Week 1

* 1.1: Explain the history and purpose of data mining across multiple disciplines.
* 1.2: Describe different data mining tasks.
* 1.3: Recognize attributes of data needed for data mining.
* 1.4: Review and summarize data exploration techniques for use in initial data analysis.

Week 2

* 2.1: Define classification and classification applications.
* 2.2: Compare and contrast different decision tree approaches.
  + In summary, ID3 and C4.5 are entropy-based approaches that are particularly suited for classification tasks. CART is a binary splitting method that can handle both classification and regression tasks, and it's particularly suited for large datasets. CHAID is a statistical-based method that is suited for categorical data. The choice of decision tree algorithm depends on the type of data, the size of the dataset, and the task at hand.
* 2.3: Apply common algorithms for decision tree-based algorithms.
  + ID3 (Iterative Dichotomiser 3): ID3 is an algorithm that uses an entropy-based approach to build decision trees. It iteratively selects the feature that results in the most information gain (i.e., the feature that reduces the most entropy) and splits the data based on that feature. The process continues until a stopping criterion is met, such as when all the data points in a node belong to the same class or when the tree reaches a maximum depth.
  + C4.5: C4.5 is an improvement over ID3 that uses a more sophisticated measure of impurity called information gain ratio. It also includes the ability to handle continuous variables and missing values, and it can prune the tree to avoid overfitting. C4.5 works by recursively splitting the data based on the feature that maximizes the information gain ratio.
  + CART (Classification and Regression Trees): CART is a binary splitting method that can handle both classification and regression tasks. It works by finding the feature and the threshold that minimize the sum of squared errors or the Gini impurity index, depending on the task. The data is then split into two subsets, and the process is repeated until a stopping criterion is met.
  + Random Forest: Random Forest is an ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting. It works by creating multiple decision trees on different subsets of the data and combining their predictions. The random forest algorithm uses the majority vote of the trees to make the final prediction.
  + Gradient Boosted Trees: Gradient Boosted Trees is another ensemble method that combines multiple decision trees. It works by creating decision trees in a sequential manner, where each subsequent tree tries to correct the errors made by the previous tree. The algorithm uses a loss function to measure the error and adjusts the weights of the samples to prioritize the ones that are misclassified.
* 2.4: Identify common hurdles of applying classification techniques.

Week 3

* 3.1: Define instance-based classifiers.
  + If you seen you can classify if not you cant. Nearest neighbor. Doesn’t need exact match. But needs to be close.
* 3.2: Use the basics of probability theory to calculate the Bayes Classifier.
  + P(A|B) = (P(B|A) \* P(A)) / P(B)
  + P(A|B) is the probability of event A occurring given that event B has occurred
  + To calculate P(Y=rich), you need to sum the joint probabilities of Y=rich over all possible values of X1 and X2.
* 3.3: Use the probability estimation to calculate Naive Bayes Classifier.
  + Naïve Bayes Classifier assumes that the attributes are conditionally independent given the class
  + P(C | x1, x2, ..., xn) = P(C) \* P(x1 | C) \* P(x2 | C) \* ... \* P(xn | C) / P(x1, x2, ..., xn) where C is the class variable and x1, x2, ..., xn are the feature values.
    - Based on these information, compute P(Y = rich | X1 = Female, X2 < 40.5) using Naive Bayes assumption of conditional independence.
    - P(Y = rich | X1 = Female, X2 < 40.5) = (P(X1 = Female, X2 < 40.5 | Y = rich) \* P(Y = rich)) / P(X1 = Female, X2 < 40.5)
    - P(Y = rich | X1 = Female, X2 < 40.5) = (0.0362188 \* 0.239181) / 0.1285086
      * P(X1 = Female, X2 < 40.5 | Y = rich) = P(X1 = Female | Y = rich) \* P(X2 < 40.5 | Y = rich)
      * P(X1 = Female, X2 < 40.5 | Y = rich) = (0.0245895 + 0.0116293) \* (0.0245895 + 0.0116293 + 0.253122 + 0.0421768)
      * P(X1 = Female, X2 < 40.5 | Y = rich) = 0.0362188
* 3.4: Recognize the basic structure of neural networks.
  + Compute the forward pass for predicted output for O with parameters shown below.
    - calculate the activation of the hidden layer neurons:
      * z1 = (I1 \* w1) + (I2 \* w2) + b1
      * z2 = (I1 \* w3) + (I2 \* w4) + b1
    - Next, calculate the output of the hidden layer neurons by passing the activation through an activation function, here using sigmoid function:
      * h1 = 1 / (1 + exp(-z1) .. exp(z1) is 2.71828^z1
      * h2 = 1 / (1 + exp(-z2))
    - Then, we compute the activation of the output neuron:
      * z3 = (h1 \* w5) + (h2 \* w6) + b2
    - Finally, we compute the output of the neural network, also using the sigmoid activation function:
      * O = 1 / (1 + exp(-z3))
    - The equation of the hyperplane can be written as Y = mX + b
  + Gradient Descent can converge into local optimum instead of global ones
  + Complex and Time-Consuming
  + Neural networks are sensitive to the presence of noise
  + Why do we need backpropagation for Artificial Neural Networks (ANN)? Why not just use perceptron learning algorithm? - In ANN, we do not have the true error at each hidden layer - Perceptron assumes availability of ground truth at the output. But in ANN we do not have it for the hidden layers.
* 3.5: Identify the Perceptron learning algorithm.
  + A perceptron is a linear classifier.
  + Since activation functions are further combined for hidden layer inputs, the MultiLayer Perceptron MLP is a non-linear classifier.
* 3.6: Use the artificial neural networks learning model.
  + the core problem solved in training an Artificial Neural Networks (ANN) is Gradient descent - It is used to search the parameter space to reduce error by changing the weights at a rate proportional to the differential of the weights.
* 3.7: Explain the underlying concepts behind support vector machines and why they work.
  + Support vectors are the training points closes to the decision boundary
  + advantages of having decision boundaries with a large margin in Support Vector Machines?
    - A large margin implies there is a high confidence in classification.
    - Decision boundaries with large margin tend to have better generalization errors.
    - If the margin is small, then any slight perturbation to the decision boundary can have quite a significant impact on its classification.
  + the key concepts underlying SVMs are maximizing the margin between the hyperplane and the closest support vectors, using kernel functions to capture complex nonlinear relationships, and the ability to handle high-dimensional data efficiently. These concepts make SVMs a powerful and widely used machine learning algorithm.

Compute the value of w6 after one iteration of backpropagation for the following neural network. The initial parameters and the ground truth value for the forward pass is given below. Take I1 = 0.5 and I2= 0.8.

Review the SVM equation below. Select all correct sentences that are true about SVM.

Compute the forward pass for predicted output for O with parameters shown below. Take I1 = 0.05 and I2= 0.1

i = {1,2}

Review the table labeled Figure 1: Food Type Prediction. A K-Nearest Neighbor (KNN) model is used to predict food type based on sugar content and crunchiness for the following training set. If K is set to 10 and the model uses Euclidean distance, which food type will be predicted for Tomato?

Review the table labeled Figure 1: Food Type Prediction. A K-Nearest Neighbor (KNN) model is used to predict food type based on sugar content and crunchiness for the following training set. If the model uses Euclidean distance, which value should be given to K in order for the model to predict Tomato as a Protein?