Case Study-1

Concept Learning

Introduction:

- Learning involves acquiring general concepts from specific training examples. Example: People continually learn general concepts or categories such as "bird," "car," "situations in which I should study more in order to pass the exam," etc.
- Each such concept can be viewed as describing some subset of objects or events defined over a larger set
- Alternatively, each concept can be thought of as a Boolean-valued function defined over this larger set. (Example: A function defined over all animals, whose value is true for birds and false for other animals).

Definition: Concept learning- Inferring a Boolean-valued function from training examples of its input and output.

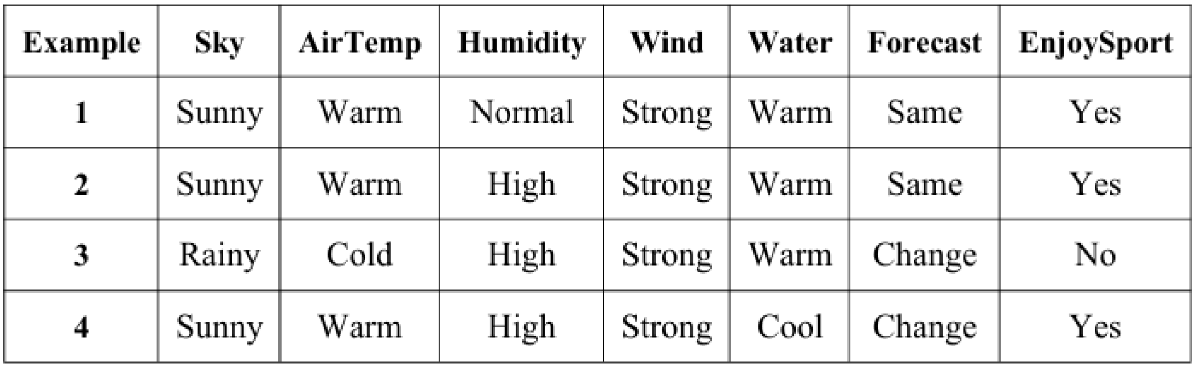
Objective

- Learn a hypothesis h*h* in the form of a conjunction of attribute constraints that matches the target concept c*c* (Enjoy Sport) for all instances.
- Achieve perfect classification accuracy on training data while generalizing to unseen examples within the defined hypothesis space.

Theory:

ACONCEPT LEARNING TASK

Consider the example task of learning the target concept "Days on which Aldo enjoys his favorite water sport”**.**



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

Table: Positive and negative training examples for the target concept

Enjoy Sport. The task is to learn to predict the value of Enjoy Sport for an arbitrary day, based on the values of its other attributes

Given:

- Instances X: Possible days, each described by the attributes
  - Sky (with possible values Sunny, Cloudy, and Rainy),
  - Air Temp (with values Warm and Cold),
  - Humidity (with values Normal and High),
  - Wind (with values Strong and Weak),
  - Water (with values Warm and Cool),
  - Forecast (with values Same and Change).
- Hypotheses H: Each hypothesis is described by a conjunction of constraints on the attributes Sky, Air Temp, Humidity, Wind, Water, and Forecast. The constraints may be "?" (any value is acceptable), “Φ” (no value is acceptable), or a specific value.

- Target concept c: Enjoy Sport: X → {0, l}
- Training examples D: Positive and negative examples of the target function

- Determine:
- A hypothesis h in H such that h(x) = c(x) for all x in X. Table: The Enjoy Sport concept learning task.

Table: The Enjoy Sport concept learning task.

Limitations

- Assumes noise-free training data and a hypothesis space HH that contains the target concept, limiting real-world applicability.
- Algorithms like Find-S produce overly specific hypotheses and cannot represent uncertainty when multiple consistent hypotheses exist.

Applications

- Binary classification tasks like spam detection or medical diagnosis where decision rules are derived from labeled attribute sets.
- Educational systems for designing personalized learning paths based on student attribute patterns (e.g., performance, engagement).

Conclusion

- Concept learning frameworks like Enjoy Sport demonstrate how inductive inference generalizes from examples but are constrained by hypothesis space design.
- Effective for well-defined, discrete attribute spaces but struggles with continuous data, noise, and evolving concepts.

Case Study-2

Inductive Bias in Decision Tree Learning

Introduction:

Inductive bias is the set of assumptions that, together with the training data, deductively justify the classifications assigned by the learner to future instances.

Objective:

- Select the shortest decision tree consistent with training data by prioritizing attributes that maximize information gain.
- Efficiently approximate optimal hypotheses through heuristic search, avoiding exhaustive breadth-first exploration

Theory:

Given a collection of training examples, there are typically many decision trees consistent with these examples. Which of these decision trees does ID3 choose?

ID3 search strategy

- Selects in favour of shorter trees over longer ones
- Selects trees that place the attributes with highest information gain closest to the root.

Approximate inductive bias of ID3: Shorter trees are preferred over larger trees

- Consider an algorithm that begins with the empty tree and searches breadth first through progressively more complex trees.
- First considering all trees of depth 1, then all trees of depth 2, etc.
- Once it finds a decision tree consistent with the training data, it returns the smallest consistent tree at that search depth (e.g., the tree with the fewest nodes).

• Let us call this breadth-first search algorithm BFS-ID3.

• BFS-ID3 finds a shortest decision tree and thus exhibits the bias "shorter trees are preferred over longer trees.

A closer approximation to the inductive bias of ID3: Shorter trees are preferred over longer trees. Trees that place high information gain attributes close to the root are preferred over those that do not.

• ID3can be viewed as an efficient approximation to BFS-ID3, using a greedy heuristic search to attempt to find the shortest tree without conducting the entire breadth-first search through the hypothesis space.

• Because ID3 uses the information gain heuristic and a hill climbing strategy, it exhibits a more complex bias than BFS-ID3.

• In particular, it does not always find the shortest consistent tree, and it is biased to favour trees that place attributes with high information gain closest to the root.

Applications

- Medical diagnosis: Building interpretable decision rules from patient attributes to classify diseases.
- Customer segmentation: Identifying key demographic or behavioural attributes for targeted marketing.

Conclusion

- ID3's preference for concise, high-information-gain trees enables practical classification but risks overfitting and struggles with continuous data.
- While effective for discrete attributes, its heuristic bias trades optimality for computational efficiency, limiting robustness in noisy or complex domains.

Case Study-3

Basics of Sampling Theory

Introduction

Sampling theory is a fundamental branch of statistics focused on how to select and analyse a subset (sample) from a larger group (population) to make valid inferences about the whole. Instead of collecting data from every member of a population—which is often impractical or impossible—sampling enables researchers to gather, analyse, and generalize findings efficiently and cost-effectively

Objective

- Estimate the true binomial proportion p*p* by analysing sample error rates across multiple trials.
- Quantify uncertainty using confidence intervals or maximum likelihood estimators to bound estimation errors.

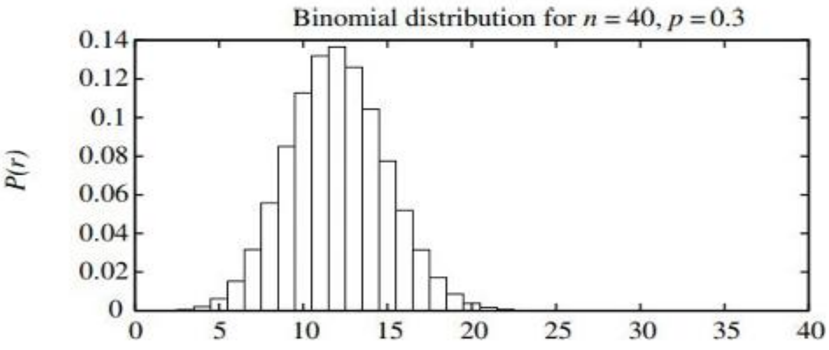
Theory

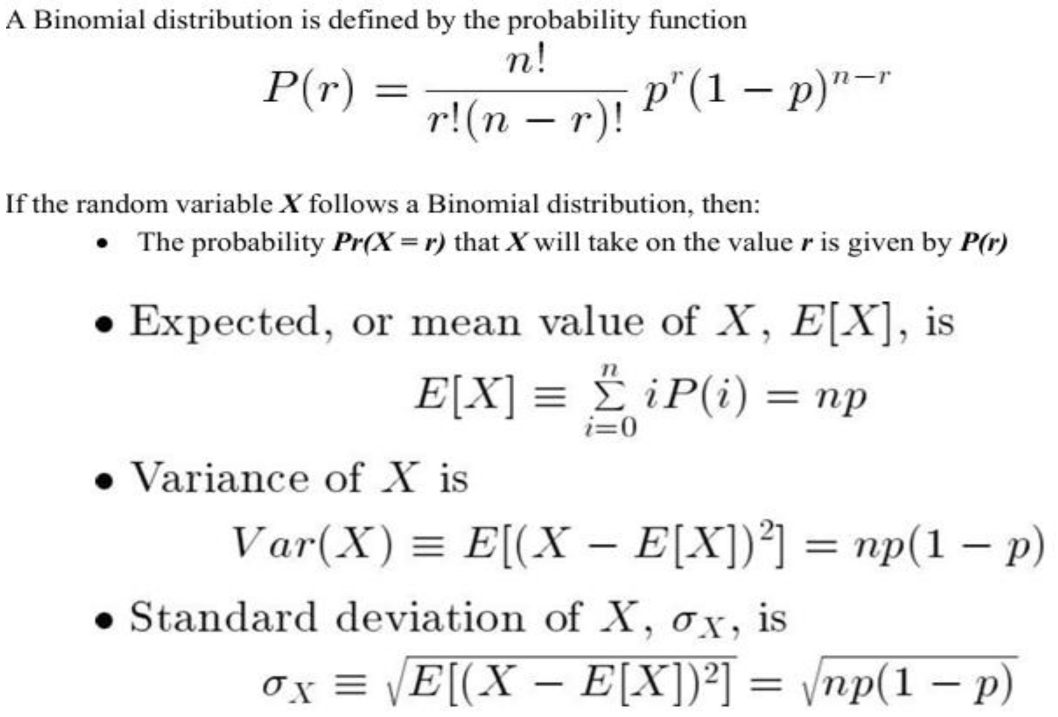
Error Estimation and Estimating Binomial Proportions

- Collect a random sample S of n independently drawn instances from the distribution D, and then measure the sample error errors(h). Repeat this experiment many times, each time drawing a different random sample Si of size n, we would expect to observe different values for the various errors i(h), depending on random differences in the makeup of the various Si. We say that errors i(h), the outcome of the ith such experiment, is a random variable**.**

- Imagine that we were to run k random experiments, measuring the random variables errors 1(h), errors 2(h) . . . errors sk(h) and plotted a histogram displaying the frequency with which each possible error value is observed.

- As **k** grows, the histogram would approach a particular probability distribution called the Binomial distribution which is shown in below figure.





A Binomial distribution is defined by the probability function

$$P(r) = \frac{n!}{r!(n-r)!}\, p^r (1-p)^{n-r}$$

If the random variable ***X*** follows a Binomial distribution, then:

- The probability ***Pr(X = r)*** that ***X*** will take on the value ***r*** is given by ***P(r)***

- Expected, or mean value of X, $E[X]$, is

$$E[X] \equiv \sum_{i=0}^{n} iP(i) = np$$

- Variance of X is

$$Var(X) \equiv E[(X - E[X])^2] = np(1-p)$$

- Standard deviation of X, σ_X, is

$$\sigma_X \equiv \sqrt{E[(X - E[X])^2]} = \sqrt{np(1-p)}$$

Applications

- Quality control: Monitoring defect rates in manufacturing by modeling pass/fail outcomes as binomial trials.
- Survey analysis: Estimating population proportions (e.g., election polls) with confidence intervals derived from sample data.

Conclusion

- Binomial error estimation provides tractable inference for binary outcomes but requires large samples for reliable confidence intervals.
- Modern data-driven methods enhance robustness by adapting to unknown correlations without prior complexity assumptions.

Case Study-4

Maximum Likelihood and Least Squared Error Hypotheses

Introduction

Learning continuous-valued target functions, such as in neural networks or linear regression, often relies on probabilistic frameworks like Maximum Likelihood Estimation (MLE) to fit models to noisy data.

Objective

Minimize the sum of squared errors between predicted and observed values, thereby finding the hypothesis that maximizes the likelihood of the observed data under the assumed noise model.

Theory

Consider the problem of learning a continuous-valued target function such as neural network learning, linear regression, and polynomial curve fitting.

A straightforward Bayesian analysis will show that under certain assumptions any learning algorithm that minimizes the squared error between the output hypothesis predictions and the training data will output a maximum likelihood (ML) hypothesis

Learner L considers an instance space X and a hypothesis space H consisting of some class of real-valued functions defined over X, i.e., $(\forall h \in H)$ [h: X → R] and training examples of the form

- The problem faced by L is to learn an unknown target function f: X →R

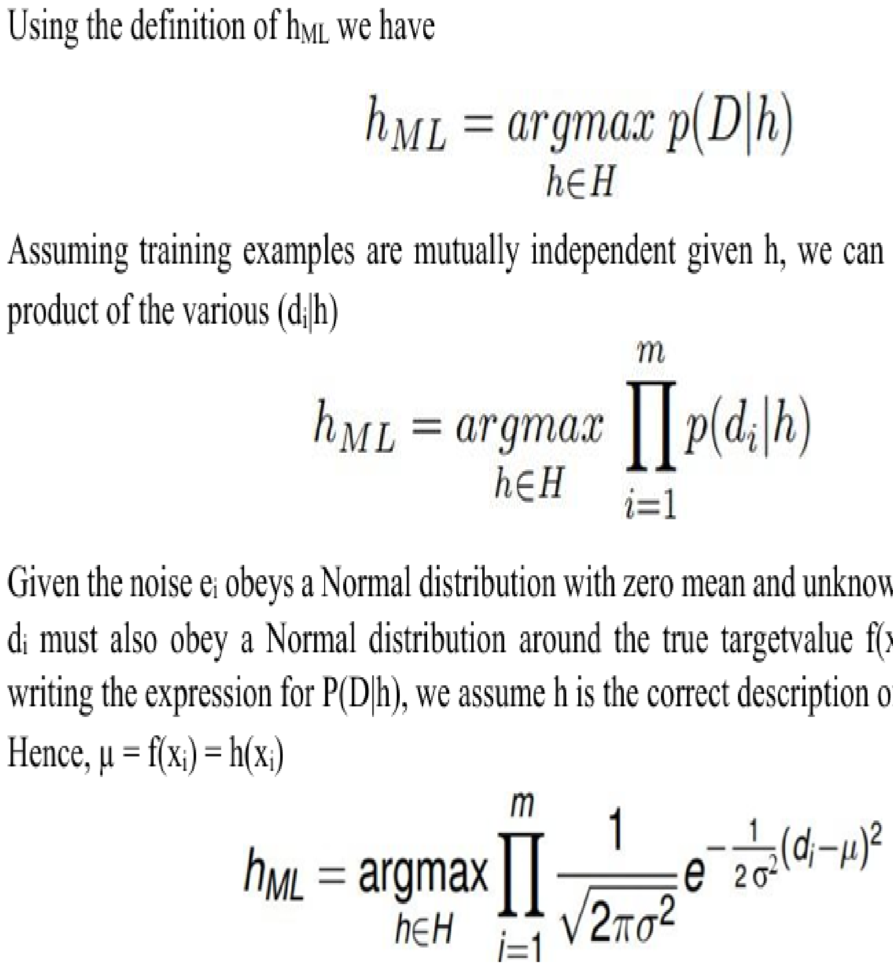
- Aset of m training examples is provided, where the target value of each example is corrupted by random noise drawn according to a Normal probability distribution with zero mean (di = f(xi) +ei)

- Each training example is a pair of the form (xi,di) where di = f (xi) + ei.

Here f(xi) is the noise-free value of the target function and ei is a random variable representing the noise.

Itis assumed that the values of the ei are drawn independently and that they are distributed according to a Normal distribution with zero mean.

● The task of the learner is to output a maximum likelihood hypothesis or a MAP hypothesis assuming all hypotheses are equally probable a priori.



Using the definition of h_{ML} we have

$$h_{ML} = \underset{h \in H}{argmax}\ p(D|h)$$

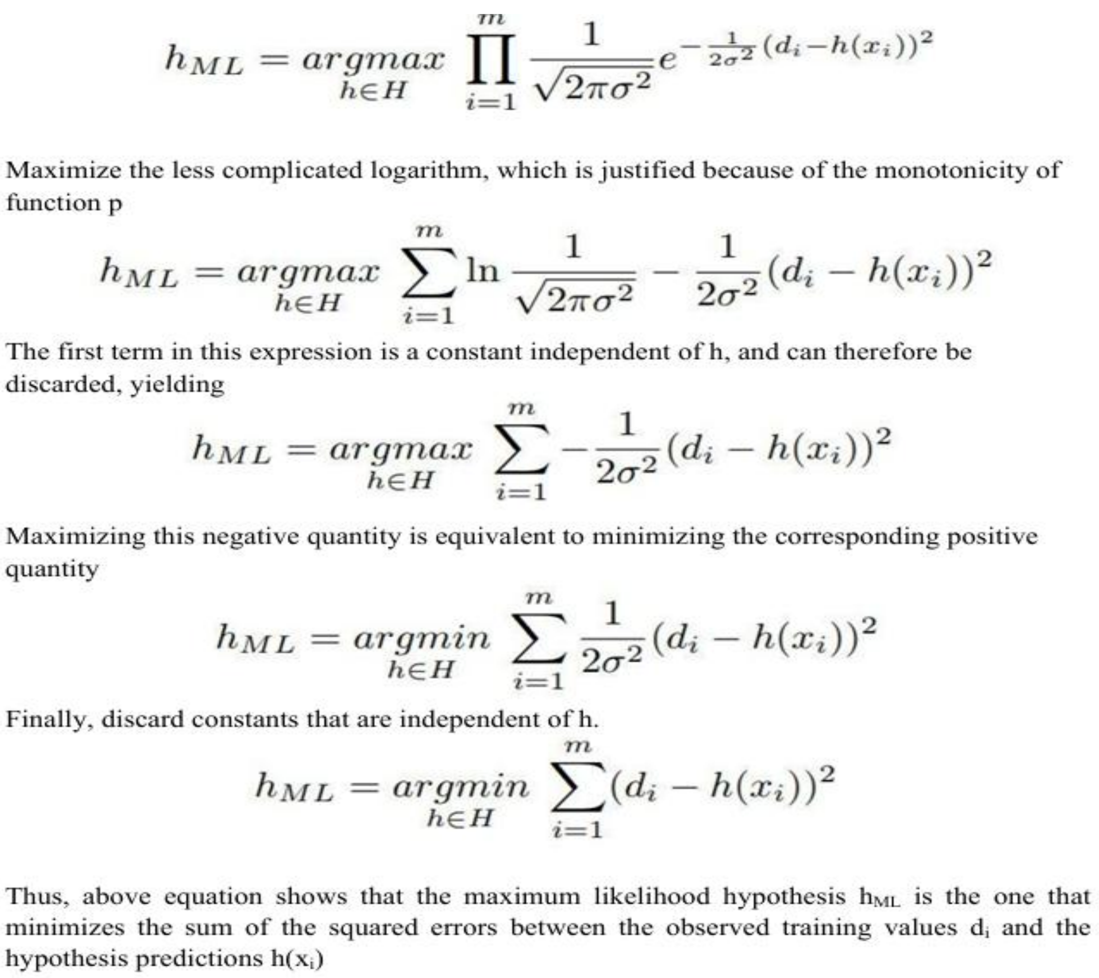
Assuming training examples are mutually independent given h, we can write P(D|h) as the product of the various $(d_i|h)$

$$h_{ML} = \underset{h \in H}{argmax} \prod_{i=1}^{m} p(d_i|h)$$

Given the noise e_i obeys a Normal distribution with zero mean and unknown variance σ^2, each d_i must also obey a Normal distribution around the true targetvalue $f(x_i)$. Because we are writing the expression for P(D|h), we assume h is the correct description of f.

Hence, $\mu = f(x_i) = h(x_i)$

$$h_{ML} = \underset{h \in H}{\text{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i-\mu)^2}$$



$$h_{ML} = \underset{h \in H}{argmax} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i - h(x_i))^2}$$

Maximize the less complicated logarithm, which is justified because of the monotonicity of function p

$$h_{ML} = \underset{h \in H}{argmax} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

The first term in this expression is a constant independent of h, and can therefore be discarded, yielding

$$h_{ML} = \underset{h \in H}{argmax} \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

Maximizing this negative quantity is equivalent to minimizing the corresponding positive quantity

$$h_{ML} = \underset{h \in H}{argmin} \sum_{i=1}^{m} \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

Finally, discard constants that are independent of h.

$$h_{ML} = \underset{h \in H}{argmin} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

Thus, above equation shows that the maximum likelihood hypothesis h_{ML} is the one that minimizes the sum of the squared errors between the observed training values d_i and the hypothesis predictions $h(x_i)$

Applications

Widely used in regression analysis, neural network training, and curve fitting for predicting real-valued outcomes in fields like finance, engineering, and healthcare.

Conclusion

MLE provides a principled, statistically sound approach for parameter estimation in machine learning models, ensuring optimal fit to data under reasonable assumptions about noise and distribution. This method is particularly effective when the data follows a well-defined distribution, such as the normal distribution, allowing for precise modeling and prediction. Additionally, MLE's reliance on maximizing the likelihood function ensures that the estimated parameters are those most likely to have generated the observed data, providing a robust basis for inference and decision-making.