FOUNDATIONS OF ALGORITHM DESIGN AND MACHINE LEARNING

- Prof. Aritra Hazra

Topic – Core Learning Principles

APRIL 17, 2025

SUBMITTED BY:

Vikash Kumar Gupta (24BM6JP60)

Learning Principles

1. **Occam Razor**

   a. **Simplest Plausible Hypothesis**

   Occam's Razor is a principle that suggests:

   Among competing hypotheses that explain the data equally well, the one with the fewest assumptions (i.e., the simplest) should be selected.

   This means preferring the simplest model that still performs well on the training data, i.e.,

   $$E_{\text{in}} \approx 0$$

   b. **What is meant by 'SIMPLE'?**

   'Simplicity' can be interpreted in several mathematical ways in ML:

   - *Less degree of polynomial*: A linear model (e.g. $y = w \cdot x + b$) is simpler than a high-degree polynomial.
   - *Shorter bit size* : According to the Minimum Description Length (MDL) principle, the best hypothesis compresses the data the most.
     $\rightarrow$ Simplicity = compressibility
   - *Lower VC-dimension*: Vapnik-Chervonenkis dimension is a measure of a model's capacity to overfit. Simpler models have a lower VC dimension.

   c. **Why simple is better?**

   We generally tend to believe that simple is better, i.e.,

   $$E_{\text{val}} \approx 0$$

   - *Generalization*: Simpler models are less likely to overfit the training data and tend to generalize better to unseen data.
   - *Fewer assumptions = fewer ways to be wrong*: Each additional parameter or assumption introduces potential for error.
   - *Efficient computation*: Simpler models require less data and less computational resources to train and infer.

2. **Sampling Bias**

   a. **Survey of Sample**

   Sampling bias occurs when the sample used to train or evaluate a model is not representative of the population. This leads to models that perform poorly on real-world data.

   Example – Truman vs. Dewey Election (1948):

   A famous case of sampling bias occurred in the 1948 US Presidential election.

   Polls conducted before the election predicted a landslide victory for Thomas Dewey over Harry Truman. However, Truman won.

What went wrong? Surveys were conducted by telephone, but at that time, the telephones were mainly owned by wealthy households. As a result, the sample was biased towards richer individuals who were more likely to support Dewey.

Lesson: We must ensure that our sample reflects the distribution of the entire population. The training data must come from the same distribution as the real-world data that the model will encounter.

b. **Partition of Test/Train**

It's a standard practice to split the dataset into training and testing sets to evaluate a model's generalization.

A key point is:

Always normalize (or standardize) data after splitting into train/test sets.

Why?

- If normalization (like scaling to zero mean and unit variance) is done before splitting, the statistics (mean, std) of the test set leak into the training process, violating the principle of isolation.
- This leads to data leakage, and the test set is no longer an unbiased measure of generalization.
- Instead, calculate normalization parameters only from the training set, and apply the same transformation to the test set.

c. **Data Imbalance**

Data imbalance occurs when classes in a classification problem are not represented equally.

For instance, in a binary classification setting:

- Class 0 (no disease): 95%
- Class 1 (disease): 5%

This kind of imbalance creates multiple issues:

- **Accuracy paradox:** A model might predict the majority class all the time and still achieve high accuracy, while ignoring the important minority class.
- **Biased learning:** The model tends to ignore the minority class due to its scarcity in the training data.

Real-world Example – Credit Card Application Bias:

In the past, banks only stored data about applicants whose credit card applications were accepted (i.e., positive class only).

So when a new application came in, the model would only consider past positive examples and try to find similar ones.

As a result, potentially good applicants who looked slightly different from historical positives were overlooked.

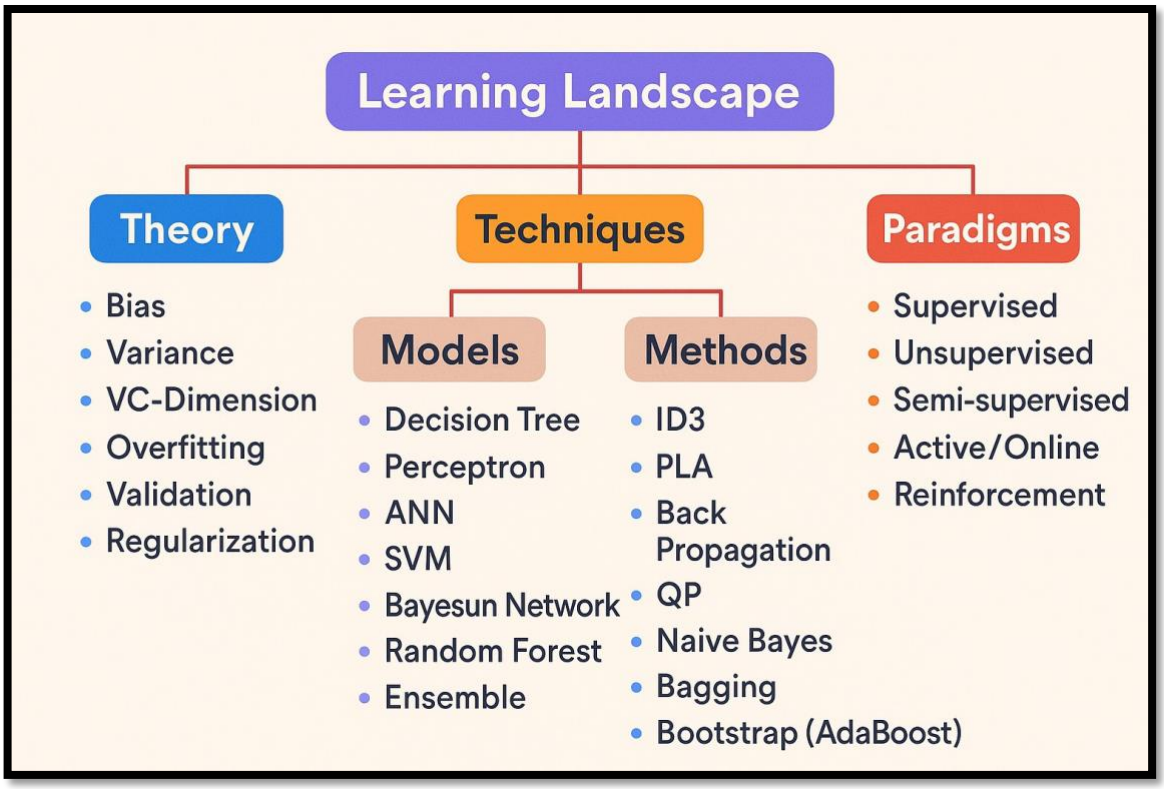
This sampling bias caused the bank to miss valuable customers, who were then targeted by other banks with more representative datasets (that is, datasets that included accepted and rejected applicants).

Ways to handle:

- *a. Resampling Techniques*
  - *Undersampling* the majority class to balance the dataset.
  - *Oversampling* the minority class by duplicating or synthetically generating data.
- *b. Use of Better Metrics*
  Instead of plain accuracy, use:
  - *Precision*
  - *Recall*
  - *F1-score*

– *ROC-AUC*

- *c. Data Augmentation*
  Data augmentation involves generating new training examples by applying transformations to the existing ones.
  By making the *minority class more diverse*, the model becomes more robust and better at generalization.



The Evolution of "Free" in the Digital Age

Over the past decades, each technological era has made a fundamental capability effectively *free* — opening new possibilities while also introducing new challenges. Here's a glimpse into that journey:

- **1950–1980: Free Computation**
  The early computing revolution drastically reduced the cost of processing power. Tasks that once required massive machines became possible on desktops and microcontrollers.
- **1965–1990: Free Storage**
  With exponential drops in the price of memory and hard drives, storing large volumes of data became cheap and routine — paving the way for data-driven innovation.
- **1991–1997: Free Internet**
  The Internet connected the world and democratized access to information. Anyone, anywhere, could now access vast knowledge bases at minimal or no cost.
- **1990s–2010: Free Communication**
  The rise of email, instant messaging, and social media allowed instant, global communication

— for individuals, businesses, and entire nations.

Together, Free Internet and Free Communication fueled the rise of Big Data: The constant generation, transmission, and storage of digital interactions created vast datasets — enabling breakthroughs in analytics and machine learning.

- **2010–2024: Free Reality**
  Technologies like Augmented Reality (AR), Virtual Reality (VR), and the Metaverse blurred the lines between physical and digital experiences.

  But Free Reality came with its perils:

  The same tools that power immersive learning and virtual collaboration also enabled *deepfakes* — hyper-realistic synthetic media that can mislead, manipulate, or impersonate individuals.

  This raised urgent concerns about **truth, identity, and digital ethics** in an era where seeing is no longer believing.

- **2018–Present: Free Intelligence**
  With the rise of AI platforms and generative models, cognitive tools (like language understanding, vision, and reasoning) became accessible to everyone.

  What once required a team of experts can now be done with a prompt — leading to a new wave of innovation across industries.