Carl Stevenson

Unit 8 notes

* Instance-based classifiers
  + Rote Learner
    - Memorizes training dataand performs classification only if the attributes of the record match one of the training examples exactly
    - Uses example cases to predict future cases
    - Nearest neighbor
      * uses k closet points for classification
    - Geometric similarity
      * Minimality
      * Symmetry
      * Triangle Inequality
    - Amos Tversky
      * Similarity between objects a and b is a function of
        + C, Commonalities
        + A, unique features in a
        + B, unique features in b
    - Issues to resolve
      * Categorization of features
      * Interpretation of features
      * Independence of features
    - Nearest-neighbor classification
      * requires:
        + set of stored records
        + Distance metric to compute distance between records
        + The value of k, the number of nearest neighbors to retrieve
      * To classify an unknown record
        + compute distance to other records
        + identify k nearest neighbors
        + use class labels of nearest neighbors to determine the class label of unknown record
      * Choosing the value of k
        + if k is too small, sensitive to noise points
        + if k is too large, neighborhood may include points from other classes
      * Determine class from nearest-neighbor list
      * take the majority vote of class labels among the k-nearest neighbors
      * may weigh the vote according to distance
      * Attributes may have to be normalized to prevent distance measures from being dominated by one of the attributes
      * Problem with Euclidian measure:
        + High dimensional data
        + “Curse of dimensionality”
        + can produce counter-intuitive results

Possible solutions: normalize vectors to unit length, use different similarity metric

* + - * k-NN classifiers are lazy learners
        + does not build models explicitly
        + classifying unknown records is relatively expensive
      * Bayes Classifier
        + uses probabilistic framework for classification
        + bayesian classifiers

Consider each attribute and class label as random variables

Naive bayes classifier

* + - * Estimating probabilities from data
        + For continuous attributes

Discretize the range into bins

One ordinal attribute per bin

violates independence assumption

Two-Way split

Choose one of the new splits as the attribute

Probability density estimation

assume attribute follows a normal distribution

Use data to estimate parameters of distribution

Once probability distribution is known, can use it to estimate the conditional probability

* + - * + if one of the conditional probabilities is zero, then the entire expression becomes zero