Exposé

Machine Learning Elective by Meghan Kane

Student: Matthias Dietrich (Interactive Media Design)

Project Idea

Before I started studying Interactive Media Design, I completed a apprenticeship as a geomatic, which is pretty much a more modern version of a cartograph. That's why I am interested in maps and geodata. So, I thought about a project idea, that uses geodata to create some kind of new map. Therefore, I searched for open source geodata. I found a dataset, which contains all municipalities of Germany with its name, population and superordinated federal state.

Visualised the dataset looks like this:

With this dataset in mind I thought about a project idea, which turned out to be more of a fun-project. I thought of a website, that takes a fictional location name as an input like „Bratheim"and through the machine learning algorithm you will see in which federal state this fictional place will most likely going to be in Germany geographically.

Learning Goals

Through this project I want to find out to what extent place names can be identified geographically. Besides getting to know the principles of machine learning, I would also like to know how machine learning can be used for geographic purposes.

Alternative Solutions

To identify the location of a place through its name, you could look at all the names manually and look for similar prefixes and suffixes like "-heim" or "Bad-". With these word-blocks, you could filter all locations and then analyse in what federal state the majority of these locations are . This approach would work, but it won't consider all types of similarity, that can be identified by a learning algorithm.

Approaching the 4 stages of Machine Learning

1. **Frame the problem**
   The Problem that needs to be solved is to assign a federal state based on a string as an input.

2. **Prepare Data**
   Since I already found the dataset I will be using, I just need to convert it into a format, which can be interpreted by the model. The found dataset contains about 10.000 municipals. I need to split this dataset into a training and a testing dataset. I also need to have the same amount of training data for every label (federal state).

3. **Train ML Model**
   First, I need to identify which type of model is suitable for my problem. I think a Recurrent Neural Network could be the right choice because I have a sequence of letters that need to be interpreted.

4. **Predict using ML Model**
   After the training I will start predicting the federal state for the test dataset to evaluate how well the model works. When I am pleased with the result, I will start predicting the coordinates for some fictional places.

Open Questions

One question I have is about how I am going to evaluate the precision of the trained model. For the test dataset I can use the distance between the predicted and the real location, but how am I going to check if the outcoming coordinates of a fictional place are reasonable?

Question answered due the 30 minute meeting:

I won't use coordinates as an output. Instead I will use the 16 federal states of germany to classify the input string.