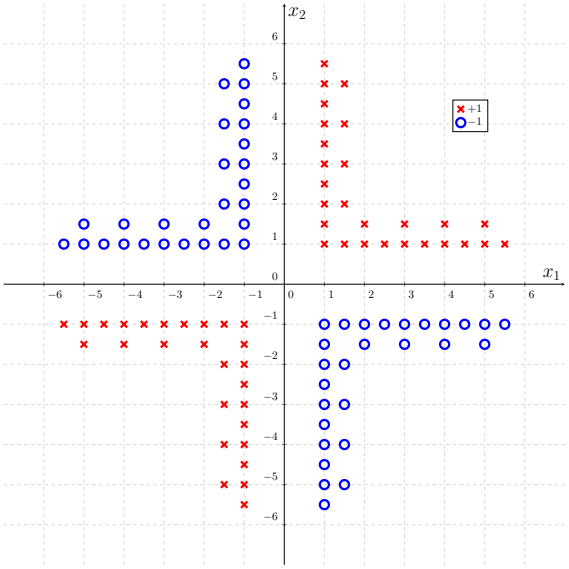
6.390 Introduction to Machine Learning

Recitation Week #5

Issued March 6, 2023

1. Consider the following data set:



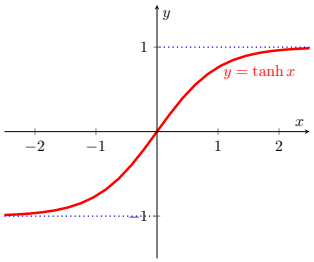
Suppose that we would like to design non-linear transformations to introduce new features to create a linearly separable dataset. Out of the choices below, determine whether these choices of features would make the dataset linearly separable or not, and explain.

Hint: Consider the data points which reside in each of the four quadrants of the plot and reason what will happen to groups of points with the proposed feature transformations.

(a) $[x_1, x_2, x_1 x_2]$.

(b) $[x_1^2, x_2^2, \frac{x_1+x_2}{2}]$.

(c) $[x_1 + x_2, x_1 \tanh(x_2), 1]$. Recall that $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$:



2. We work for an investment banking firm Silver Bags, and we are trying to build several predictive models about the stocks of companies.

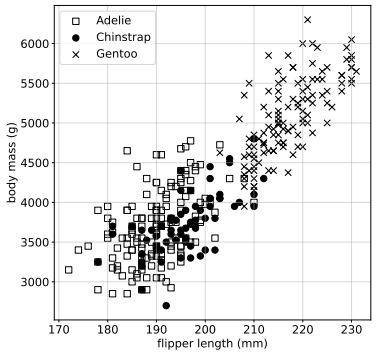
   Companies are described to us in terms of 3 features. For each feature, describe a transformation to make a new feature vector where every element is in $\mathbb{R}$. Ultimately, we will concatenate all these new feature vectors to represent the company in a machine learning algorithm, so you should choose wisely with that goal in mind. It is totally reasonable for more than one transformation to exist, so please explain your reasoning!

   (a) Market segment (one of "service," "natural resources," or "technology.")

   (b) Number of countries in which it operates $(1 - 50)$.

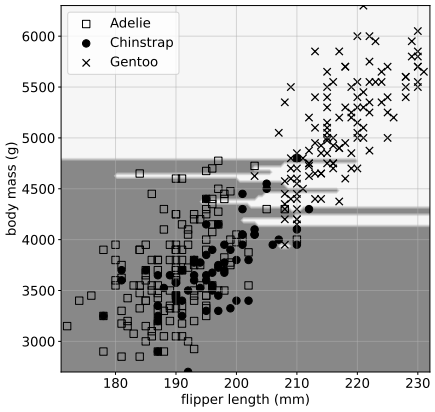
   (c) Total valuation (-1 billion to $+1$ billion).

3. In this problem we will analyze the Palmer Archipelago (Antarctica) penguin data set for classification.



We consider two features: body mass and flipper length. This dataset consists of three different species of penguins: Adelie, Chinstrap and Gentoo. There are 152 Adelie, 124 Gentoo and 68 Chinstrap in the dataset. The dataset is plotted to the left, where Adelie are labeled with a square, Gentoo with an x and Chinstrap with an o. Our goal will be to classify a penguin given its body mass and flipper length.

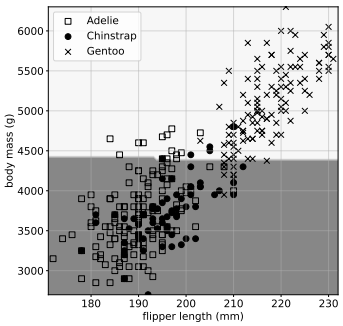
(a) First, consider binary classification for identifying Gentoo penguins. In these below, we've plotted the classification decision and boundary from 1-Nearest Neighbors using Euclidean distance. The classifier labels an input feature vector to be the same as the closest neighbor.

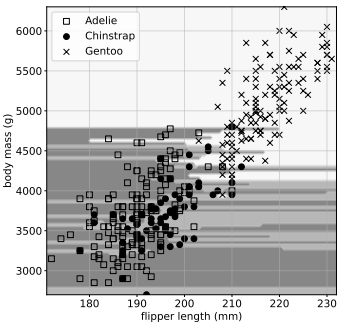


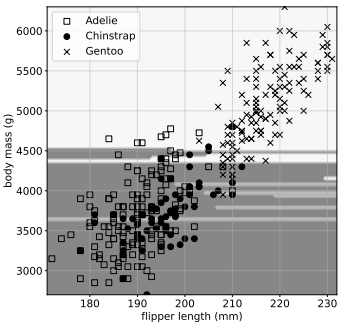
i. Which feature—body mass or flipper length—seems to dominate the decision?

ii. Why are there thin slices in the 1-Nearest Neighbors boundaries? Is this desirable behavior? If not, is there anything we can change to get better behavior from 1-Nearest Neighbors?

(b) Now, we will return to the multiclass classification problem. We run the k-Nearest Neighbors algorithm using Euclidean distance on the Penguins dataset with varying number of neighbors: $k = 1, 10, 100$. The classifier labels an input feature vector to be the *majority vote* of the k closest neighboring feature vectors. Match the plot with the k value used.

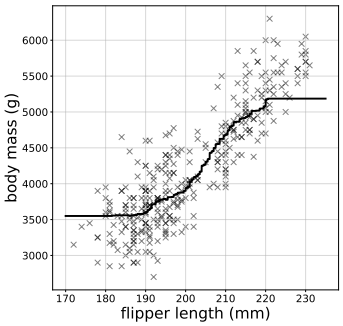


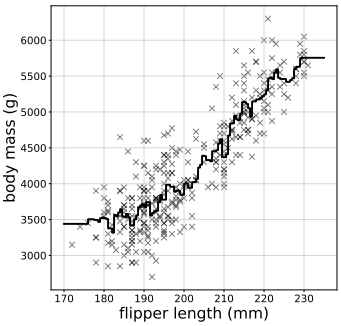


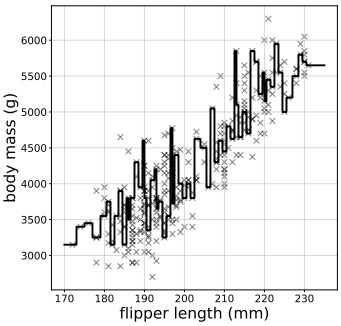


4. Now we will consider the same Penguins dataset for a regression task. Our goal will be to predict the body mass of a penguin—independent of its species—given the length of its flipper.

   (a) We run the k-Nearest Neighbors algorithm using Euclidean distance on the Penguins dataset with varying number of neighbors: $k = 1, 10, 100$. The regressor labels an input feature vector by taking the *average* of the labels of the k nearest neighbors. Match the plot with the k value used.







   (b) Is it possible to extrapolate to higher or lower flipper lengths than those seen in the data? Can we trust the predictions outside of the observed range? Why or why not?