Chapter 1 - Basic Concepts

Non-parameteric models:

- Number of parameters grow with the size of the training data
- More flexible but computationally intractable for large datasets
- Example: KNN with a suitable distance metric - the number of pairs to be considered grows as $O(n^2)$.
- Also the idea of a D-dimensional unit cube with uniform data distributed inside. if we want to "grow" a hypercube that contains 10% of the data, what should its edge-length be?
- $V = e^3 \implies e = (0.1)^{\frac{1}{3}} = 0.8$, i.e, a the edge of the hypercube has to be 80% of the edgesize of the original unit cube! Hence, in high dimension, nearest neighbours are not truly near anymore.

Parametric models:

- Fixed number of parameters
- Advantage - faster
- Disadvantage - make strong assumptions about the data distribution (also known as inductive bias)
- Example: Linear regression - assumption that the noise $\epsilon \sim \mathrm{N}(0, 1)$
- In other words linear regression models the conditional distribution $p(y|\mathbf{x}, \theta) = \mathrm{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$
- To go from linear regression to logistic regression
  - The variable y is now modelled as a Bernoulli variable (i.e., with two outcomes instead of continuous outcome space)
  - The output has to be in the range [0,1]
- Logistic Regression: $p(y|\mathbf{x}, \mathbf{w}) = Ber(y|\mu(\mathbf{x}))$ where $\mu(\mathbf{x}) = p(y = 1|\mathbf{x})$, the probability of success. Hence, we model, $\mu(\mathbf{x}) = sigmoid(\mathbf{w}^T\mathbf{x})$

Overfitting:

- To explain overfitting, the best example to use is that of KNN. When K=1, the model has zero error on the training set as it perfectly classifies each training datapoint. But test set error is likely to be high. We can slowly increase K, when K=n, we always predict the majority class (underfitting).
- K-fold cross validation procedure for model selection
- Bias-variance tradeoff: U-shaped curve. Bias (simple model - high bias); Variance (low variance if the model parameters dont change across different subsets of the data - e.g., linear regression line)
- No free lunch: Speed-accuracy-complexity-interpretability tradeoffs