Lecture: Information Theory in Pattern Recognition

Dr. Amit Ranjan

School of Computer Science

Basics Concepts of Information Theory

Entropy

Definition:

Entropy is a measure of uncertainty or randomness in a probability distribution. It quantifies the amount of information required to describe the state of a system. The higher the entropy, the greater the disorder or uncertainty.

Mathematical Formulation:

For a discrete random variable X with probability distribution $P(X)$, the entropy $H(X)$ is defined as:
$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$ where $P(x_i)$ is the probability of the event x_i.

Interpretation:

- If entropy is high, the system is more unpredictable.
- If entropy is low, the system has less uncertainty (e.g., a biased coin has lower entropy than a fair coin).

Basics Concepts of Information Theory

Example:

Consider a fair coin flip, where the probability of heads (H) and tails (T) are both 0.5.

$$H(X) = -(0.5 \log_2 0.5 + 0.5 \log_2 0.5)$$

$$H(X) = -(0.5 \times -1 + 0.5 \times -1) = 1$$

Thus, the entropy of a fair coin toss is 1 bit.

Now, consider a biased coin where the probability of heads is 0.8 and tails is 0.2:

$$H(X) = -(0.8 \log_2 0.8 + 0.2 \log_2 0.2)$$

$$H(X) = -(0.8 \times -0.3219 + 0.2 \times -2.3219) = 0.72$$

The entropy here is lower because the outcome is more predictable.

Basics Concepts of Information Theory

Mutual Information

Definition:

Mutual information (MI) quantifies the reduction in uncertainty of one random variable due to the knowledge of another. It measures the dependency between two variables.

Mathematical Formulation:

For two discrete random variables X and Y, the mutual information $I(X;Y)$ is given by: $I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$ where $P(x,y)$ is the joint probability distribution and $P(x), P(y)$ are the marginal distributions.

Properties:

- Mutual information is always non-negative: $I(X;Y) \geq 0$
- $I(X;Y) = 0$ if and only if X and Y are independent.
- Mutual information is symmetric: $I(X;Y) = I(Y;X)$

Example:

Consider a dataset where X represents symptoms and Y represents disease presence. If knowing the symptom significantly reduces the uncertainty about the disease, then $I(X;Y)$ will be high. For example, if all patients with a fever (X) have flu (Y), then the mutual information is maximized.

Applications in Pattern Recognition

Feature Selection Using Information Theory

Feature selection is crucial in pattern recognition to remove redundant and irrelevant features, improving model performance. Information-theoretic methods use entropy and mutual information for feature selection.

Mutual Information for Feature Selection

Mutual information can quantify how much information a feature provides about the class label. The goal is to select features that maximize mutual information with the target variable.

Mathematical Formulation: Given a feature set F and class variable C, the relevance of a feature X can be measured as: $I(X;C)$ A commonly used feature selection criterion is the **Max-Relevance and Min-Redundancy (mRMR)** approach:

$$\max \sum_{X_i \in S} I(X_i; C) - \lambda \sum_{X_i, X_j \in S} I(X_i; X_j)$$

where S is the selected feature subset, and λ controls redundancy.

Example: In a spam email classification task, words like "free," "win," and "offer" may have high mutual information with the spam label, making them useful features.

Applications in Pattern Recognition

Information Gain

Information gain measures the reduction in entropy after splitting a dataset based on a feature. It is used in decision tree algorithms like ID3 and C4.5.

$$IG(X) = H(Y) - H(Y|X)$$

where $H(Y)$ is the entropy of the target variable, and $H(Y|X)$ is the conditional entropy after splitting on feature X.

Example: If we split a dataset of patients based on "fever" (yes/no), and it significantly reduces uncertainty about "flu" presence, the feature has high information gain.

Applications in Pattern Recognition

Clustering Using Information Theory

Clustering involves grouping data points based on similarity. Information theory helps in evaluating clustering quality using entropy-based measures.

Mutual Information for Clustering Evaluation

Mutual information can measure the similarity between a clustering result and a ground truth classification.

Normalized Mutual Information (NMI): $NMI(X,Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$ where X and Y are the predicted and actual clusters.

Example: If a clustering algorithm groups patients based on symptoms and the result closely aligns with actual disease categories, NMI will be high.

Minimum Description Length (MDL) Principle

The MDL principle suggests choosing the model that provides the best compression of the data.

- A good clustering minimizes within-cluster entropy while maintaining meaningful groupings.
- This is useful for selecting the optimal number of clusters.

Example: If clustering DNA sequences, an MDL-based method might select the number of clusters that best compresses genetic variations.