## **Attention based on scale-dot product**

In machine learning, “attention” refers to assigning relevance and importance to specific parts of the input data while ignoring the rest to mimic the cognitive attention process. Below is a small dataset that illustrates ow the attention mechanism works.

Assume we have the predictor to determine the value of the target variable . If you use a least squares linear regression model, you can predict the value of the target variable by using a fixed weight regardless of the input values . An example expression is given below, with as a constant.

While the attention mechanism takes into account the relevance of the input values to the predictors in the dataset. Instead of using a static value , the weights are generated according to how similar is to a data point in the training instances. The generalized attention model predicts the target as:

Where, the main ingredients of this equation are:

: These correspond to the values of the predictors in the training dataset.

: These correspond to the values of the target values y in the training dataset

: This is the value of the test data point

: This function determines the similarity of keys with the query.

The main idea here is to compute the similarity between the and and assign a high value if the query is close to the key, and a low value if they are very different. You can choose the function that fit your application.

: This function normalizes the weights.

This function is used to ensure that the attention weights lie between 0 and 1 and normalize to sum to 1.

Text

Description automatically generated