Week 6 - Introduction to Classification

Introduction:

Welcome to this lecture on basic classification models used in statistical modelling, with a particular emphasis on the k-Nearest Neighbors (k-NN) algorithm. Classification is a fundamental task in statistical modelling, where the goal is to assign input data points to predefined categories or classes.

Basic Classification Models:

There are several basic classification models, each with its own characteristics and applications:

1. Logistic Regression:

Logistic regression is a linear model used for binary classification (two classes). It models the probability of an input belonging to one of the two classes using a logistic (S-shaped) curve.

2. Decision Trees:

Decision trees are versatile models that can be used for both binary and multi-class classification. They partition the input space into regions based on a series of if-then-else rules.

3. Naïve Bayes:

Naïve Bayes is a probabilistic model that is particularly useful for text classification and spam detection. It relies on Bayes' theorem and assumes independence between features.

4. k-Nearest Neighbors (k-NN):

k-NN is a non-parametric and instance-based classification algorithm. It classifies data points based on their proximity to other data points in the feature space.

Let's dive deeper into k-NN:

k-Nearest Neighbors (k-NN):

k-NN is a simple yet effective classification algorithm that works based on the idea of proximity or similarity between data points. It classifies a data point by considering the class labels of its k-nearest neighbors in the feature space.

The k-NN Algorithm:

1. Choose the value of k, the number of nearest neighbors to consider (an odd number is often chosen to avoid ties).
2. Compute the distance between the input data point and all other data points in the dataset (common distance metrics include Euclidean, Manhattan, and Minkowski).
3. Select the k-nearest neighbors based on the computed distances.
4. Count the class labels of the k-nearest neighbors and assign the majority class label to the input data point.

The intuition behind the k-nearest-neighbours model.

Lets think about the k-NN model intuitively. What we are basically saying is, if two objects share similar characteristics, they may be the same class of object. Now, of course that may also not be true for many reasons and we can be critical of this as a model later, but for now lets think about

Parameters of k-NN:

Two main parameters in k-NN are:

k: The number of nearest neighbors to consider (a hyperparameter that needs to be tuned). *Distance metric:* The measure used to calculate the distance between data points.

Distance Metrics

Strengths and Weaknesses of k-NN:

Strengths: Simplicity and ease of implementation. No assumptions about the underlying data distribution. Can handle both binary and multi-class classification. Weaknesses: Sensitive to the choice of k. Computationally expensive for large datasets. Affected by irrelevant or noisy features. Applications of k-NN:

k-NN is used in various applications, including: Image classification. Recommender systems. Anomaly detection. Medical diagnosis. Text categorization. And more. Conclusion:

Classification is a fundamental task in machine learning, and there are several basic models to choose from. k-NN is an intuitive and straightforward classification algorithm that relies on proximity in the feature space. Understanding the strengths, weaknesses, and parameters of k-NN is essential for effective model selection and application.