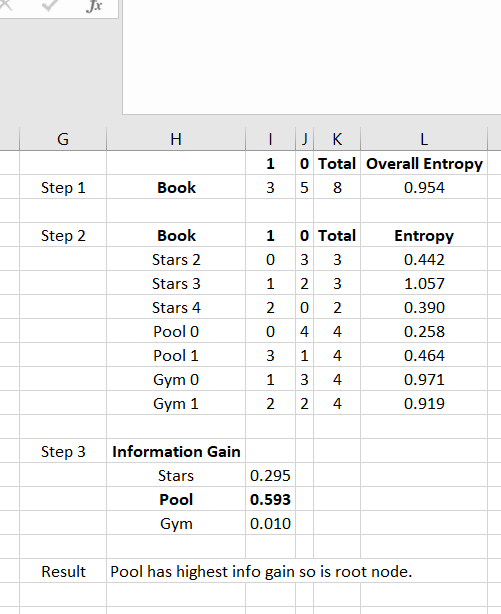
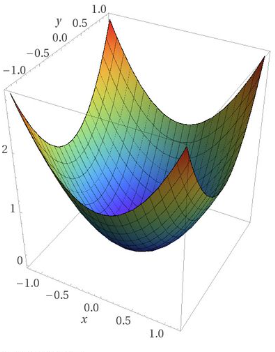
1. Question 1
   * 1. In relation the terms ML prediction models needing to be **consistent** with training data and **generalise** well:
        1. ML algorithms search through all possible patterns that exist between a set of descriptive features and a target feature to find the best model that is **consistent** with the training data (i.e., agrees with all the training instances).
        2. Useful predictive models must be able to **generalise** well, i.e., make predictions for queries that are not present in the training data
     2. **Overfitting** occurs when the model is so complex that it fits to the data too closely and becomes sensitive to noise (e.g., mislabelled feature values). This is different to **Underfitting** which occurs when the prediction model is too simplistic to represent the underlying relationship between the descriptive features and the target feature.
     3. Distance functions that could be used when comparing:
        1. Numeric features are **Euclidean distance** which is the length of a straight line between two points.
        2. Categorical features are **Hamming distance** which is the sum of the overlap differences across all features.
     4. **Non-deterministic** means running the algorithm several times on the same data, could give different results. Such as different starting positions can lead to different local minima or different clustering of the same data.
     5. **Data Leakage** might lead a system to train a suboptimal model that performs significantly worse in practice but not known during training because test data was present in training data. To remedy this the data should be split into training and testing data prior to any imputer or model training.
     6. **K-fold** cross validation provides more robust evaluation than a single random train-test split because hold-out sampling requires that we have enough data available for suitable training, testing and if required validation sets. Any partition too small will result in poor evaluation. Also, performance measured using hold-out sampling can be misleading if we make a lucky split of the data that places the difficult instances into the training set and easy ones in the test set. This makes the model seem more accurate than it is when deployed. K-fold cross validation attempts to address these two issues.
   1. Decision Trees and Nearest-Neighbour classifiers for supervised learning system:

|  |  |
| --- | --- |
| Decision Trees | Nearest-Neighbour |
| - Uses Informative metrics (Info Gain)  - Very interpretable  - Not great with continuous of features  - Not great with lots of features  - Not great with small datasets  - Eager learner, not good if model changes over time but faster | - Uses Distance Metrics (Manhattan)  - Interpretable  - Can handle continuous of features  - Not great with lots of features  - Sensitive to noise  - Lazy Learner, good with concept creep but slow with large data. |

* 1. Decision Tree Calculation
     1. Overall entropy = 0.954
     2. Pool has highest info gain so is root node.

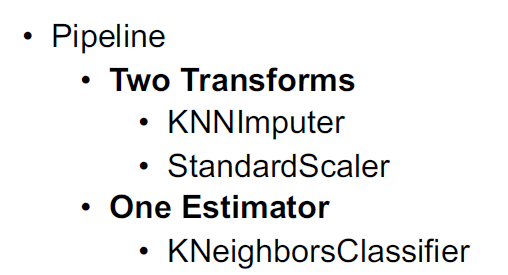
1. Question 2
   1. Naïve Bayes
      1. Smoothing take some of the probability from the events with lots of the probability share and gives it to events with no probabilities
      2. 
   2. Dimension Reduction
      1. Three reasons to reduce number of features in dataset
         1. Computational Cost can be high with significantly increased running time and memory usage.
         2. Financial Cost can be high running experiments to generate feature values can be expensive, especially in clinical medicine manufacturing.
         3. Interpretability can be poor because feature set is not compact to help better understand underlying process that generated the data.
      2. Difference between filters and wrappers are:
         1. Filters have a pre-processing step that ranks and evaluates features to filter them. Evaluation is done using information gain, gini index, chi-square, etc. Can be faster than wrappers but filters have no model bias and no feature dependencies.
         2. Wrappers wrap the classifier in feature selection. Features subsets are then evaluated directly based on their performance when used with the specific classifier. There is a computational cost and risk of overfitting if wrappers are used.
      3. Filters and wrappers select different sub features from a dataset because filters have no model bias and no feature dependencies whereas wrappers do.
   3. Evaluation
      1. Performance measures to use in this binary classification of spam/not-spam are Precision (TP/TP+FP) and Recall (TP/TP+FN) to plot a PR curve because of the imbalanced confusion matrix an ROC curve would not be suitable.
      2. This classifier is performing reasonably well because glancing at the confusion matrix both the true positives and true negatives are high. It’s misclassification rate ({FP+FN}/{TP+TN+FP+FN}) is 21%, it’s wrong under ¼ of the time or 79% accurate. The model seems to incorrectly predict not-spam 32% of the time and incorrectly predict spam 23% of the time, showing an issue not correctly labelling spam as spam but letting it through. This model is better at predicting non-spam levels than spam levels. The model is 81% precise with recall of about 80%. 19% incorrectly marking non-spam as spam and 20% letting spam through. The F1 harmonic mean, useful alternative to misclassification, tending towards smaller numbers, is 0.81. F1 values closer to 1 are better. This indicates that the model is somewhat decent at predicting spam.
      3. Classification accuracy might not always be adequate because it can mask poor performance where a dataset in imbalanced. Example KNN has 90 non-churn and naïve bayes has 70 non churns. The 90 overwhelms the performance of the churn level and is misleading. Average class accuracy can resolve this issue.
2. Question 3
   1. Gradient decent
      1. Gradient decent is an optimisation algorithm that introduces preference bias, favouring models that use gradient decent approach through weight space and can optimise the parameters of an ml model by selecting a random point in the weight space so each weight is assigned a random value within some sensible range. SEE is calculated for training data which defines a point on the error surface. Slope of error is determined using derivative of a point on the surface. Weights are adjusted using the direction of error surface gradient moving to a new point. The algorithm repeats until a global minimum is reached.



* + 1. Learning rate (alpha) in gradient decent determines the size of the adjustment made to each weight at each step of the process and could learn a sub optimal set of weights if set too high or get stuck in a local minimum if set too low.
  1. k-Nearest Neighbour using Manhattan distance and hyperparameter k=1 to predict if weather is suitable to go out.
     1. What is prediction given query and K=1; No. Going out has one vote for No.



* + 1. Comment on suitability of k=2; Tie. Going out has one vote for No and one for yes.
    2. Comment on suitability of k=3; No. Going out has two votes for No and one vote for yes.
    3. The impact using Model use un-normalised data is that features having different variance can mean a larger value feature dominates the distance calculation. For example, Cloudy feature has a range of 1 while Temp have a range of 18. Temperature would have a higher rate in similarity metric.
  1. Model building pipelines
     1. To build an evaluation pipeline using hold-out testing, we randomly select an instance from our ABT dataset. The remaining is used to training the algorithm while the one held out is used to test and calculate a performance measure. Using hold out tests avoids the issue of peaking where performance of the model is evaluated on the same data used for training which will cause issues with generalisation.



* + 1. To avoid data leakage a clear separation between testing data and training data is needed and that this split is done before normalisation is carried out.
    2. Grid-Search helps to find the optimal hyperparameters of a model which can improve accuracy of prediction. This is because it builds a model for every combination of hyperparameter specified and evaluates each model. Grid-Search can be an exhausted operation which can take a long time and does not take past experience into consideration.
    3. Randomised Search is an alternative to Grid-Search. It is faster than Grid-Search but may not return the best combination of hyperparameters. It also does not take into account past evaluations and will continue regards of the result.