MOTIVATION

- Inductive methods, such as decision tree induction and neural network BACKPROPAGATION, seek, general hypotheses that fit the observed training data.
- Analytical methods, such as **PROLOG-EBG ,seek general** hypotheses that fit prior knowledge while covering the observed data.
- Purely analytical learning methods offer the advantage of generalizing more accurately from less data by using prior knowledge to guide learning. However, they can be misled when given incorrect or insufficient prior knowledge.

- Purely inductive methods offer the advantage that they require no explicit prior knowledge and learn regularities based solely on the training data. However, they can fail when given insufficient training data, and can be misled by the implicit inductive bias they must adopt in order to generalize beyond the observed data.
- **The difference between inductive and analytical learning methods** can be seen in the nature of the ***justifications that can be given for their learned hypotheses.***
- Hypotheses output by purely analytical learning methods such as PROLOGEBG carry a ***logical justification; the output hypothesis follows deductively from*** the domain theory and training examples. Hypotheses output by purely inductive learning methods such as BACKPROPAGATION carry a ***statistical justification;***

- The output hypothesis follows from statistical arguments that the training sample is sufficiently large that it is probably representative of the underlying distribution of examples. This statistical justification for induction is clearly articulated in the PAC-learning results.
- Analytical methods provide logically justified hypotheses and inductive methods provide statistically justified hypotheses, it is easy to see why combining them would be useful: Logical Justifications are only as compelling as the assumptions, or prior knowledge, on which they are built. They are suspect or powerless if prior knowledge is incorrect or unavailable. Statistical justifications are only as compelling as the data and statistical assumptions on which they rest.

	Inductive learning	Analytical learning
Goal:	Hypothesis fits data	Hypothesis fits domain theory
Justification:	Statistical inference	Deductive inference
Advantages:	Requires little prior knowledge	Learns from scarce data
Pitfalls:	Scarce data, incorrect bias	Imperfect domain theory

TABLE 12.1
Comparison of purely analytical and purely inductive learning.

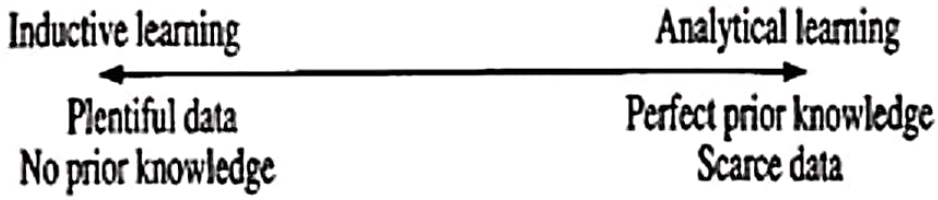

FIGURE 12.1
A spectrum of learning tasks. At the left extreme, no prior knowledge is available, and purely inductive learning methods with high sample complexity are therefore necessary. At the rightmost extreme, a perfect domain theory is available, enabling the use of purely analytical methods such as PROLOG-EBG. Most practical problems lie somewhere between these two extremes.

- Similarly, in analyzing **a stock market** database to learn the target concept "companies whose stock value will double over the next 10 months," one might have approximate knowledge of economic causes and effects, suggesting that the gross revenue of the company is more likely to be relevant than the color of the company logo.

In both of these settings, our own prior knowledge is incomplete, but is clearly useful in helping discriminate relevant features from irrelevant.

- For example, when applying **BACKPROPAGATION** to a problem such as **speech recognition**, one must choose the encoding of input and output data, the error function to be minimized during gradient descent, the number of hidden units, the topology of the network, the learning rate and momentum, etc.
- purely inductive instantiation of BACKPROPAGATION, ***specialized by the designer's choices to the task of speech recognition.***
- We are interested in systems that take prior knowledge as an ***explicit input to the learner, in the same sense that***the training data is an explicit input, so that they remain general purpose algorithms,even while taking advantage of domain-specific knowledge

- **Some specific properties we would like from such a learning method include:**

- **Given no domain theory**, it should learn at least as effectively as **purely inductive methods**.

- **Given a perfect domain theory**, it should learn at least as effectively as purely analytical methods.

- Given an imperfect domain theory and imperfect training data, it should combine the two to outperform either purely inductive or purely analytical methods.

- It should accommodate an unknown level of error in the training data.
- It should accommodate an unknown level of error in the domain theory.
- For example, accommodating errors in the training data is problematic even for statistically based induction without at least some prior knowledge or assumption regarding the distribution of errors. Combining inductive and analytical learning is an area of active current research.

2.1 The Learning Problem

- **Given:**
- A set of training examples D, possibly containing errors.
- A domain theory B, possibly containing errors.
- A space of candidate hypotheses H.
- **Determine:**
- A hypothesis that best fits the training examples and domain theory.

- It is not clear what values to assign to **kB and kD** to specify the relative importance of fitting the data versus fitting the theory. If we have a very poor theory and a great deal of reliable data, it will be best to weight **errorD(h)** more heavily.
- Given a strong theory and a small sample of very noisy data, the best results would be obtained by weighting **errorB(h**) more heavily. Of course if the learner does not know in advance the quality of the domain theory or training data, it will be unclear how it should weight these two error components.

- ***2.Bayes theorem perspective.***
- Bayes theorem computes this posterior probability based on the observed data ***D, together with prior knowledge*** in the form of ***P(h), P(D), and P(Dlh).***
- The Bayesian view is that one should simply choose the hypothesis whose posterior probability is greatest, and that Bayes theorem provides the proper method for weighting the contribution of this prior knowledge and observed data.

- Unfortunately, Bayes theorem implicitly assumes pefect knowledge about the probability distributions P(h), P(D), and P(Dlh). When these quantities are only imperfectly known, Bayes theorem alone does not prescribe how to combine them with the observed data.
- (One possible approach in such cases is to assume prior probability distributions over P(h), P(D), and P(D1h) themselves, then calculate the expected value of the posterior P (h / D) .
- we will simply say that the learning problem is to minimize some combined measure of the error of the hypothesis over the data and the domain theory.

2.2 Hypothesis Space Search

- We can characterize most learning methods as search algorithms by describing the hypothesis space H they search, the initial hypothesis **ho** at which they begin their search, the set of search operators **O t**hat define individual search steps, and the goal criterion **G** that specifies the search objective.

Three different methods for using prior knowledge to alter the search performed by purely inductive methods.

- ***Use prior knowledge to derive an initial hypothesis from which to begin the search.:***
- In this approach the domain theory B is used to construct an initial hypothesis ***ho that is consistent with B. A standard inductive method*** is then applied, starting with the i ***ho.***
- ***For example, the*** **KBANN system learns artificial neural networks in** It uses prior knowledge to design the interconnections and weights for an initial network, so that this initial network is perfectly consistent with the given domain theory. This initial network hypothesis is then refined inductively using the BACKPROPAGATaIlOgNor ithm and available data. Beginning the search at a hypothesis consistent with the domain theory makes it more likely that the final output hypothesis will better fit this theory.

- ***Use prior knowledge to alter the objective of the hypothesis space search:***
- The goal **criterion G** is modified to require that the output hypothesis fits the domain theory as well as the training examples.
- **For example**, the **EBNN** system learns neural networks ,Whereas inductive learning of neural networks performs gradient descent search to minimize the squared error of the network over the training data, EBNN performs gradient descent to optimize a different criterion.
- This modified criterion includes an additional term that measures the error of the learned network relative to the domain theory.

- ***Use prior knowledge to alter the available search steps.***
- In this approach, the set of search **operators 0** is altered by the domain theory.
- **For example, the FOCL** system learns sets of Horn clauses. It is based on the inductive system **FOIL**, which conducts a greedy search through the space of possible Horn clauses, at each step revising its current hypothesis by adding a single new literal.
- **FOCL** uses the domain theory to expand the set of alternatives available when revising the hypothesis, allowing the addition of multiple literals in a single search step when warranted by the domain theory.
- **FOCL** allows single-step moves through the hypothesis space that would correspond to many steps using the original inductive algorithm. These "macro-moves" can dramatically alter the course of the search, so that the final hypothesis found consistent with the data is different from the one that would be found using only the inductive search steps.