Some Learning Methods

PSY 322/ORF 322: Human Machine Interactions

Gilbert Harman
Department of Philosophy
Princeton University

Wednesday, March 3, 2004

Pattern Recognition or Classification

- Given some information about a case, we want to classify it in some way
- Examples
  - Automatic mail sorting based on zip codes
  - Computer speech recognition of commands
  - Email spam detection
  - Automatic medical diagnosis based on X-rays or blood samples
- Quality of the system: the accuracy of the classifications as measured for example by the percentage of errors
- We might want to distinguish different sorts of errors.

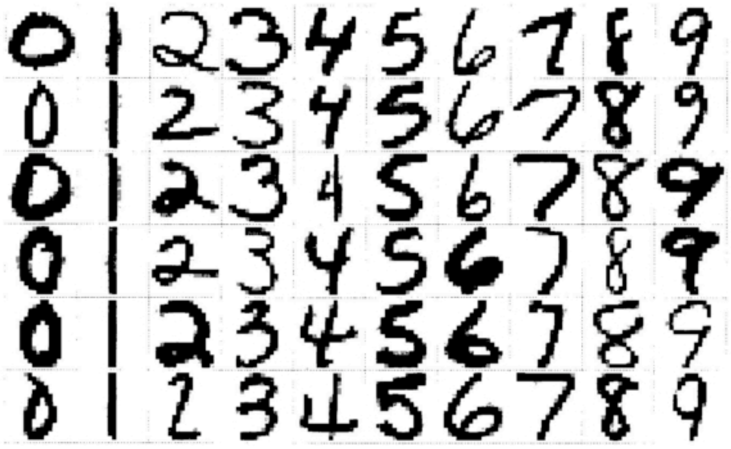


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

Feature Spaces and Feature Vectors

- Items to be classified have certain “observable” features, color, size, mass, temperature, etc.
- Each feature can take any of several values. Color might be red, green, blue, yellow, etc.
- There is a *feature space*: with a number of dimensions, color, size, etc.
- An observation of features locates an object in feature space.
- The features of the object can be represented by a *feature vector* in feature space.

Scene Recognition

- Task is to recognize a scene given values of the pixels on a CRT.
- Each pixel can take eight possible values
- There are $1024 \times 768 = 768,432$ pixels.
- The features are the values of the pixels.
- The feature space has 768,433 dimensions.
- A feature vector is an assignment of values to each pixel.

How Good Is the System?

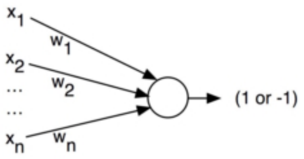
- The true classification of the object is not in general determined by its features. But there will be some probabilistic relation between feature vectors and correction classifications.
- If the probabilities are known, the classification is binary (yes/no), and our goal is to minimize the number of errors, then the best method is to choose each time the most likely classification given the feature vector.
- What if the probabilities are not known?
- We might use data to learn the probabilities, or at least to get a system that does as well as possible.

Very Simple Pattern Recognition Learning Problem

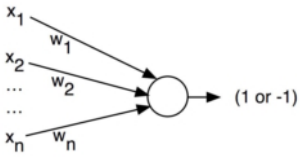
- Assume that there is a background probability distribution that produces objects of a variety of unobserved types with certain observable features.
- Assume that the probabilities with respect to the each object are independent of what previous objects have been produced.
- Assume we do not know anything (else) about the distribution.
- A helpful tutor tells us the correct classifications of the first N objects that are produced. So, we have data indicating associating certain feature vectors with correct classifications.
- We want to use that data in order to be able to do as well as we can in classifying new objects.

Perceptron Classification

- Suppose we want YES/No classification of items
- Perceptron takes inputs from each feature, calculates their weighted sum, and outputs +1 if the sum is greater than 0 and –1 if the sum is less than 0.
- Each input link has a positive or negative weight associated with it.



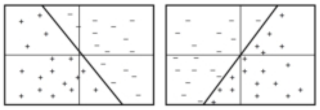
Perceptron Learning



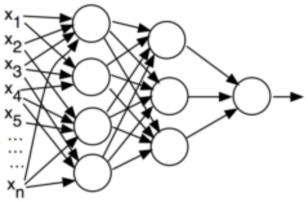
- Perceptron is changed by increasing or decreasing weights to get better results on the data.
- Any classification that can be represented using a perceptron can be learned by this method.

Problem with Perceptrons

- Only linearly separable classifications can be represented.

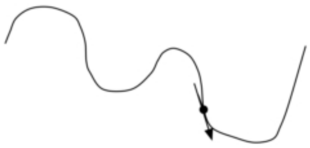


Multi-layer Feed-forward Neural Nets



- If we allow several layers of perceptrons and put thresholds on outputs, any classification can be represented
- If we use an unsharp threshold so the output is a differentible function of the input, we can use gradient descent to train the network on the data.

Features of Neural Net Learning



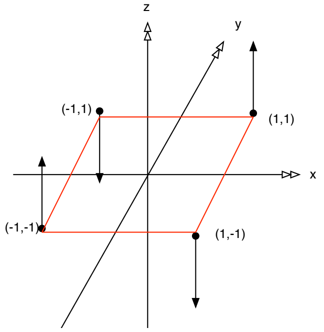
- Lots of data may be needed
- Curse of dimensionality: need separate input node for each pixel in scene analysis problem
- Can get stuck in a local minimum

Alternative: Support Vector Machines

- Problem: not all classifications are linearly separable
- Multi-layer neural nets is one way to go
- Different way: transform the feature space into a different space in which the classification is linearly separable
- Linearly separated classifications are easy to learn (VC dimension is low).
- Example: the XOR problem
  - The data indicate that we should say YES iff either $x > 0$ and $y < 0$ or $x < 0$ and $y > 0$.
  - These data are not linearly separable.
  - One solution is to map points in 2D space to points in 3D space, so that (x, y) is mapped to $(x, y, (x \times y))$.
  - The XOR data are linearly separable in the new space.

XOR Transformed

After mapping (x, y) to $(x, y, (x \times y))$, the xy plane separates the data correctly.



General problem: finding a good transformation as in this case, but there are methods for doing this.

Nearest Neighbor Learning

- Suppose we have a topology and distance measure on the feature space.
- A new item is given the same classification as the nearest item among the data
- This method does better and better the more data we have.
- Sensitive to the topology and distance measure.
- Requires remembering all the data
- Sensitive to dimensions of the feature space ("the curse of dimensionality")
- Determination of nearest neighbor is computationally complex

Transduction

- The methods discussed so far use data to get a principle of classification and use that principle to characterize new data.
- The principles are implicit in neural net and nearest neighbor approaches
- Current work aimed at methods that go directly from data to a new case without even implicitly using the data to find a rule to classify new data
- Transduction versus induction
- In principle it should be harder to find a rule than to classify one or a few new cases

Legal Example

- The Supreme Court wants to decide a case before it using facts of the case, written law, various legal principles, past precedents, etc.
  - The Court does not want to decide certain nearby possible cases.
  - So, the Court tries to settle this case in a way that leaves the other cases maximally unsettled.
- This is like one method currently being tried out to do transduction
- The decision about in a given case depends not just on the data but also the fact that this particular case is being considered and certain other cases are not to be considered.
- The system does not go from data to a general rule for deciding new cases and then decide the next case in accord with that rule

Other Possible Cases of Transduction

- Understanding a poem
- Understanding ordinary conversation in context
- Understanding other people: much ordinary commonsense reasoning
- Moral reasoning

END