Group-26 AAD-Assignment-2

Omar, Eloise, Alina, Sue

2025-05-13

Contents

2.1 Nonlinear model vs non-parametric model

GAM:

Pros

- **Flexibility to capture non-linearity**. In a GAMM we replace each linear term $\beta_k x_k$ with a smooth function $f_k(x_k)$, meaning we can automatically model non-linear relationships between each predictor and the response without having to manually try out different transformation on each variable individually.

- **Non-linear fits may be more accurate**. The ability of creating this non-linear predictor and response relationship may make our model more accurate when predicting the medianHousingValue.

- **Interperability from additivity**. Since a GAMM is additive, we can examine the effect of each X_j on Y individually while holding all of the other variables fixed. Therefore, we can understand each variables individual effect on house value.

- **Control over smoothness**. Each f_j comes with an associated smoothing parameter (or degrees of freedom), making it straightforward to trade bias and variance (e.g. via cross-validation).

Cons

- **Additivity assumption restriction**. If there are strong synergistic effects like between location (longitude/latitude) and income –an additive model will not capture them unless explicit interaction terms $X_j \times X_k$ are included or low-dimensional interactions function of the form $f_{jk}(X_j, X_k)$ are manually introduced.
- **Computational cost**. Fitting this model to over ~$20,600$ observations and multiple smoothers can be very slow when compared to fitting a single parametric method.

KNN:

Pros

- **Completely non-parametric**. KNN makes no assumptions about the form of $f(X)$, allowing the model to potentially fit better than a parametric model. At a point x_0 KNN averages the responses of the K closest training blocks: $\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$, where N_0 is the set of the K nearest neighbors.
- **Control of bias-variance**. A small K yields a very flexible, low-bias but high-variance fit; large K yields a smoother, lower-variance fit.

Cons

- **Dimensoniality constraints**. As the number of predictors grows, the "nearest" neighbors tend to be far away in a high-dimensional space, so KNN's performance degrades rapidly as the number of predictors grow.
- **Distance metric sensitivity**. With a KNN you must scale the numeric features and encode the categorical features (e.g. oceanProximity) carefully. Otherwise poorly scaled or encoded features can dominate the distance of the nearest neighbors calculation.
- **Computationally intensive with predictions**. For each new group of predictions all ~$20,600$ must be computed to find the K nearest, which can be very intensive and slow.
- **Low interperability**. There is no simple way to explain a KNN prediction beyond pointing to the raw neighbors and their average.

In this housing-price context:

- **GAMM:** it is likely to give an interpretable model with, as we can analyse the partial-effects (e.g. how median income or ocean proximity individually affects price), and we can capture smooth non-linear trends.
- **KNN:** can capture complex interactions, between predictors, automatically, but with eight predictors (including a categorical one) it may run into high-dimensionality issues, making distance-based averaging unstable and slow.

2.1.1 GAM vs KNN approach

2.1.2 GAM vs KNN Regression Model

2.1.3 Which model performs better?

2.2 Classification models

2.2.1 Two classification methods

2.2.2 Suggested classifiers

2.2.3 Which classification method performs better?

2.3 A hybrid approach

2.3.1 Dicussing feasibility of the approach

2.3.2 test MSE of medianHousingValue using the approach

2.3.3 Comparison of the accuracy of this procedure to model in 2.1.3