5.2 Projecting onto a function basis

Given the Legendre polynomials $P_4(x) = \frac{35x^4-30x^2+3}{8}$ and $P_5(x) = \frac{63x^5-70x^3+15x}{8}$ over domain $[-1, 1]$, show that $\mathcal{F} = \left\{ \frac{P_4}{\|P_4\|}, \frac{P_5}{\|P_5\|} \right\}$ forms an orthonormal function basis over $[-1, 1]$, assuming $\|P_n\| = \sqrt{\frac{2}{2n+1}}$. Then, by projection, find coefficients c_0 and c_1 such that $\left(\frac{P_0(x)}{\|P_0\|} = \frac{1}{\sqrt{2}}, \frac{P_1(x)}{\|P_1\|} = \frac{x}{\sqrt{2/3}} \right) (c_0 \ c_1)^\top = 1$. To show that you found the right coefficients, work out $\left(\frac{1}{\sqrt{2}} \frac{x}{\sqrt{2/3}} \right) (c_0 \ c_1)^\top$ and show that it is equal to 1.

$$\mathcal{F} = \left\langle \frac{P_4}{\|P_4\|} , \frac{P_5}{\|P_5\|} \right\rangle$$

$$\|P_4\|^2 = \int_{-1}^{1} \left(\frac{35x^4-30x^2+3}{8} \right)^2 dx = \frac{2}{9} \Rightarrow \|P_4\| = \sqrt{2/9}$$

$$\|P_5\|^2 \int_{-1}^{1} \left(\frac{63x^5-70x^3+15x}{8} \right)^2 dx = \frac{2}{11} \Rightarrow \|P_5\| = \sqrt{\frac{2}{11}}$$

$$\mathcal{F} = \left\langle \frac{\left(\frac{35x^4-30x^2+3}{8} \right)}{\sqrt{2/9}} , \frac{\left(\frac{63x^5-70x^3+15x}{8} \right)}{\sqrt{2/11}} \right\rangle$$

$$= \int_{-1}^{1} \left(\frac{(35x^4-30x^2+3)(63x^5-70x^3+15x)}{(8\sqrt{2/9})(8\sqrt{2/11})} \right) dx$$

$$= 0$$ with help from Wolfram ☺

$f_1(x) = \frac{1}{\sqrt{2}}$ $\quad f_2(x) = \frac{x}{\sqrt{2/3}}$ $\quad c_0 = ?$ $\quad c_1 = ?$

$$(f_1(x), f_2(x))\,(c_0, c_1)^T = 1$$

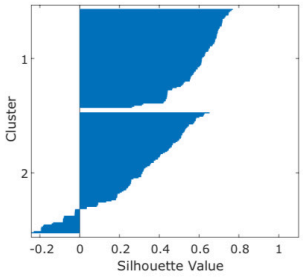
$$c_0 = \int_{-1}^{1} f_1(x)\,dx = \sqrt{2}$$

$$c_1 = \int_{-1}^{1} f_2(x)\,dx = 0$$

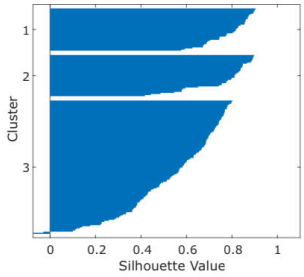
$$\left(\frac{1}{\sqrt{2}}, \frac{x}{\sqrt{2/3}}\right)(\sqrt{2}, 0)^T = \frac{1}{\sqrt{2}} \cdot \sqrt{2} = \underline{\underline{1}}$$

5.5 Silhouettes

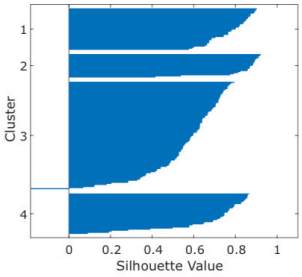
Assume we have a dataset containing information about our customers. The dataset contains various information, including the annual income of the customers and spending scores. A spending score is a number between 1-100, and it is based on the amount of money spent in the mall. We would like to use these two measures now to identify groups of customers so we can plan our marketing strategies accordingly. However, we do not know how many groups of customers we have in the data. We thus computed several versions of k-means clustering with different Ks. **Your task is to pick a K, which is the best one based on the below silhouette charts, i.e., decide how many clusters we have in the data and explain your answer.**



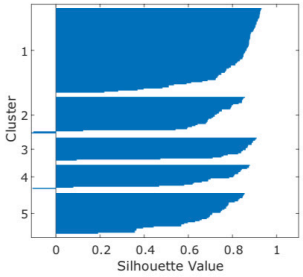
(a) K==2



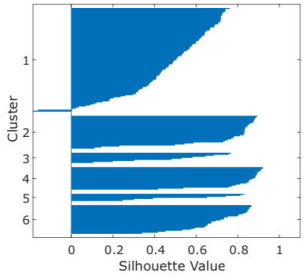
(b) K==3



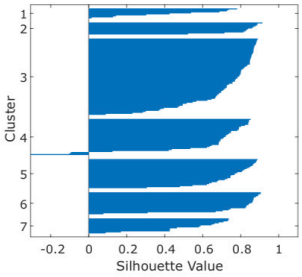
(c) K==4



(d) K==5



(e) K==6



(f) K==7

Silhouette values are how close clusters are to other clusters. 1 = good/far away 0 = bad/close clusters.

K = 5 seems to be the best option, where all clusters seem fairly seperate.

K = 3 has the fewest errors (negative values)

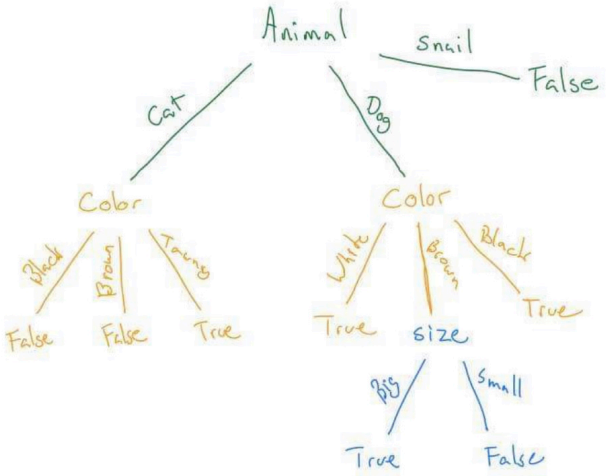
5.6 Decision Tree

At the police station, the officers collected a dataset containing information about various encounters with animals in the city. They noted the animal's type, color, size, and whether the animal had fur or not for each encounter. Finally, they marked the encounter to be dangerous or safe. Today, you were hired as a data analyst to create a decision tree that the task force can use in the future to avoid possible casualties when encountering dangerous animals. **Thus, your task is to build the most effective decision tree based on the below dataset and explain why it is the most effective one and how you constructed it.**

Hint: To determine the variables that best splits the items on each level, you can use the Gini impurity metric (this metric was intuitively explained in the video for the L7a lecture), but feel free to experiment with others if you want: `https://en.wikipedia.org/wiki/Decision_tree_learning#Metrics`

ID	Animal	Color	Size	Fur	Dangerous
1	Snail	white	small	no	FALSE
2	Cat	black	small	yes	FALSE
3	Dog	white	small	yes	TRUE
4	Dog	brown	big	yes	TRUE
5	Dog	black	big	yes	TRUE
6	Cat	brown	small	yes	FALSE
7	Cat	tawny	small	no	TRUE
8	Snail	brown	big	no	FALSE
9	Dog	brown	small	yes	FALSE
10	Snail	black	small	no	FALSE
11	Cat	tawny	big	yes	TRUE
12	Snail	brown	big	no	FALSE

Using Informationg Gain reduction



$H(T) = -\left(\frac{5}{12} \cdot \log_2 \frac{5}{12} + \frac{7}{12} \cdot \log_2 \frac{7}{12}\right) = .9799$

$IG(H(T)|\text{Animal}) = H(T) - H(T|\text{Animal}) = .0376$

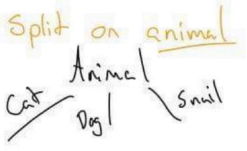
$IG(H(T)|\text{Color}) = 0.283$

$IG(H(T)|\text{Size}) = 0.22$

$IG(H(T)|\text{Fur}) = 0.104$

$$H(T|a) = \sum_{v \in \text{vals}(a)} \frac{|S_a(v)|}{|T|} \cdot H(S_a(v))$$

Split on animal



Then we do this for each subtree

Cat:

$IG(H(T|cat)|color) = 1$

$IG(H(T|Cat)|size) = 0.311$

$IG(H(T|Cat)|Fur) = 0.311$

Pick color to explain all cases

Snail: Fully explained already

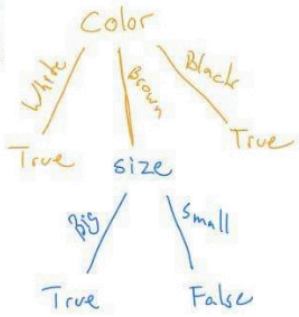
Dog:

$IG(H(T|Dog)|color) = 0.311$

$IG(H(T|Dog)|size) = 0.311$

$IG(H(T|Dog)|Fur) = 0$

arbitrary choice will need both



5.7 Quiz

Answer the following questions (and explain your answers):

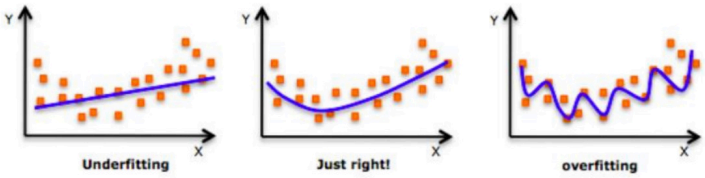
1. Explain why it is important to select the starting position for centroids in the K-means algorithm correctly. What strategies are used for such a selection?
2. Explain what overfitting is, why it is good or bad, and how it is connected to the complexity of the trained model? Propose some strategies that can be used to increase its positive or reduce its negative effect.
3. Explain the similarities and differences between Linear discriminant analysis (LDA) and Principal component analysis (PCA).
4. Explain what a Perceptron is and how it is used in neural networks.

1.

As k-means clustering aims to converge on an optimal set of cluster centers (centroids) and cluster membership based on distance from these centroids via successive iterations, it is intuitive that the more optimal the positioning of these initial centroids, the fewer iterations of the k-means clustering algorithms will be required for convergence.

There are a number of initialization strategies:
random data points: highly volatile, not well positioned throughout the entire data space.
k-means++: Dispersed set.First assigning centroid to the location of a randomly selected data point, then choosing the subsequent centroids from the remaining data points based on a probability proportional to the squared distance away from a given point's nearest existing centroid. Pushes the centroids as far from one another as possible, covering as much of the occupied data space as they can from initialization.
Sampling: cluster a smaller subset of data using a different clustering algorithm, then pick representatives from each cluster. Can be effective, includes hierarchical clustering.

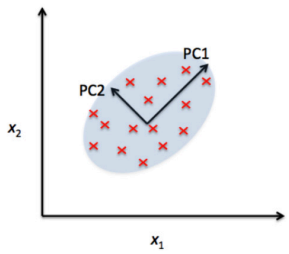
2.



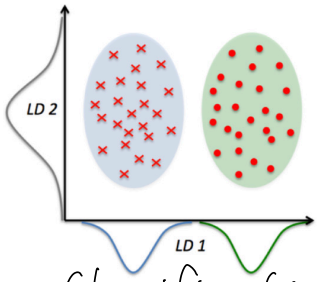
Overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably". An overfitted model contains more parameters than can be justified by the data, in that sense the model has a higher complexity than the data. It can be bad since it can be bad at predicting for data not seen before, which often is the purpose of models. To reduce overfitting one can reduce the complexity of the model (number of parameters/degrees and sudh) or use regularization algorithms to the dataset by introducing additional information to simplify models.

3.

a. Both LDA and PCA are dimensionality reduction algorithms meaning that they reduce the number of random variables to a set of principal variables, and find structures in data to reduce dimensionality and lower dimensional variables. The difference is that LDA is supervised, which means that the developer must specify class labels. PCA on the other hand ignores class labels , and is therefore unsupervised.



Variance



Classification

4.

A Perceptron loosely mimics a neuron in the human brain. More specifically it is a binary classifier that decides whether or not an input, represented by a vector of numbers, belongs to some specific class. It combines a set of weights with the feature vector to calculate whether to return 1 or 0. The 4 parts of a Perceptron are input values, weights and bias, net sum and activation function. Neural networks are a collection of Perceptrons combined to solve a specific input -> output problem.

5.8 Haar

Compute Haar Wavelet decomposition for the following discrete signal **[8,4,1,9,7,7,4,5,9,1]**. You are required to hand in the whole wavelet pyramid, including all intermediate steps/levels from top to bottom, until you get a constant signal. You are required to deliver both the averages and differences. Remember to normalize the results not to lose the energy and correctly handle the intermediate steps with odd signal lengths. In the report, explain your strategy for these odd cases.

Signal = $[8, 4, 1, 9, 7, 7, 4, 5, 9, 1]$

$CA1 = [6, 5, 7, 4.5, 5] / \sqrt{2}$
$CD1 = [2, -4, 0, -0.5, 4] / \sqrt{2}$

Odd length: copy last value

$CA2 = [5.5, 5.75, 5] / \sqrt{2^2}$
$CD2 = [0.5, 1.25, 0] / \sqrt{2^2}$

Odd length: copy last value

$CA3 = [5.6250, 5] / \sqrt{2^3}$
$CD3 = [-0.125, 0] / \sqrt{2^3}$

$CA4 = [5.3125] / \sqrt{2^4}$
$CD4 = [0.3125] / \sqrt{2^4}$