next 24 hours. What kind of model would fit best?

a) A recurrent neural network since they work well with temporal data
b) A support vector machine since they work well with non-linear data
c) A convolutional neural network since they work well with spatial data
d) A gradient boosting model since they generalize well on complex data

15. Recurrent neuron



$$S_2 = \tanh(W^T x_1 + U^T s_1)$$

A standard recurrent neuron in an RNN receives 4 values in [0,1] from the input layer as a vector x_i. It also has a state with 3 values in [0,1] as a vector s_i. How many weights do we need to learn for this neuron? Include bias weights in both U and W.

a) 27
b) 12
c) 7
d) 35

16. Confusion matrix

You are in a business where you need to take good care of your customers and develop a machine learning model that indicates whether a customer is a good candidate for a new marketing campaign (this is the positive case). But you know that many of your most loyal customers hate this kind of marketing and may actually leave your business if you run campaigns like this towards them. Which values in a confusion matrix would you want to reduce to ensure that your loyal customers are not disturbed?

a) True positives
b) True negatives
c) False positives
d) False negatives

17. Assessment of model

You have a data set about a rare disease where to variations need different medication. You have only 34 data points and want to estimate how good your logistic regression model can become. What kind of technique among these would you use to estimate how good logistic regression will perform on this data set?

a) Stratified 5-fold cross-validation
b) Leave one out cross-validation
c) Non-stratified 10-fold cross-validation
d) Random grid search

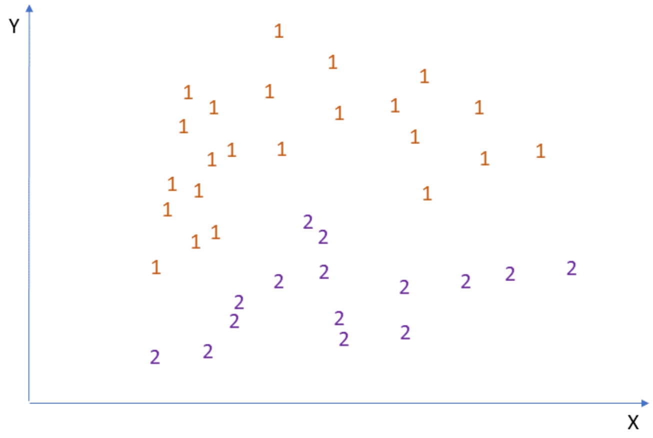
18. Dimension Reduction

The data you have has 212 features, all of them with values [0.0,10.0], and the target value is a number in the interval [0.0,50.0]. You want to reduce your data set so that each data point is represented by exactly 10 different features, each feature indicating how much each data points fits a particular aspect of the data. What would you use?

a) PCA
b) NMF
c) t-SNE
d) k-Means

19. Clustering algorithm

Your friend did a clustering process on a data set with two features X and Y, and ended up with the clustering shown below. All the 1-s are data points in one cluster, and the 2-s are data points in the other cluster. Which algorithm have your friend most likely used?



a) DBSCAN
b) K-Means clustering
c) Agglomerative clustering
d) t-SNE

20. Q-learning

You have built a robot which has no knowledge about the environment it shall act in. Your robot starts up in state s3, and it has no knowledge about any of the other states of the domain and their possible actions. The q-algorithm is adapted so it adds new states and actions as they are encountered, and any new state-action pair will get an initial q-value of 1.0. For a state where you have not registered any actions, the best action is assumed to have a q-value of 1.0.

The robot now randomly selects action a2, gets a reward of 4, and moves to a new state s1 (having a new set of actions available). With a discount factor of 0.8 and a learning rate of 0.1, what is the q-value for the state-action-pair (s3,a2) after the reward has been received?

a) 1.38
b) 4.0
c) 1.29

d) 1.0