Similarity-Based Methods

1. What is the most common (dis)similarity metric for numeric vectors?

   Solution: Euclidian distance

2. What is Hamming distance for vectors of categorical features?

   Solution: The total number of features for which two vectors disagree on the values.

3. In many text mining systems documents are represented as vectors of word counts, where the length of a document vector is the number of words appearing in the entire text corpus and each index in a document vector represents the number of times a particular word appears in the document. Is Euclidian distance a good (dis)similarity metric? If not, why not, and which (dis)similarity metric would work better?

   Solution: No. Euclidian distance does not take scale into account. A longer document is "more" "about" a subject than a shorter document about the same topic. Normalized euclidian distance or cosine similarity are better document similarity metrics.

4. What is a Voronoi tesselation?

   Solution: A paritioning of the space into regions in which the boundaries are maximally different from the points in each region.

5. What is the nearest-neighbor rule for classification, in words and mathematically?

Solution:

$$g(\vec{x}) = y_{[1]}$$

where $y_{[1]}$ is the label of the nearest neighbor to $\vec{x}$

6. What is the VC-dimension of nearest neighbor classifiers? Why?

Solution:

7. What is the generalization bound of nearest neighbor classifiers?

Solution: For any $\delta > 0$, and any continuous noisy target $\pi(\vec{x})$, there is a large enough N for which, with probability at least $1 - \delta$, $E_{out} \leq 2E^*_{out}$.

-

8. What is the main advantage of nearest-neighbor classification?

Solution: Simplicity

9. What is the main disadvantage of nearest-neighbor classification?

Solution: Computational cokmplexity

10. What is the k-Nearest Neighbor classification rule, in words and mathematically?

Solution:

$$g(\vec{x}) = sign\big(\sum_{i=1}^{k} y_{[i]}(\vec{x})\big)$$

where $g(\vec{x})$ is the final hypothesis, k is the number of nearest neighbors, and $y_{[i]}$ is the classification ($+1$ or -1) of the ith nearest neighbor.

11. What happens to the complexity of the hypothesis as k is decreased or increased?

Solution: As k increases, the complexity of the k-NN hypothesis class increases.

12. List three ways the hyperparameter k can be chosen.

Solution:

- $k = 3$
- $k = \sqrt{N}$
- CV (cross-validation)

13. What is a parametric model?

Solution: A model with paramters to learn (estimate), such as μ and σ^2 for a Gaussian model or the weights $\vec{w}$ of a linear model.

14. What is a non-parametric model?

Solution: A model wich does not include any parameters to be learned.

15. Which is k-NN: parametric, or non-parametric?

Solution: Non-parametric

16. What is the difference between an eager learning algorithm and a lazy learning algorithm?

Solution: An eager learning algorithm learns a model of the data from samples before being queried on an unseen data sample. A lazy algorithm does not learn a model but rather simply uses the training samples to answer queries about unseen samples.

17. Which is k-NN: eager, or lazy?

Solution: Lazy

18. What is the memory complexity of k-NN?

Solution: $O(Nd)$

19. What is the time complexity of k-NN?

Solution: $O(Nd)$

20. List two categories of approaches to improving the efficiency of k-NN.

Solution:

- Reduce the amount of data – represent class boundaries with only necessary samples
- Store data in specialized data structures to speed-up nearest neighbor searches.

21. Describe the basic idea behind the condensed nearest neighbor algorithm.

Solution: Retain only the data samples necessary to define the class boundaries.

22. Describe the basic idea behind improving nearest-neighbor search efficiency through clustering.

Solution: If clusters are good (see below) then nearest neighbor search can be reduced to nearest cluster search.

23. Describe the two primary characteristics of good clusters?

Solution:

- "Tightess" – cluster members are close to the cluster centers
- "Separation" – clusters are distant from each other

24. What are the three primary ways to choose k for k-means clustering?

Solution:

- $k = 3$
- $k = \sqrt{N}$
- CV (cross-validation)

25. Describe the E-step and M-step in expectation-maximization algorithms.

Solution:

-