Capstone Project

This is the proposal for the capstone project for the Machine Learning Nanodegree in Udacity.

Domain background

This project is from the field of Anthropology. Kung! is a tribe which lives in the southern part of Africa, on the western part of the Kalahari sand system. They are hunter-gatherers with a total population somewhere between 50,000 and 100,000. The !Kung language, commonly called Ju, is one of the larger click languages. Here is a wikipedia link for more information about Kung! tribe.

Project Aim

The project aims at understanding the relationship between the height and age for the Kung! tribe. Below is a snapshot of the data from a public website.

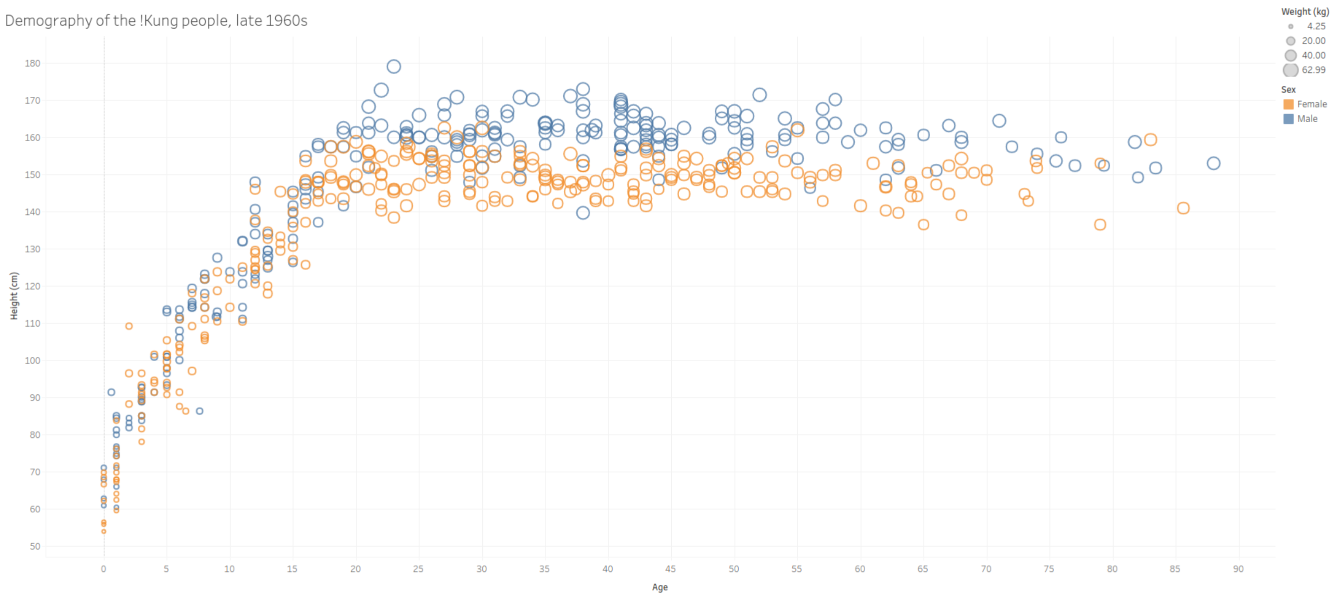

Fig 1: Plot of height vs age data for Kalahari Kung! San people collected by Nancy Howell.

Problem statement

We will built a probabilistic machine learning model for the height vs age relationship for the Kung! tribe.

This problem is picked from the book Statistical Rethinking by Richard McElreath.

Dataset

The dataset was collected by Nancy Howell. It can be found at this location.

It consists of following columns:

- height: Height in cm
- weight: Weight in kg
- age: Age in years
- male: Gender indicator

- age.at.death: If deceased, age at death
- alive: Indicator if still alive

For our purpose, we will only use the height and age columns.

Solution statement

Given the non-linear relationship between the height and age data in Fig 1, we will fit a set of polynomial models to the data.

Benchmark model

We will use a simple linear model as a benchmark for this.

Evaluation metrics

We will compare different polynomial models using Widely Applicable Information Criterion (WAIC) and test-sample deviance.

Project design

The project will follow the following workflow:

- Download the data.
- Select only the age and height columns
- Standardize the data
- Split the data into train and test data sets.
- Define polynomial models up to degree p and weakly regularizing priors for the parameters
- Fit the p models on the training data set
- Compare the models using WAIC
    - Choose the model with the best WAIC (M_{WAIC})
- Compare the models using test-sample deviance
    - Choose the model with the best test-sample deviance (M_{best})
- Does WAIC do a good job of estimating the test deviance?
    - M_{WAIC} = M_{best}
- Choose the best model (M_{best})

We will use the PyMC3 package for this work. PyMC3 is a Python package for Bayesian statistical modeling and Probabilistic Machine Learning.