Statistical Learning Methods
Questions Lecture 4

Tobias Famos

June 19, 2022

Questions

1. What are the advantages of K-NN Regression?
2. What are the disadvantages of K-NN Regression?
3. How to select distance measure for K-NN Regression?
4. How are similarity and distance connected?
5. Show a few distance measures for numerical values
6. How do we handle boolean values for distances?
7. Show a few distance measures for sets
8. Why do we normalize?
9. Show a few ways to normalize
10. What hyperparameters are there for KNN?
11. Explain how KNN regression works

Answers

1.
   - It is non-parametric, meaning there are no parameters to be determined, only the data it is performed upon.
   - No strong assumptions about $f(x)$ thus it can generate complex boundaries

2.
   - As non parametric model, it needs a lot of data
   - It is slow compared to other models
   - No simple statistical tests (e.g. p-values for coefficients)
   - No internal feature weighting. All features are taken into consideration, thus a feature selection must be performed.

3.
   - There is no clear way, no best distance measure (no free lunch)
   - Can only be determined by trial and error.

4.
   - similarity = 1 − distance

5.
   - Manhattan Distance = $\sum_{i=0}^{n} |a_i - b_i|$
   - Euclidean Distance = $\sqrt[2]{\sum_{i=0}^{n}(a_i - b_i)^2}$ Is most videly used.
   - Canbera = $\sum_{i=0}^{n} \frac{|A_i - B_i|}{|A_i| + |B_i|}$
   - Tantimoto = $\sum_{i=0}^{n} \frac{|A_i - B_i|}{max(A_i, B_i)}$

6.
   - Jaccard = $\frac{|A \cap B|}{|A \cup B|}$
   - Dice = $\frac{2 \cdot |A \cap B|}{|A| + |B|}$

7.
   - When working with distances (such as in K-NN) the normalization is important.
   - We normalize to make sure all the attributes have the same weight in the distance.

8.
   - MinMax Norm: $w'_k = \frac{w_k - min}{max - min}$
   - Z-Score: $w'_k = \frac{w_k - \mu}{\sigma}$ (μ: Mean, σ: Standard Deviation)

9.
   - Distance Metric
   - K (number of nearest neighbours to consider)

10.
    - To predict the value for x we take the average of the K nearest neighbours of x(N_i).

- $f(x) = \frac{1}{k} \cdot \sum_{x_i \in N_i} y_i$
- Intuitively: We take all the k neighbours, take their values and take average for x

11.

-