19Z601- MACHINE LEARNING

UNIT- 1 INTRODUCTION

INTRODUCTION : Types of Learning - Designing a learning system - concept learning - Find-s Algorithm - Candidate Elimination - Data Preprocessing - Cleaning - Data Scales - Transformation - Dimensionality Reduction. (9)

Presented by

Ms.Anisha.C.D

Assistant Professor

CSE

TYPES OF LEARNING

- Learning = Improving with experience at some task
  - Improve over Task T
  - With respect to performance measure P
  - Based on Experience E
- Examples of Learning Problems:

Checkers Learning Problem	**Task T :** Playing Checkers **Performance Measure P :** Percent of games won against opponents **Training Experience E :** Playing practice games against itself
Handwriting Recognition Learning Problem	**Task T :** Recognizing and classifying handwritten words within images **Performance Measure P :** Percent of words classified correctly **Training Experience E :** A database of handwritten words with given classification.

DESIGNING A LEARNING SYSTEM

1. Choosing a Training Experience
2. Choosing the Target Function
3. Choosing the Representation of the Target Function
4. Choosing a Function Approximation Algorithm
    - 4.1 Estimating Training Values
    - 4.2 Adjusting weights
5. The Final Design

CONCEPT LEARNING

- Concept Learning is the learning process carried out by **inferring a Boolean valued function** from training examples of its input and output.

FIND S ALGORITHM

1. Initialize h to the most specific hypothesis in H
2. For each positive training instance x
   - For each attribute constraint a_i in h

     If the constraint a_i in h is satisfied by x

     Then do nothing

     Else replace a_i in h by the next more general constraint that is satisfied by x
3. Output hypothesis h

CANDIDATE ELIMINATION ALGORITHM

- **Input : Set of instances in the training dataset**
- **Output : Hypothesis G and S**

Step 1: Initialize G, to the maximally general hypotheses.

Step 2: Initialize S, to the maximally specific hypotheses.

- Generalize the initial hypothesis for the first positive instance

Step 3: For each subsequent new training instance

- If the instance is positive
  - Generalize S to include the positive instance
  - ➢ Check the attribute value of the positive instance and S
    - If the attribute value of positive instance and S are difference, fill that field value with '?'
    - If the attribute value of positive instance and S are same, then do no change
- If the instance is negative
  - Specialize G to exclude the negative instance,
  - Add to G all minimal specialization to exclude the negative example and be consistent with S
    - If the attribute value of S and the negative instance are different, then fill that attribute value with S value
    - If the attribute value of S and negative instance are same, no need to update 'G' and fill that attribute value with '?
  - Remove from S all inconsistent hypotheses with the negative instance.

CANDIDATE ELIMINATION METHOD - PROBLEM

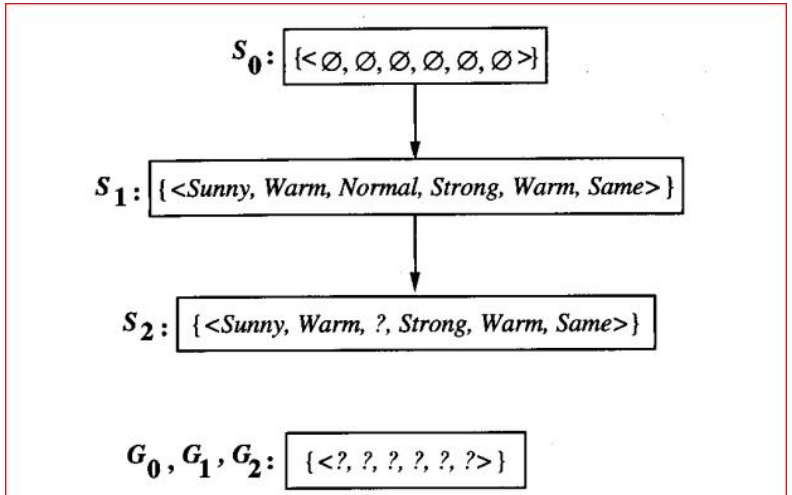
PROBLEM (QUESTION):

Example	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

STEP 1 (CANDIDATE ELIMINATION TRACE 1) :

Define the initial boundary sets S_o and G_o corresponding to the most specific and most general hypotheses.

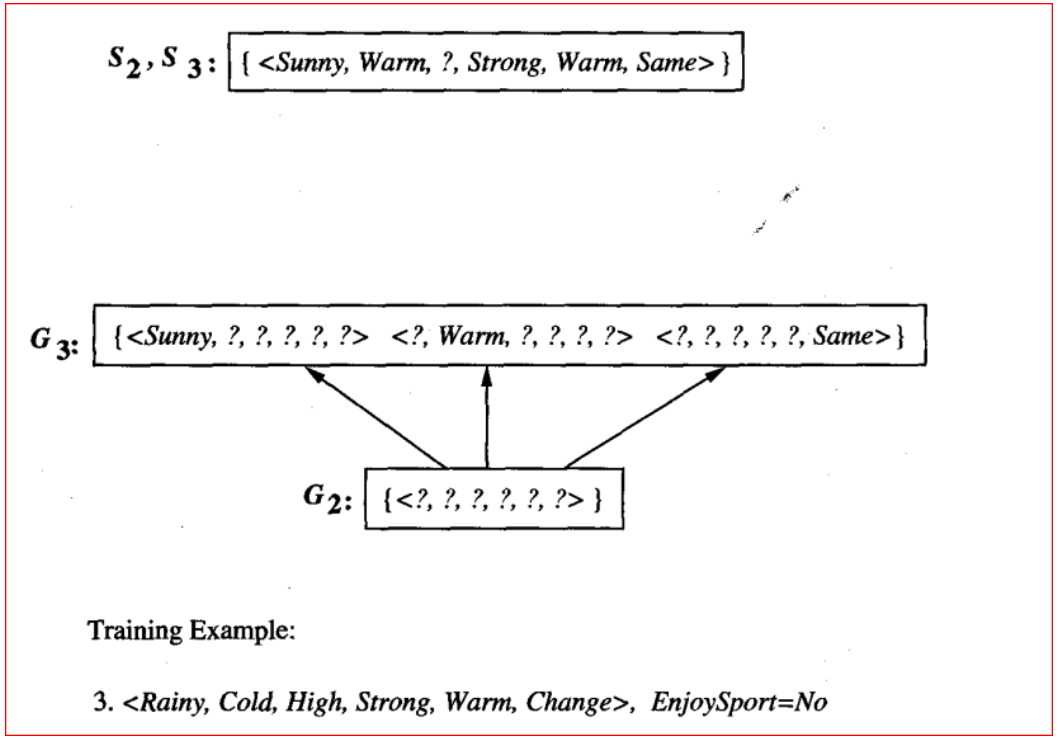
Training examples 1 and 2 are positive and force the S boundary to be more general and has no effect on G boundary.



CANDIDATE ELIMINATION METHOD – PROBLEM (CONTD..)

STEP 2 (CANDIDATE ELIMINATION TRACE 2) :

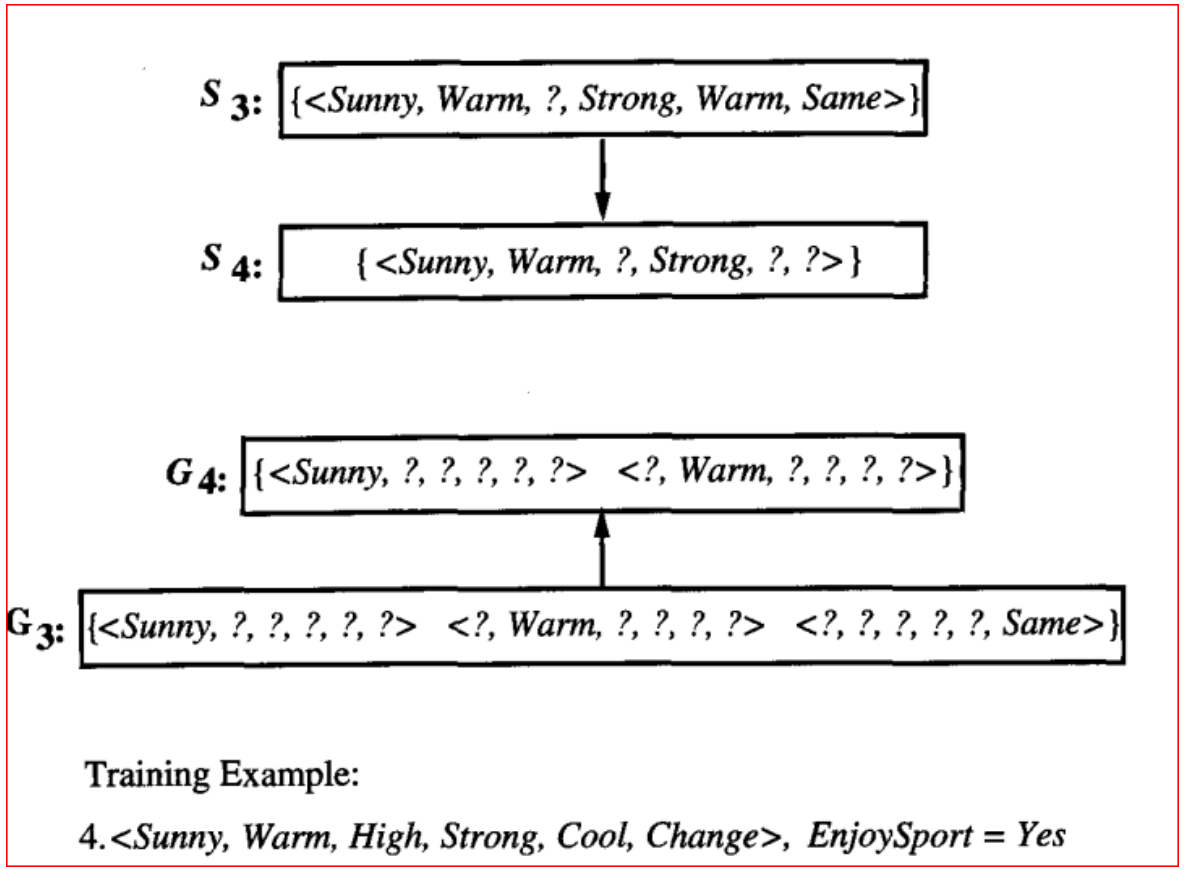
- **Training example 3 is a negative example that forces G_2 boundary to be more specialized.**
- **Maximally general hypothesis are included in G_3.**



CANDIDATE ELIMINATION METHOD – PROBLEM (CONTD..)

STEP 3 (CANDIDATE ELIMINATION TRACE 3) :

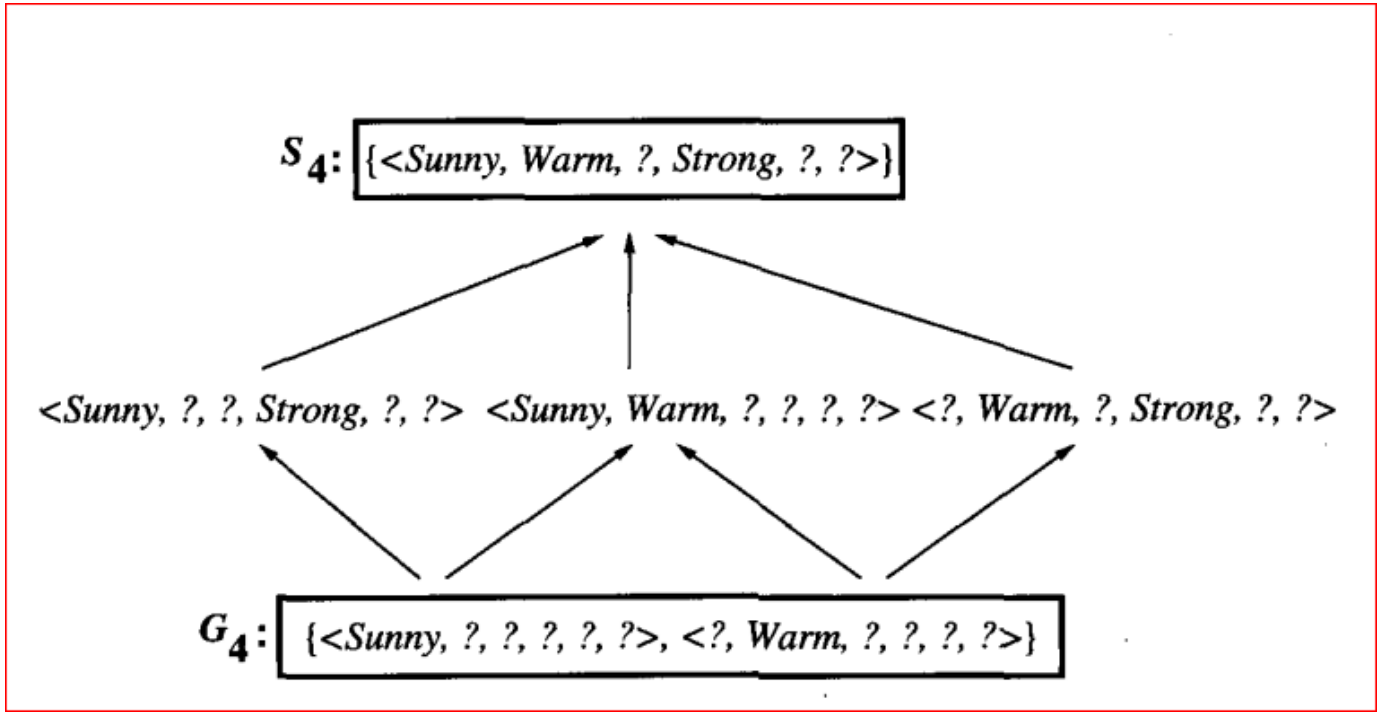
- **Training example 4 is a positive example, positive training example generalize the S boundary from S_3 to S_4**
- **One member of G_3 must also be deleted, because it is no longer more general than the S_4 boundary.**



CANDIDATE ELIMINATION METHOD – PROBLEM (CONTD..)

FINAL STEP (FINAL VERSION SPACE)

The final version space for the EnjoySport concept learning problem.



DATA PREPROCESSING

- Data cleaning :
  - Handling Missing data – Data imputation
  - Handling Noisy data – Data filtration
- Data Transformation :
  - One Hot Encoding
  - Binning
  - Normalization
  - Standardization
- Data Reduction
  - Feature Selection
  - Feature Extraction

CLEANING

- Handling Missing Values – Data Imputation
  - Replacing the missing values with mean or median.
  - Replacing the missing values with constants.
  - Replacing the missing values with information from other columns.
- Handling Noisy Data – Denoising or filtering (Removes outliers)
  - Smoothing
  - Binning

DATA SCALES

- **Nominal Scale :** Named Variables
- **Ordinal Scale:** Named + Ordered Variables
- **Interval Scale :** Named + Ordered + Proportionate Intervals between variables
- **Ratio Scale :** Named + Ordered + Proportionate Interval between Intervals + accommodate absolute zero

TRANSFORMATION

- The problem of transforming raw data into dataset is called feature engineering.

One Hot Encoding	Transformation of categorical feature into several binary codes is called One Hot Encoding . Example : Categorical Feature “Colors” with three possible values : Red, Yellow and Green, Transform this feature into a vector of three numerical values. Red = [1,0,0] Yellow = [0,1,0] Green = [0,0,1]
Binning	Transformation of numerical feature into categorical one. Binning is also called bucketing is the process of converting a continuous feature into multiple binary features called bines or buckets.
Normalization	Process of converting an actual range of values which a numerical feature can take into a standard range of values, typically in the interval [-1,1] or [0,1]
Standardization	Z-score Normalization is the procedure during which the feature values are rescaled so that they have the properties of a standard normal distribution with mean = 0 and standard deviation =1 .

DIMENSIONALITY REDUCTION

- There are two methods for dimensionality reduction :

METHODS	DESCRIPTION
Feature Selection	Finding K of the d dimension which gives more information and discard (d-k) dimension. Example : Subset Feature Selection
Feature Extraction	Finding new set of k dimensions which is a combination of original d dimensions. **Examples :** **Linear Projection Methods :** • Principal Component Analysis (PCA) – Unsupervised Learning • Linear Discriminant Analysis – Supervised Learning