Terminology

1 Agnostic Learning

Agnostic learning, also sometimes referred to as "agnostic machine learning," is a theoretical framework in machine learning that makes fewer assumptions about the data or the function to be learned. The term "agnostic" is used to indicate that the learning algorithm **does not know or "believe" in advance the form or structure of the target function or the distribution of the data.**

1.1 Assumptions

Unlike traditional statistical learning models that may assume a specific form for the data distribution or the target function, agnostic learning makes minimal assumptions. It is designed to be robust to unknown or adversarial data distributions.

1.2 PAC Learning

Agnostic learning extends the Probably Approximately Correct (PAC) learning framework. In the PAC model, learning algorithms aim to find a hypothesis that is probably approximately close to the true function. The agnostic extension allows for the possibility that no hypothesis may be a close approximation, and instead seeks the hypothesis that performs best among the class, even if none are particularly close to the true function.

1.3 Robustness

Because of its fewer assumptions, agnostic learning can be more robust to model misspecification and can be applied to a wider range of problems, including those with noisy or complex data distributions.

2 Bayesian Statistics

Bayesian statistics is an approach to statistical inference that is grounded in Bayes' theorem. It provides a probabilistic framework for updating beliefs in the light of new evidence. In Bayesian statistics, probability is interpreted as **a degree of belief or confidence in an event occurring**, which contrasts with the frequentist approach where probability is defined through long-run frequencies of events.

2.1 Baye's Theorem

This theorem is the cornerstone of Bayesian statistics, which states that the posterior probability of a hypothesis is **proportional to its prior probability and the likelihood of the current evidence** given that hypothesis.

2.2 Prior Probability

The prior reflects the beliefs about the value of a parameter before any new data is considered. It represents what is known or believed about the parameter prior to observing the current data.

2.3 Posterior Probability

The posterior is the updated belief about the parameter after taking into account the new evidence (data). It combines the prior and the likelihood of the observed data.

3 Bias (Error)

In the context of model prediction error, bias refers to the error introduced by approximating a real-world problem, which may be complex and non-linear, with a simplified model.

1. **High Bias**: Models with high bias pay very little attention to the training data and oversimplify the model, which can't capture the underlying trends. This is often associated with models that are too simple (linear models trying to capture non-linear relationships).

2. **Low Bias**: Models with low bias are usually more complex and fit the training data very closely. They are likely to capture the true relationship between features and the outcome variable better.

4 Convexity

Convexity in machine learning is a property of certain mathematical functions and sets that is central to the field of optimization, which is foundational for training machine learning models. A convex function is one where a line segment drawn between any two points on the function's graph **does not intersect the graph at any point other than the endpoints**.

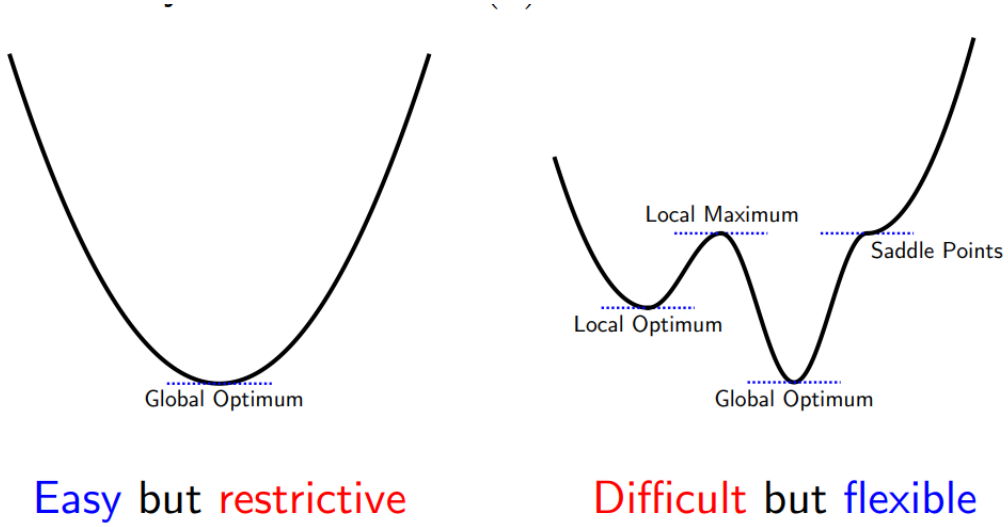

Figure 1: Convexity

As shown in the graph, the LHS is **convex**, and the RHS is **non-convex**. **Convex optimization** are 'nice' in the sense that they are easier to solve than non-convex problems because they have a global minimum that optimization algorithms can find reliably and efficiently. **Non-convex optimization** can have multiple local minima and saddle points, making it difficult to find the global minimum, but the model will be more flexible.

5 Features

A feature is a measurable property or characteristic of a phenomenon being observed. In the context of a dataset, features are often referred to as columns or variables. Features are fundamental to the performance of machine learning algorithms. They directly influence the predictive models by providing the algorithm with input data that can be used to infer patterns from and make decisions.

5.1 Feature Selection

Feature selection is the process of identifying and selecting a subset of input variables that are most relevant to the target variable. Good feature selection can improve model accuracy, reduce overfitting, and decrease computational cost.

5.2 Feature Engineering

Feature engineering is the process of using domain knowledge to create new features from raw data that make machine learning algorithms work. This can involve combining features, transforming features, or extracting new information from existing data.

5.3 Normalization and Standardization

Features often need to be normalized or standardized so that they have a similar scale. This is important for models that are sensitive to feature scale, like SVMs or k-nearest neighbors.

5.4 Dimensionality Reduction

Techniques like Principal Component Analysis (PCA) can reduce the number of features by transforming them into a smaller set that still captures most of the variability in the data.

6 Maximum Likelihood Estimation

6.1 Likelihood

Likelihood is a measure of the plausibility of a statistical model parameter given the observed data. It is not the probability of the data itself, but rather the probability of the data given specific parameter values of the model.

Likelihood function is a function of the parameters given fixed data. It quantifies how well different parameter values explain the data that you have observed. Mathematically, for a set of data $\mathbf{X}$ and a statistical model with parameters θ, the likelihood function is defined as the probability of the observed data given the parameters:

$$L(\theta|\mathbf{X}) = P(\mathbf{X}|\theta) \tag{1}$$

For **independent and identically distributed (i.i.d.)** data points, the likelihood is often the product of the probabilities of individual data points.

In practice, it's common to work with the natural logarithm of the likelihood function, known as the **log-likelihood**. This is because the log transformation turns products into sums (making calculations easier), and it preserves the location of the maximum (since logarithm is a monotonically increasing function)

6.2 MLE

In MLE, the objective is to find the parameter values that maximize the likelihood function. The estimated values are those for which the observed data is most probable:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta|\mathbf{X}) \tag{2}$$

7 Response

Response typically refers to the output or target variable that a model is trying to predict or explain. It is the dependent variable in the context of a dataset or a learning algorithm. The response variable could be the result of a classification process (such as 'spam' or 'not spam'), a continuous value in a regression task (like the price of a house),

or the next sequence in a pattern for sequence prediction tasks (such as the next word in a sentence)

8 Risk

In machine learning, the term "risk" often denotes the expected loss or cost of a particular model with respect to a given loss function. It is a measure of how predictions from the model deviate from the true outcomes.

8.1 Empirical Risk

This is the average loss over the training sample. It is what machine learning models typically minimize directly during the training process.

8.2 Bayes Risk

The **lowest possible risk** that can be achieved by any model. This is also known as the irreducible error and represents the best possible performance given the inherent noise in the data.

8.3 Excessive Risk

It quantifies how much more loss a particular model incurs in comparison to the optimal model that achieves the Bayes risk:

$$\text{Excessive Risk} = \text{Empirical Risk} - \text{Bayes Risk} \tag{3}$$

Minimizing excessive risk is important because it directly relates to improving a model's performance on future, unseen data. A model with lower excessive risk is expected to generalize better.