Data Science

Deriving Knowledge from Data at Scale

Data Retrieving, Exploration, and Transformation

Hang Zhang, Ph.D.
Oct 15th, 2018

Agenda for Lesson 3

- Capstone Project kick-off
- How to retrieve data from different resources (in Python)
  - SQL Server
  - URL
  - Azure Blob Storage (Cloud)
- Data Exploration
- How to handle categorical variables
  - One-hot-encoding
  - Risk values
- Data Cleaning
- Data Normalization
- Feature engineering (Brief Overview))

Capstone Project

- PLAsTiCC Astronomical Classification
  - Can you help make sense of the Universe?
- Task:
  - The Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC) asks Kagglers to help prepare to classify the data from this new survey. Competitors will classify astronomical sources that vary with time into different classes, scaling from a small training set to a very large test set of the type the LSST will discover.
- Timeline:
  - December 10, 2018 - Entry deadline. You must accept the competition rules before this date in order to compete.
  - December 10, 2018 - Team Merger deadline. This is the last day participants may join or merge teams.
  - December 17, 2018 - Final submission deadline.
  - January 15, 2019 - LSST Workshop entry deadline.
  - February 15, 2019 - LSST Workshop announcement.

Timeline of Capstone Project

- Teaming up (<= 5 members per team)
  - Submit 1-page with team name, team member names with team lead in bold font
  - Team name convention: DS420_<your team name>

- Mid-term checkup report
  - Due midnight of the Sunday before the 7th class.
  - 3-page reports: what has been done, what is to be done, where are you on the leaderboard

- Final Report:
  - Due midnight of the Sunday before the last class (Dec 9th).
  - 6-page reports:
    - Show me how you executed this project by following data science process

Several Ways to Fail the Capstone Project

- I cannot find your team on the leaderboard
  - It means you do not ever have a valid submission
- Your leaderboard score does not improve from the baseline model
  - Baseline model is usually shared online
  - It means you do not do your own work
- Your team breaks the rules
- Lack of details in your report

In-class Lab

- Section 1: Retrieving Data from URL, SQL Server, and Cloud

Data Exploration and Visualization

- How do your datasets look like?
- What is the quality of data?
- What is the general statistics of the entire datasets?
- What is the distribution of each individual variable (univariate analysis)?
- What are the relationship between variables (multivariate analysis)?
- Any clustering pattern in your data?
- Is this a relevant dataset for your machine learning task? In another way, is the machine learning task easy or difficult based on the provided datasets?

Keep in Mind: Do Not Make Any Commitment Too Early about the Performance of the Machine Learning Model to Your Business Partner

- Since the data they provide to you might be irrelevant to the machine learning model they are expecting from you
  - Run an iteration of the data exploration and visualization
  - Create a report to present your data exploration and visualization results
  - Discuss with your business partner
  - Get confirmation and/or clarification from them about what you have found
  - Do multiple rounds till both parties agree that you have the right data for the right business problem

Data Schema

- How your data looks like?
  - How many columns each dataset has?
  - What are the types of columns?
  - Easiest way to get this is print out the top k rows of the data
  - For example:
    - In Python, assuming that you have read the data into pandas dataframe, use:

      df_name.head(k)

Data Quality

- What is the % of data that is missing?
  - Missing values might be represented as NA (default) in a data frame
  - Case by case, it might also be represented by some specific values (strings or numbers) in a dataset

- More importantly, we want to know:
  - Which variables have the highest rate of missing values
  - Where they are in the data frame

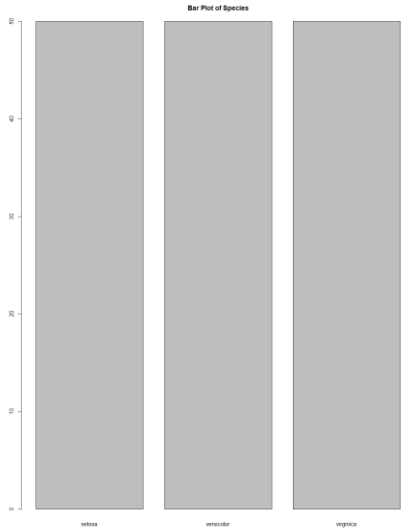
General Statistics of A Dataset

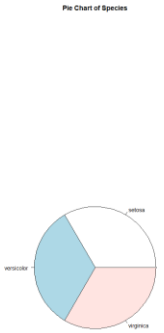
- Python:

```
import pandas_profiling
pandas_profiling.ProfileReport(df)
```

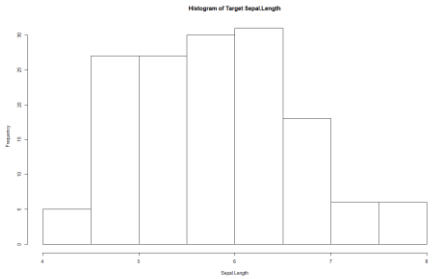
Investigation of Individual Variables

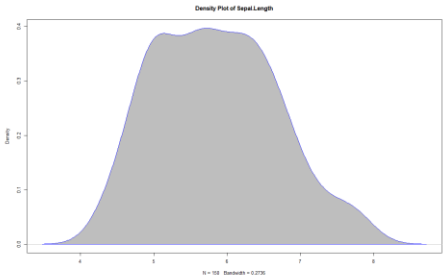
- For categorical variables:
  - Bar chart and pie chart

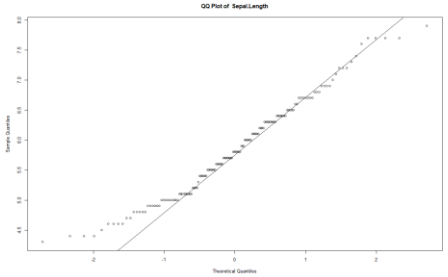


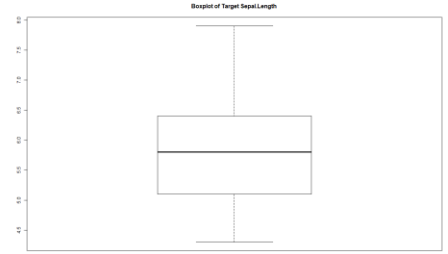


- For numerical variables:
  - Histogram
  - Probability density plot
  - QQ plot
  - Box plot



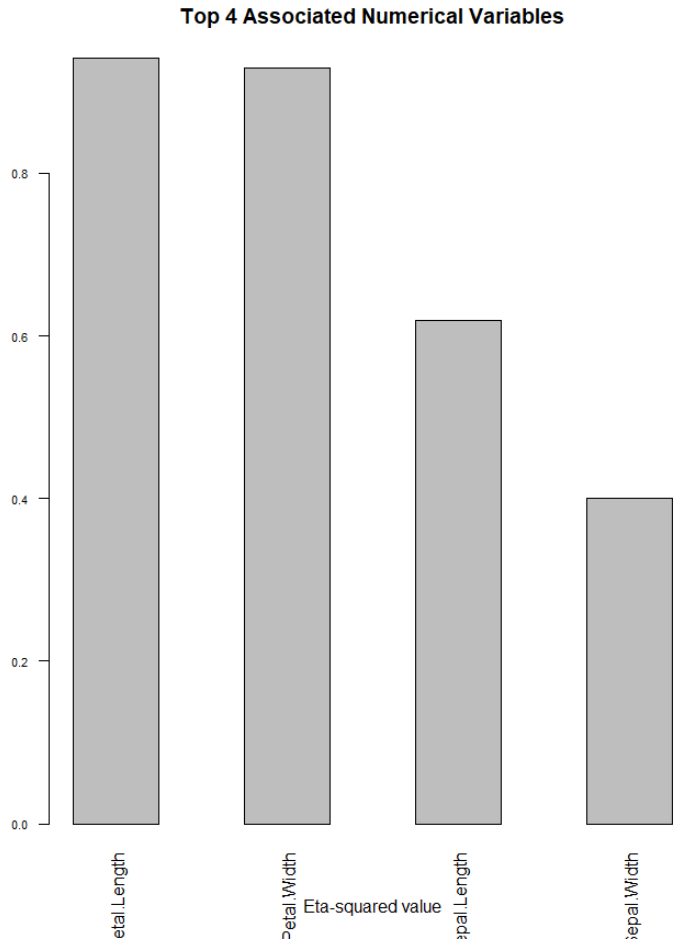






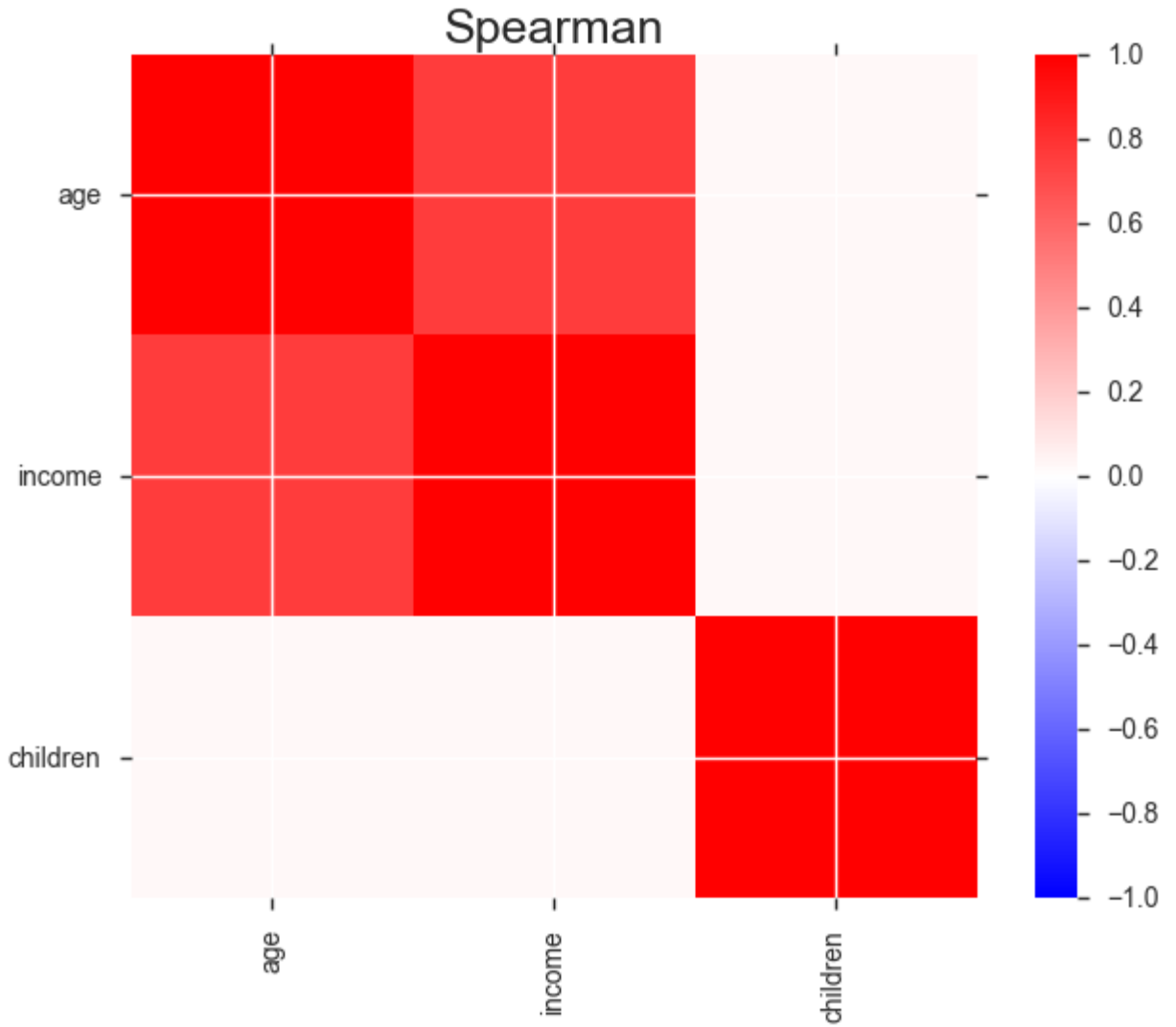
Investigation of Relationships between Variables

- Which variables are most important in my supervised learning?



Target Variable	Categorical X Variable	Numerical X Variable
Categorical (Classification)	Chi-square test	ANOVA
Numerical (Regression)	ANOVA	Correlation

What are the correlation structure of numerical variables?



In-class Lab

- Section 2: Data Exploration

Data Cleaning

Missing values are common– *UCI machine learning repository, 31 of 68 data sets reported to have missing values.*

"Missing" can mean many things… You need to have a discussion with the data provider or experts who understand the data collection/preparation process to understand why data are missing

It might be just a mistake when data is prepared

Dealing With Missing Data

- Throw away cases with missing values
  - in some data sets, most cases get thrown away
  - If missing data actually carries information, throwing away cases may lose information

- Impute (fill-in) missing values
  - Once filled in, data set is easy to use
  - However, if missing values poorly predicted, may hurt performance of subsequent uses of data set

- Treat “missing” as a new attribute value
  - Replace (fill-in) missing values with some value, and add an indicator variable to let the model know that this variable is missing at this observation
  - What value should we use to code for missing with continuous or ordinal attributes?

Missing Values: Imputing

Fill-in with mean, median, or most common value

Predict missing values using machine learning

Expectation Maximization (EM):

- Build model of data values (ignore missing values)
- Use model to estimate missing values
- Build new model of data values (including estimated values from previous step)
- Use new model to re-estimate missing values
- Re-estimate model
- Repeat until convergence

Data Cleaning

Outliers – *may indicate 'bad data' or it may represent something scientifically interesting in the data…*

Simple working definition: an outlier is an element of a data sequence S that is inconsistent with expectations, based on the majority of other elements of S.

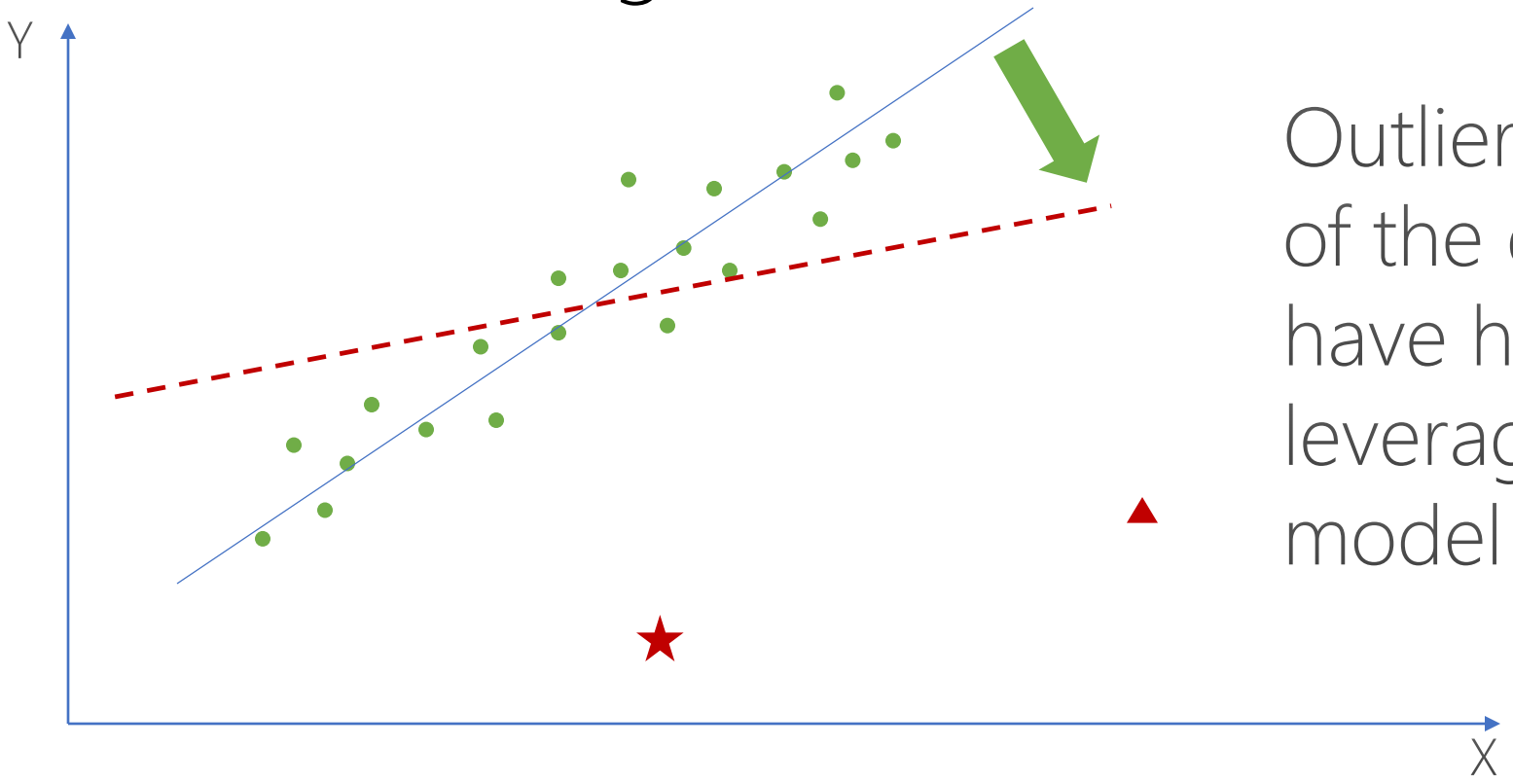
Sources of outliers

- Measurement error
- There does exist some extreme cases, for instance, some patients in healthcare insurance policies are 120 years old;

Data Cleaning

Outliers – *may indicate 'bad data' or it may represent something scientifically interesting in the data…*

Outliers can distort the regression results.



Outliers at the edge of the distribution have higher leverage on the model than others

Data Cleaning

Outliers – *may indicate 'bad data' or it may represent something scientifically interesting in the data…*

Identify outliers

- Question origin, domain knowledge invaluable
- Dispersion – *"spread" of a data set, departure from central tendency, use a box plot…*

Deal with outliers

- Winsorize – Set all outliers to a specified percentile of the data. Not equivalent to trimming, which simply excludes data. In a Winsorized estimator, extreme values are instead replaced by certain percentiles (the trimmed minimum and maximum). Same as clipping in signal processing.

  {92,19,101,58,1053,91,26,78,10,13,-40,101,86,85,15,89,89,28,-5,41}

Winsorize

-40
-5
10
13
15
19
26
28
41
58
78
85
86
89
89
91
92
101
101
1053

10th percentile

{92,19,101,58,1053,91,26,78,10,13,-40,101,86,85,15,89,89,28,-5,41}

Mean =101.5

{92,19,101,58,101,91,26,78,10,13,-5,101,86,85,15,89,89,28,-5,41}

Mean = 55.65

90th percentile

In-class Lab

Section 3. Handle Missing Values and Outliers

Scaling of Continuous Variables

- Many ML algorithms rely on measuring the distance between 2 samples
- There should be no difference if a length variable is measured in cm, inch, or km
- To remove the unit of measure (e.g. kg, mph, …) each variable dimension is normalized:
  - subtract mean
  - divide by standard deviation

Normalization

- **Min-max normalization:** linear transformation from v to v’
  - v’ = (v – min)/((max - min)*(newmax - newmin)) + newmin
  - Ex: transform $30000 between [10000..45000] into [0..1]

  ==> (30000 – 10000)/(35000(1)) + 0 = 0.5714

- **z-score normalization:** normalization of v into v’ based on attribute value mean and standard deviation
  - v’ = (v-Mean)/StandardDeviation

- **Normalization by decimal scaling**
  - moves the decimal point of v by j positions such that j is the minimum number of positions moved so that absolute maximum value falls in [0..1].
  - $v' = v / 10^{j}$
  - Ex: if v ranges between -56 and 9976, j=4 ==> v’ ranges between -0.0056 and 0.9976

Discretization/Binning

Less features, more discrimination ability

- Discretization is used to reduce the number of values for a given continuous attribute
  - usually done by dividing the range of the attribute into intervals
  - interval labels are then used to replace actual data values
- Discretization can have some physical, business, or biomedical meanings
  - For instance, in biomedical/healthcare area, age can be split into 0-5, 6-15, 16-21, etc.

Discretization Methods

- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well

- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples

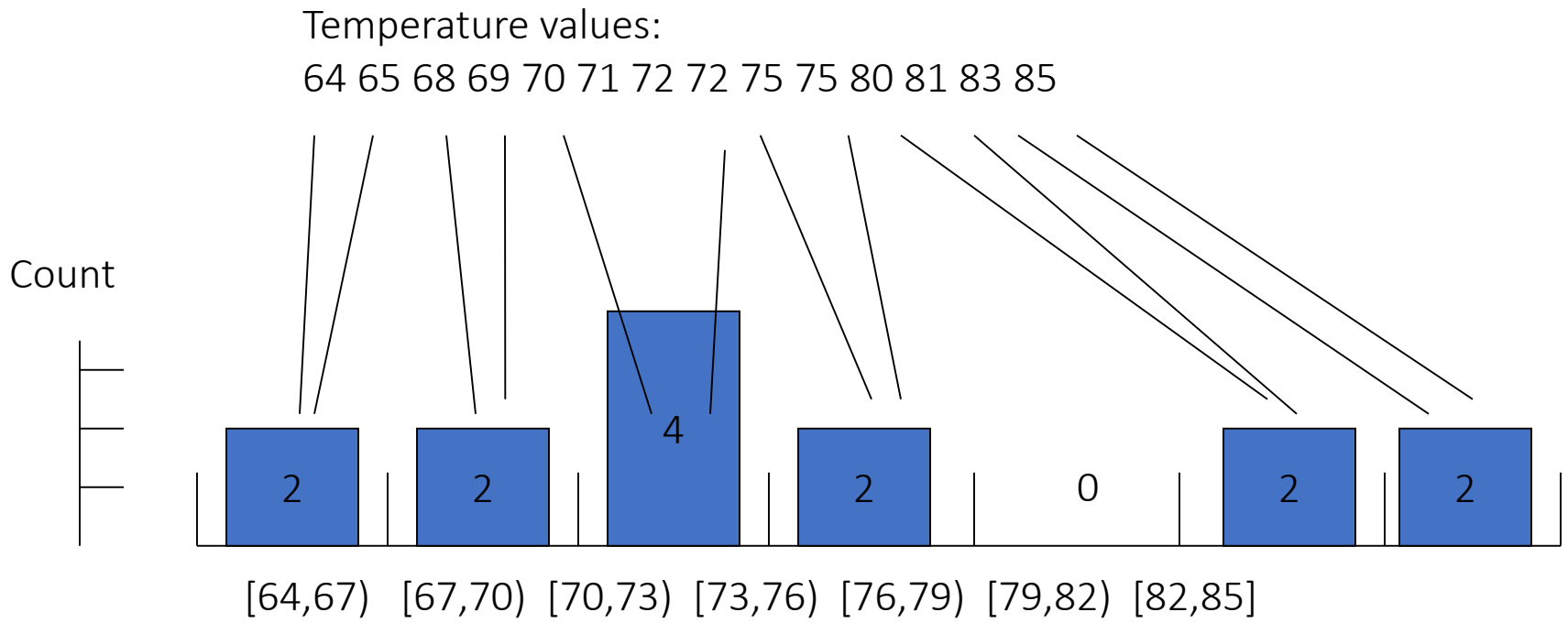
Equal width partitioning

1. Find the minimum and maximum values for the continuous feature/attribute F_i

2. Divide the range of the attribute F_i into the user-specified, n_{Fi} ,equal-width discrete intervals

[min, min+Δ), [min+Δ, min+2Δ), …, [max-Δ, max]

Equal-Width Partitioning



Equal Width, bins Low <= value < High

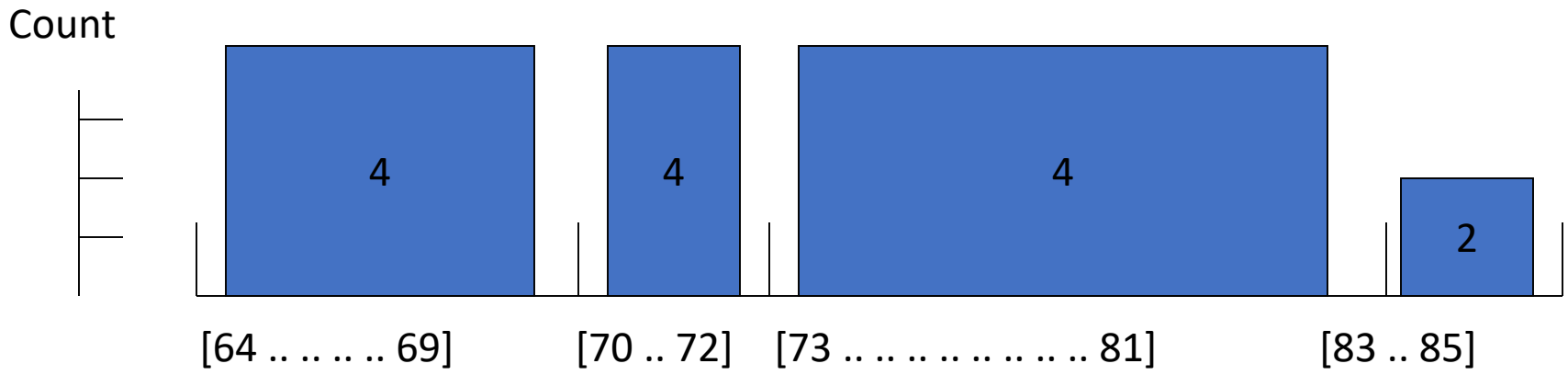
Equal-Width partitioning can produce clumping

Equal Height partitioning

1. Sort values of the discretized feature F_i in ascending order
2. Find the number of all possible values for feature F_i
3. Divide the values of feature F_i into the user-specified n_{Fi} number of intervals, where each interval contains the same number of sorted sequential values, and use the average between the two edging numbers of two consecutive bins as the edge dividing these two bins.
4. Assign the same bin labels to all observations falling in the same bin.
5. Apply the edges of the bins to allocate new observations into bins, and assign bin labels accordingly.

Equal-Height partitioning

Temperature values:
64 65 68 69 70 71 72 72 75 75 80 81 83 85



Equal Height = 4, except for the last bin

Equal-height partitioning: advantages

- Generally preferred because avoids clumping
- In practice, “almost-equal” height binning is used which avoids clumping and gives more intuitive breakpoints
- Additional considerations:
  - don’t split frequent (same) values across bins: e.g., 10, 11, 11, 11, 13, 14. If you want to partition into 3 bins with equal heights, bin 1: [10, 11], bin 2: [11, 11], bin 3: [13, 14]. Value 11 is put into 2 bins. Not good.
  - create separate bins for special values (e.g. 0)
  - readable breakpoints (e.g. round breakpoints)

Derived Variables

- Better to have a fair modeling method and good variables, than to have the best modeling method and poor variables
- Credit Risk Example: People are eligible for pension withdrawal at age 59 ½. Create it as a separate Boolean variable!
- Advanced methods exist for automatically examining variable combinations, but they can be computationally very expensive!

Special Transformations

Domain expertise, play a hunch in terms of feature discrimination

Example: *Date/Time* attribute

- Hour of a day
- Day of the week
- Day of the month
- Month of the year
- Day of the year
- Quarter of the year
- A holiday or not

Which ones to use depends on the prediction problem being solved

- Ex: For prediction of traffic on a freeway, Time of day, Day of the week, A holiday or not etc. will be useful

In-class Lab

- **Section 4. Scaling, Binning, and Data Transformation**