Random Experiment: The act of measuring a process whose outcome is uncertain.
Sample Space: Set of all possible outcomes of an experiment.
Event: Subset of outcomes of an experiment.
Conditional Probability: P(Y|X) probability of Y given X. = P(X,Y)/P(X)
Nominal: (=, !=) Names, ID numbers, Eye Color, Zip Codes. These have names.
Ordinal: (<, >) Rankings, grades, height. These have order.
Interval: (+, -) dates, temperature in C or F. Differences have meanings.
Ratio: (x, /) Length, Time, Counts. Differences and ratios are meaningful.
Noise: Unwanted values, ie distortion on a recording using a poor phone.
Outliers: Real values that are just very different.
Similarity Measure: Value from [0, 1], closer to 1 represents more similar objects.
Simple Matching Coefficient(SMC): Compare matching attributes, divide by all attributes.
$(f_{11}+f_{00})/(f_{11}+f_{10}+f_{01}+f_{00})$
Jaccard: $(f_{11})/(f_{10}+f_{01}+f_{11})$
Entropy: How many bits it takes to represent an occurrence of X. Sum(i..n) $p_i \log_2 p_i$
Classification: Given a collection of records, map each attribute x to an outcome y.
Splits: Binary vs multiple.Gini, entropy, misclassification error are split methods.
Gain: Measure purity before and after split. Gain = P(before) – M(after)
Gini: Minimum 0, with all records in 1 class, implying most interesting info.
Classification: Rule: (Condition) → y, where condition is a conjunction of attributes, y is a class.