Module 5: Generalisation Theory and the Bias–Variance Trade-Off

Learning outcomes

1. Apply probabilistic reasoning to evaluate candidate functions based on limited data.
2. Apply probability approaches to handle uncertainty and make predictions based on data.
3. Evaluate how stochastic assumptions affect prediction accuracy.
4. Analyse the feasibility of machine learning in relation to its three essential conditions.
5. Select the model that best represents the relationship between variables for a given goal.
6. Apply the 'training set–validation set–test set approach' to design an ML workflow.
7. Evaluate the feasibility of a machine learning problem by evaluating the assumptions made about the data set.
8. Analyse how tuning hyperparameters on one data set impacts the model results.

Generalisation and feasibility

- Learning from data requires assumptions. Without assumptions, generalisation is not possible.
- The ‘problem of induction’ highlights the limits of learning from repeated experience alone.
- According to the ‘no free lunch’ theorem, no algorithm performs best across all problems.
- For learning to be feasible, three conditions must be met:
  - A probabilistic setting: typically, all samples are independent and identically distributed.
  - Stationarity: the future data is drawn according to the same rules as the data you have seen already.
  - A priori knowledge: the function f to be learned is not completely arbitrary.

Probabilistic approaches in ML

Real-world data includes noise and uncertainty. The following probabilistic approaches can be used to make informed predictions.

- **Bayesian inference:** updates beliefs using prior knowledge and observed data

$$P(\theta \mid D) = \frac{p(D|\theta)P(\theta)}{P(D)}$$

- **Laplace's rule of succession:** adds pseudo-counts to smooth probabilities in small samples

$$P(X\_n+1 = 1|X_1 + \cdots + X_n = s) = \frac{s+1}{n+2}$$

- **Frequentist approach:** probability based on long-run frequencies

$$P_{freq} = \frac{s}{n}$$

- **Maximum likelihood estimation (MLE):** chooses parameters that maximise the likelihood
  $L(\theta; D) = P(D \mid \theta)$

The stochastic perspective

- A stochastic setting assumes that data is drawn randomly from a fixed distribution.
- Feasible learning requires:
  - A true function f that generates the labels.
  - A finite hypothesis space that contains candidate functions. A hypothesis space is a set of all possible models (or hypotheses) that an ML algorithm can learn from a given data set.
  - Randomly drawn data samples that represent the data-generating process.
- Incorrect functions $f_i \neq f$ are assumed to make errors with probability at least ϵ.

Generalisation bound (PAC learning)

- A generalisation bound provides a high-probability guarantee about how well a model trained on finite data will perform on new data.
- **Formula:** $N \geq \frac{\log(\frac{\delta}{H-1})}{\log(1-\epsilon)}$
  where:
  - N = minimum number of samples needed
  - H = number of candidate functions
  - ϵ = error rate of incorrect functions
  - δ = maximum allowable probability of selecting a poor model
- More data reduces the chance that a bad model appears good by random chance.

Bias–variance trade-off

The goal of the bias–variance trade-off is to find the right balance to ensure good generalisation.

- **Bias:** error from assumptions that make the model too simple
- **Variance:** error from sensitivity to fluctuations in the training set

Data splitting strategy

- **Training set:** used to train the model
- **Validation set:** used to tune the model
- **Test set:** used to evaluate performance on unseen data
- The training set–validation set–test set approach improves model selection and evaluation by separating these tasks.
- **Implementing in Python**
  - Use `train_test_split()` `from sklearn.model_selection` to split data.
  - Two-step split for training, validation and test sets:
    - `X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.4)`
    - `X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.5)`
  - Note that you will train on `X_train`, validate on `X_val` and test on `X_test`.

Other key terms

- **Hypothesis space:** all candidate functions considered by the model
- **PAC learning:** 'probably-approximately-correct' framework
- **Overfitting:** model captures noise, not patterns
- **Underfitting:** model is too simple to learn the signal
- **Stationarity:** statistical properties of data remain stable
- **i.i.d. assumption:** data points are independent and identically distributed