CS 189/289A Introduction to Machine Learning
Fall 2023 Jennifer Listgarten, Jitendra Malik DIS8

1 Curse of Dimensionality in Nearest Neighbor Classification

We have a training set: $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$, where $\mathbf{x}^{(i)} \in \mathbb{R}^d$. To classify a new point $\mathbf{x}$, we can use the nearest neighbor classifier:

$$\text{class}(\mathbf{x}) = y^{(i^*)} \quad \text{where } \mathbf{x}^{(i^*)} \text{ is the nearest neighbor of } \mathbf{x}.$$

Assume any data point $\mathbf{x}$ that we may pick to classify is inside the Euclidean ball of radius 1, i.e. $\|\mathbf{x}\|_2 \leq 1$. To be confident in our prediction, in addition to choosing the class of the nearest neighbor, we want the distance between $\mathbf{x}$ and its nearest neighbor to be small, within some positive ϵ:

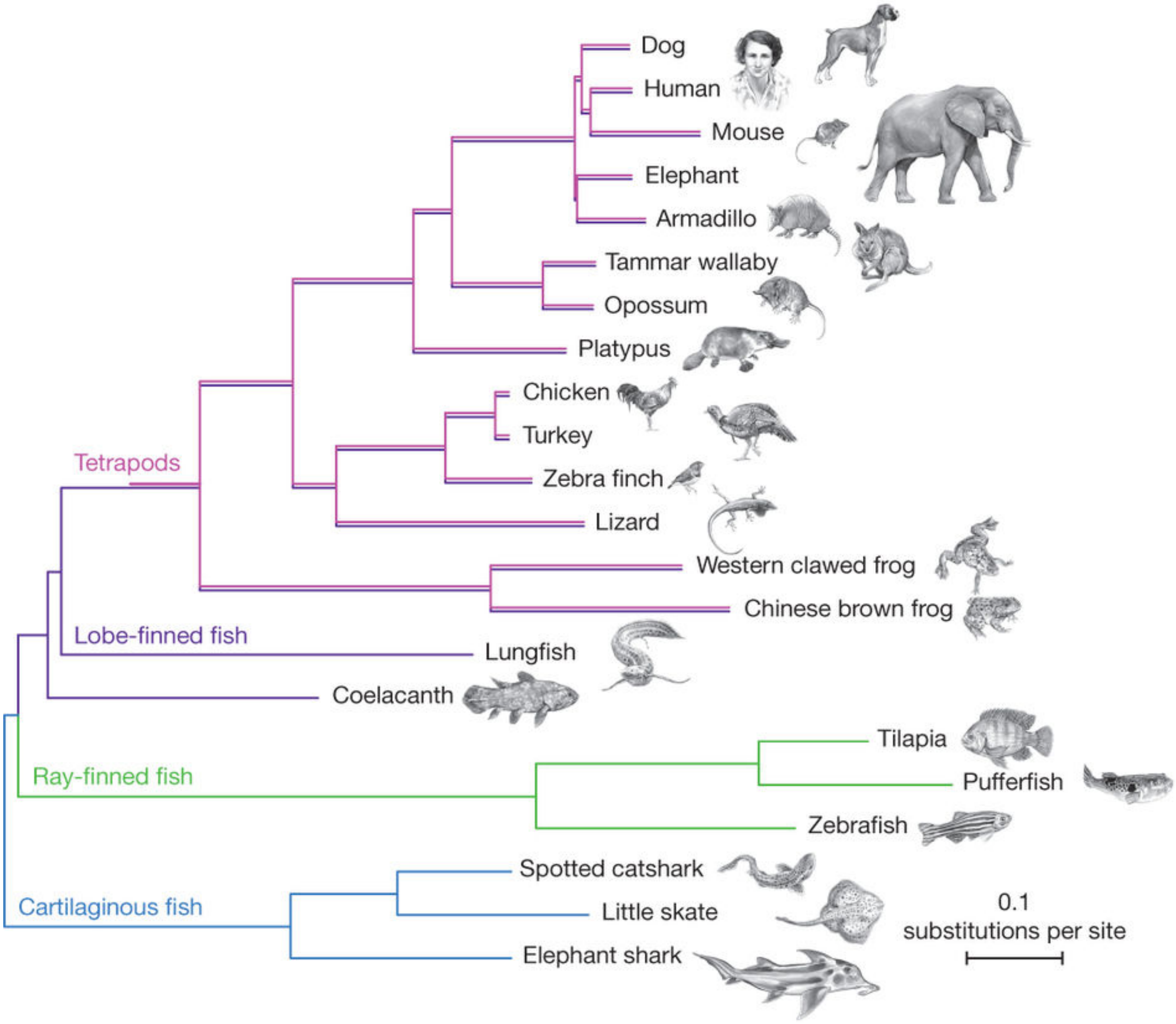
$$\|\mathbf{x} - \mathbf{x}^{(i^*)}\|_2 \leq \epsilon \quad \text{for all } \|\mathbf{x}\|_2 \leq 1. \tag{1}$$

What is the minimum number of training points we need for inequality (1) to hold (assuming the training points are well spread)? How does this lower bound depend on the dimension d?

Hint: Think about the volumes of the hyperspheres in d dimensions.

2 Hierarchical Clustering for Phylogenetic Trees

A phylogenetic tree (or "evolutionary tree") is way of representing the branching nature of evolution. Early branches represent major divergences in evolution (for example, modern vertebrae diverging from modern invertebrate), while later branches represent smaller branches in evolution (for example, modern humans diverging from modern monkeys). An example is shown below.



Creating phylogenetic trees is a popular problem in computational biology. We are going to combine what we know about clustering, decision trees, and unsupervised learning.

We start with all the samples (in this case, animals) in a single cluster and gradually divide it up. This should remind you of decision trees! After k steps, we have at most 2^k clusters. Since we do not have labels, we need to find some way deciding how to split the samples (other than using entropy).

We will use the same objective as in k-means clustering to determine how good our proposed clustering is:

$$\forall i \leq k, \mu_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j$$

$$H(S_1, \ldots, S_k) = \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2$$

At each iteration, we will split each cluster with more than one element into two clusters. The algorithm terminates when everything is in its own cluster.

(a) Consider the following six animals and their two features. Create the resulting decision tree.

Animal	Lifespan	Wings
Dog	12	0
Human	80	0
Mouse	2	0
Elephant	60	0
Chicken	8	2
Turkey	10	2

(b) Prove that an optimal clustering on $k + 1 < n$ clusters has an objective value that is at least as small as that of the optimal clustering on k clusters.

(c) What is the value of $H(S_1, \ldots, S_k)$ when $k = n$ (the number of samples)?

3 Surprise and Entropy

In this section, we will clarify the concepts of surprise and entropy. Recall that entropy is one of the standards for us to split the nodes in decision trees until we reach a certain level of homogeneity.

(a) Suppose you have a bag of balls, all of which are black. How surprised are you if you take out a black ball?

(b) With the same bag of balls, how surprised are you if you take out a white ball?

(c) Now we have 10 balls in the bag, each of which is black or white. Under what color distribution(s) is the entropy of the bag minimized? And under what color distribution(s) is the entropy maximized? Calculate the entropy in each case.
Recall: The entropy of an index set S is a measure of expected surprise from choosing an element from S; that is,

$$H(S) = -\sum_{C} p_C \log_2(p_C), \text{ where } p_C = \frac{|i \in S : y_i = C|}{|S|}.$$

(d) Draw the graph of entropy $H(p_c)$ when there are only two classes C and D, with $p_D = 1 - p_C$. Is the entropy function strictly concave, concave, strictly convex, or convex? Why? What is the significance?
Hint: For the significance, recall the information gain.