4.1 Distance Metrics

May 9, 2019

1 Chapter 4. Relationships between Observations

The previous chapter discussed ways to measure relationships between variables, or the *columns* of a `DataFrame`. This chapter is about how to measure relationships between observations, or the *rows* of a `DataFrame`.

2 Chapter 4.1 Distance Metrics

How do we quantify how "similar" two observations are? We will use the Ames housing data set, but to keep things simple, we will work with just three quantitative variables from that data set: the number of bedrooms, the number of bathrooms, and the living area (in square feet).

```
In [1]: %matplotlib inline
        import numpy as np
        import pandas as pd
        pd.options.display.max_rows = 5

        housing_df = pd.read_csv("https://raw.githubusercontent.com/dlsun/data-science-book/mas
                                 sep="\t")

        # extract 3 quantitative variables
        housing_df_quant = housing_df[["Bedroom AbvGr", "Gr Liv Area"]].copy()
        housing_df_quant["Bathrooms"] = (
            housing_df["Full Bath"] +
            0.5 * housing_df["Half Bath"]
        )
        housing_df_quant
```

```
Out[1]:       Bedroom AbvGr  Gr Liv Area  Bathrooms
        0                 3         1656        1.0
        1                 2          896        1.0
        ...             ...          ...        ...
        2928              2         1389        1.0
        2929              3         2000        2.5

        [2930 rows x 3 columns]
```