Nearest Neighbours Quiz

1. We choose the number of neighbours whose class is used to compute the prediction to generalise in a standard KNN classification method.
2. Distance metrics: Minkowski p-norm, Euclidean, Hamming
3. In the KNN context, a Voronoi tessellation
    A. Puts boundaries at equal distances between neighbouring training points
    B. Is the basis for constructing the decision boundary for 1NN
    C. Partitions the space into regions around training points
    D. Creates a complex decision boundary
4. A 1NN solution is sensitive to outliers
5. A 1NN solution does not allow us to asses the confidence of our prediction P(y|x)
6. The kNN approach to classification, where k>1...
    A. Selects the closest k training points and predicts the most frequent class in that subset
    B. Is less sensitive than 1NN to outliers
7. When classifying using a kNN, sometimes you find that two classes are equally predicted by the k nearest neighbour, e.g. if k = 9 and class A is predicted by 4 NN, B by 4 NN and C by 1 NN. The solution to finding the prediction is get the class predicted by 1NN, and if that is a tie then at random between the tied 1NN classes.
8. The difference between kNN classification and kNN regression is that in kNN classification the output is the most frequent label or class amongst the k nearest neighbours, whereas in kNN regression the output is the average of the labels of the k nearest neighbours.
9. If we select k to be too high, then we'll find that all the instances will be classified in the most probable class, and if we set it too low we'll find that the decision boundary is unstable. So one way to pick k is to try different k, and pic the k that performs best on a validation set.
10. In the Parzen version of nearest neighbour, we consider the training instances that fall in a fixed region or volume around the training instance.
11. In the kernelised nearest neighbour method, every training instance contributes to the answer in proportion to the kernel value for that training instance and the testing instance.
12. kNN method
    A. Sensitive to outliers
    B. Computationally expensive at testing time
    C. Good in an online streaming setting where new instances are arriving frequently, just add the new datapoint to the list of training instances
    D. In comparison to other methods it makes few assumptions about the data
    E. Missing data needs to be filled in
    F. Non-parametric approach
13. kNN speedup methods
    A. K-D tree: low-dimensional and real-valued data
    B. Inverted lists: high-dimensional and sparse data
    C. Locality sensitive hashing: high-dimensional and real-valued data