Instance Based Learning

Original slides: Raymond J. Mooney

University of Texas at Austin

Instance-Based Learning

- Unlike other learning algorithms, does not involve construction of an explicit abstract generalization but classifies new instances based on direct comparison and similarity to known training instances.
- Training can be very easy, just memorizing training instances.
- Testing can be very expensive, requiring detailed comparison to all past training instances.
- Also known as:
  - Case-based
  - Exemplar-based
  - Nearest Neighbor
  - Memory-based
  - Lazy Learning

Similarity/Distance Metrics

- Instance-based methods assume a function for determining the similarity or distance between any two instances.
- For continuous feature vectors, Euclidian distance is the generic choice:

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{n} (a_p(x_i) - a_p(x_j))^2}$$

Where $a_p(x)$ is the value of the pth feature of instance x.

- For discrete features, assume distance between two values is 0 if they are the same and 1 if they are different (e.g. Hamming distance for bit vectors).
- To compensate for difference in units across features, scale all continuous values to the interval [0,1].

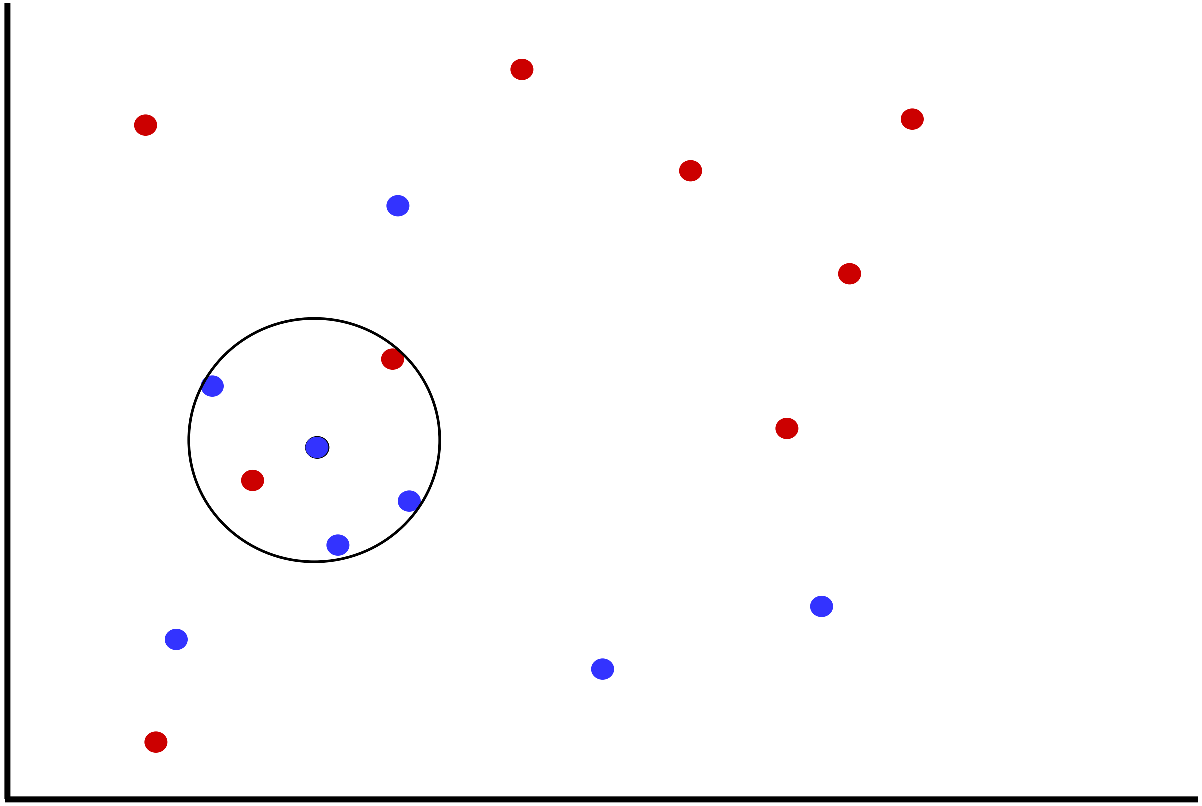
Other Distance Metrics

- Mahalanobis distance
  - Scale-invariant metric that normalizes for variance.
- Cosine Similarity
  - Cosine of the angle between the two vectors.
  - Used in text and other high-dimensional data.
- Pearson correlation
  - Standard statistical correlation coefficient.
  - Used for bioinformatics data.
- Edit distance
  - Used to measure distance between unbounded length strings.
  - Used in text and bioinformatics.

K-Nearest Neighbor

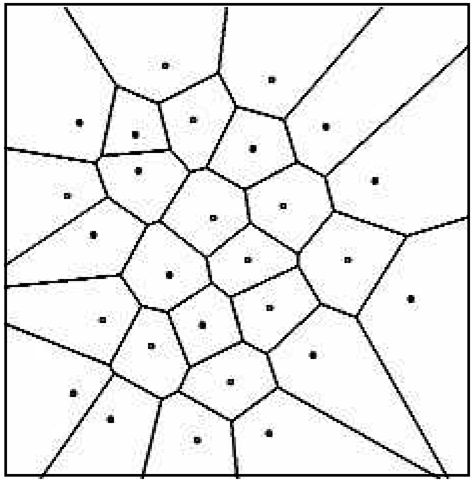
- Calculate the distance between a test point and every training instance.
- Pick the *k* closest training examples and assign the test instance to the most common category amongst these nearest neighbors.
- Voting multiple neighbors helps decrease susceptibility to noise.
- Usually use odd value for *k* to avoid ties.

5-Nearest Neighbor Example



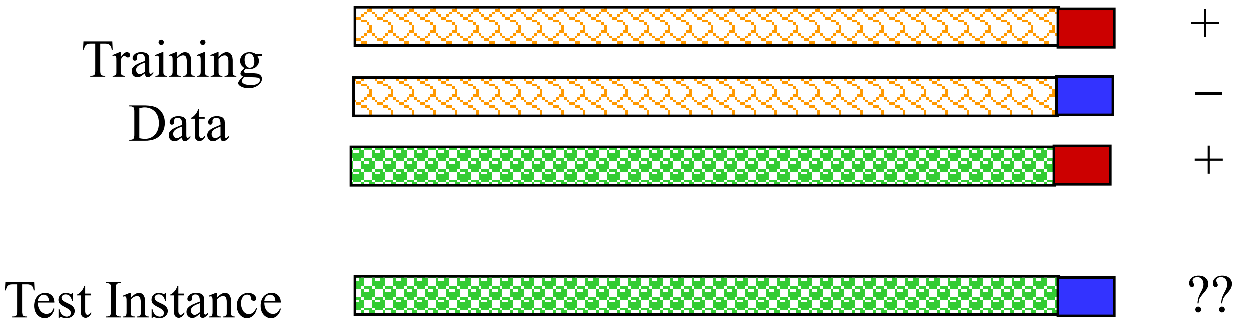
Implicit Classification Function

- Although it is not necessary to explicitly calculate it, the learned classification rule is based on regions of the feature space closest to each training example.
- For 1-nearest neighbor with Euclidian distance, the **Voronoi diagram** gives the complex polyhedra segmenting the space into the regions closest to each point.



Feature Relevance and Weighting

- Standard distance metrics weight each feature equally when determining similarity.
  - Problematic if many features are irrelevant, since similarity along many irrelevant examples could mislead the classification.
- Features can be weighted by some measure that indicates their ability to discriminate the category of an example, such as information gain.
- Overall, instance-based methods favor global similarity over concept simplicity.



Training Data

+

−

+

Test Instance

??

Conclusions

- IBL methods classify test instances based on similarity to specific training instances rather than forming explicit generalizations.
- Typically trade decreased training time for increased testing time.