*What are the **advantages** of a nearest neighbour classifier?*

1

*What are the **disadvantages** of a nearest neighbour classifier?*

2

What is the most important concept in machine learning?

3

What are the three 'ingredients' of a machine learning algorithm?

4

What does this equation calculate?

$$a = \sum_{i=1}^{F} x_i w_i$$

5

What is the perceptron learning rule?

6

What does this equation calculate?

$$a = \frac{1}{1 + exp(-\sum_{i=1}^{d} w_i x_i)}$$

7

Decision trees are good at handling _____ *data but worse at handling* _____ *data.*

8

- Very computationally expensive for every classification
- Complexity depends on the number of dimensions

2

- Very accurate
- No learning process

1

The model, the error function and the learning algorithm.

4

Never assume that you have all the data.

3

$$newWeight = oldWeight + 0.1 \times (trueLabel - output) \times input$$

6

The activation of the perceptron

5

Decision trees are good at handling categorical data but worse at handling continuous data.

8

The activation of the perceptron for non-linearly separable data

7

What does this equation calculate?

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$$

9

The 'information' contained in a variable is called the _____.

10

Explain the process of cross validation.

11

What factors should affect our decision on the best value of k?

12

What is the ensemble approach to machine learning?

13

Briefly describe bootstrapping

14

On average, what is the percentage of data points that are left unselected?

15

Explain bagging

16

The 'information' contained in a variable is called the entropy.

10

The entropy of a variable X

9

1. *Accuracy*
2. *Training time and space complexity*
3. *Testing time and space complexity*
4. *Interpretability*

12

1. *Break the data evenly into N chunks*
2. *Leave one chunk out*
3. *Train on the remaining $N-1$ chunks*
4. *Test on the chunk you leave out*
5. *Repeat until all chunks have been used to test*
6. *Plot the average and error bars for the N chunks*

11

Bootstrapping is the process of generating multiple data sets from an original.

14

Select a class of models, fit multiple models to training data (called base learners), use the models as a committee to vote on testing data.

13

Generate m bootstraps and train a model on each one. When the testing data arrives a simple majority vote takes place.

16

36.8%

15

Explain boosting

17

What type of classifier models a classification rule directly and models the probability of class memberships based on input data?

18

What type of classifier makes a probabilistic model of data within each class?

19

What type of classifier uses probabilities to classify data?

20

What is the formula to work out $P(c|X')$ Where c is a class and X' is an example?

21

What is the formula to work out a Gaussian model?

22

What are the two data representation methods that we talk about in clustering analysis?

23

What is the formula to work out Minkowski distance.

24

A discriminative classifier

18

Get a data set, take a bootstrap and train a model on it. See which examples the model got wrong then upweight those 'hard' examples and downweight the 'easy' ones. Now go back to training a model, but now you have a weighted bootstrap.

17

A probabilistic classifier

20

A generative classifier

19

$$\frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{(x-\mu)^2}{2\times\sigma^2})$$

22

$$P(c|X') = [P(x_1|c)P(x_2|c)....P(x_n|c)]P(c)$$

Where x is a feature in the example.

21

$$d(x,y) = \sqrt[p]{(x_1-y_1)^p + (x_2-y_2)^p ... + (x_n-y_n)^p}$$

24

Data matrices and distance matrices.

23

What is the formula for Manhattan distance?

25

What is the formula for Euclidean distance?

26

What is the cosine measure equation

27

What is the formula for the distance between symmetric binary attributes?

28

What is the formula for the distance between asymmetric binary attributes?

29

Briefly explain the partitioning clustering approach

30

Briefly explain the hierarchical clustering approach

31

How does the K-means clustering algorithm work?

32

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 ... + (x_n - y_n)^2}$$

26

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + |x_n - y_n|$$

25

$$d(x,y) = \frac{b+c}{a+b+c+d}$$

28

$$\frac{x_1y_1 + ... + x_ny_n}{\sqrt{x_1^2 + ... + x_n^2}\sqrt{y_1^2 + ... + y_n^2}}$$

27

Various partitions are constructed and then evaluated by some criterion, r.g. minimizing the sum of square distance cost. Typical methods are k-means, k-medoids, CLARANS.

30

$$d(x,y) = \frac{b+c}{a+b+c}$$

29

1. *Pick K random seed points (from the data)*
2. *Assign each data object to the cluster with the nearest seed point*
3. *Compute the mean points of the clusters (aka centroids)*
4. *Go back to step 1, but the seed points are now the means calculated in step 3. Do this until the assigned sets don't change between iterations.*

32

Create a hierarchical decomposition of the set of data (or objects) using some criterion. Typical methods are Agglomerative, Diana, Agnes, BIRCH, ROCK.

31

What is the runtime of the K-means algorithm?

33

What are some problems with the K-means algorithm?

34

What are the two sequential strategies used to construct a tree of clusters?

35

Explain the single link method of measuring the distance between clusters

36

Explain the complete link method of measuring the distance between clusters

37

Explain the average method of measuring the distance between clusters

38

How does the agglomerative hierarchical algorithm work?

39

What are some weaknesses of the agglomerative approach?

40

- *The initial seed points can cause 'local optimum' clusters, which may not be representative of the whole data.*
- *If a cluster has a non-convex (i.e. concave) shape, then it won't be detected.*
- *Unable to classify categorical data (unless you modify the algorithm).*
- *It's hard to evaluate the performance of the algorithm,*

34

$O(tKn)$

Where t is the number of iterations
K is the number of clusters
n is the number of objects

33

Use the smallest distance between an element in one cluster and an element in another.

36

Agglomerative and divisive

35

Find the average distance between the points in the two clusters and use that.

38

Use the largest distance between an element in one cluster and an element in another.

37

- *It is very sensitive to noise.*
- *It is less efficient than k-means clustering, with a runtime of $O(n^2)$*

40

1. *Convert the object attributes into a distance matrix, and make each object a cluster (of size 1).*
2. *Merge the two closest clusters together.*
3. *Update the distance matrix to take into account the new cluster.*
4. *Repeat from step two until we've got the desired number of clusters.*

39

In order to calculate the F-ratio index we divide the [redacted] *variance by the* [redacted] *variance.*

41

What is the formula to calculate the F-ratio index?

42

$$F(m) = \frac{mSSW(m)}{SSB(m)} = \frac{m \sum_{i=1}^{m} \sum_{j=1}^{n_i} d^2(x_{ij}, c_i)}{\sum_{i=1}^{m} n_i d^2(c_i, c)}$$

43

What is additive smoothing used for? Give an equation for it.

44

$$F(m) = \frac{mSSW(m)}{SSB(m)} = \frac{m \sum_{i=1}^{m} \sum_{j=1}^{n_i} d^2(x_{ij}, c_i)}{\sum_{i=1}^{m} n_i d^2(c_i, c)}$$

In order to calculate the F-ratio index we divide the intra-cluster variance by the inter-cluster variance.

When a $P(x|\omega_i)$ has a probability of 0 (it's 0 for just class i , additive smoothing can ensure the probability of the data object isn't 0.

$$P(x|\omega_i) = \frac{N_{i,\omega} + \alpha}{N_i + \alpha d}$$

Where N_i is the count of that feature, $N_{i,\omega}$ is the count for that class, α is the smoothing parameter (1 for laplace smoothing) and d is the number of dimensions.

m	*The number of clusters generated by the algorithm.*
n_i	*The number of data points in the ith cluster.*
c_i	*The centroid for the ith cluster.*
x_{ij}	*The jth datapoint in cluster c_i.*
c	*The mean centroid for the whole data set.*
$d(x, y)$	*The distance function we're using.*