Data Exploration and Preparation

(Based on *Fundamentals of Machine Learning for Predictive Data Analytics* (Kelleher et al., 2015))

Overview of the lecture

- Getting to know your data
  - Why?
  - How?
- Data Quality Issues
- Handling Data Quality Issues
- Data Preparation

Getting to know your data – Why?

- Prior to running machine learning algorithms on your data, it is a good idea to analyze the data in order to know whether it presents any particularities worth considering.
- This is useful since it may allow you to pre-process the data in order not to obtain substandard results caused by these particularities.
- It may also help you understand why the results you obtain may be sub-optimal and guide you on how to design new learning methods able to get around the problems.

Getting to know your data – How?

- By using standard statistical measures of central tendency and variation.
- Central tendency can be calculated using measures such as the mean, mode and median.
- Variation includes standard deviation and percentiles.
- The data can also be visualized using bar plots, histograms and box plots.
- (See in class description)

Getting to know your data – How?

- For each feature, we should analyze the central tendency and variation to understand the type of values that each feature can take.
- For categorical features, we should examine the 1st and 2nd modes as well as the percent of the values they represent.
- For continuous features, we should look at the mean and standard variation, as well as minimum and maximum values.
- We can also visualize the categorical features using bar plots and the continuous features using histograms.

Bar plots of categorical and continuous features

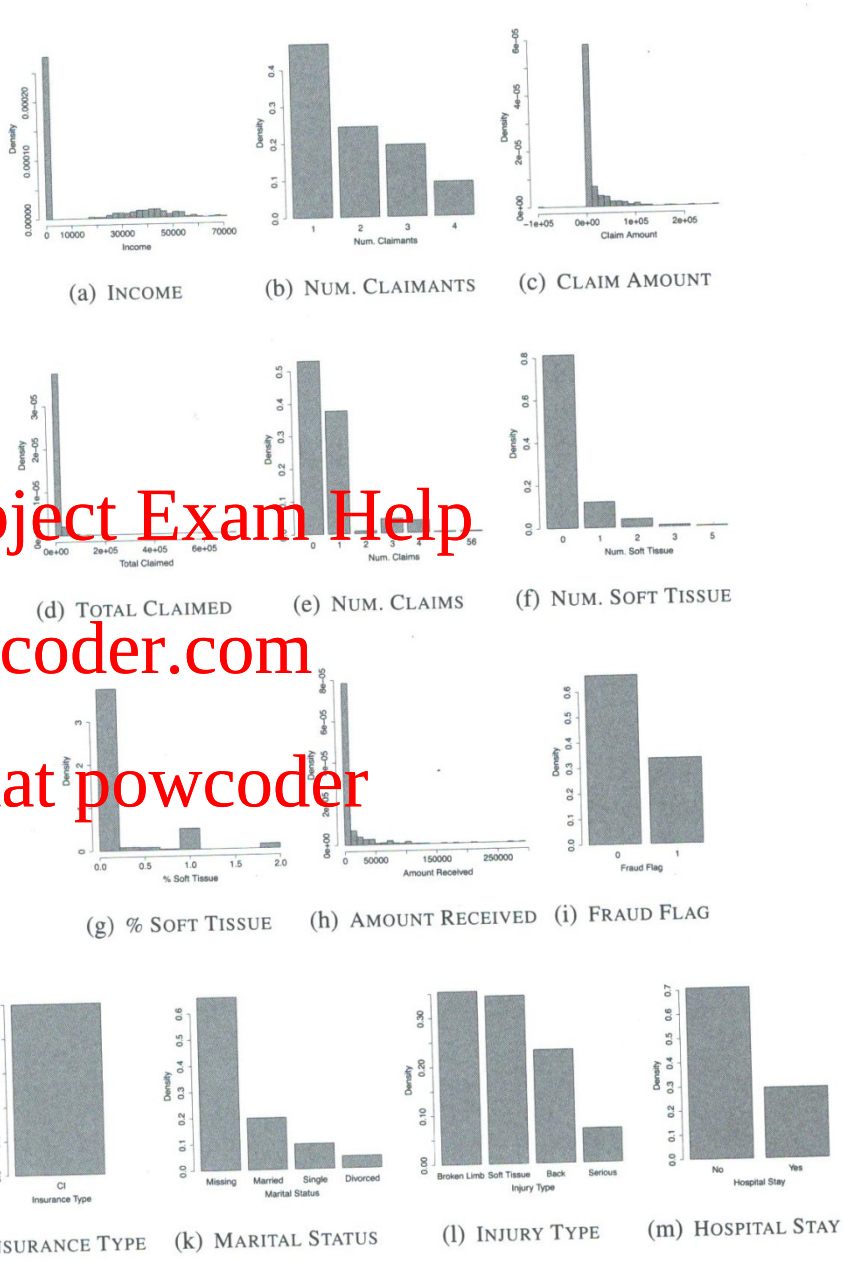


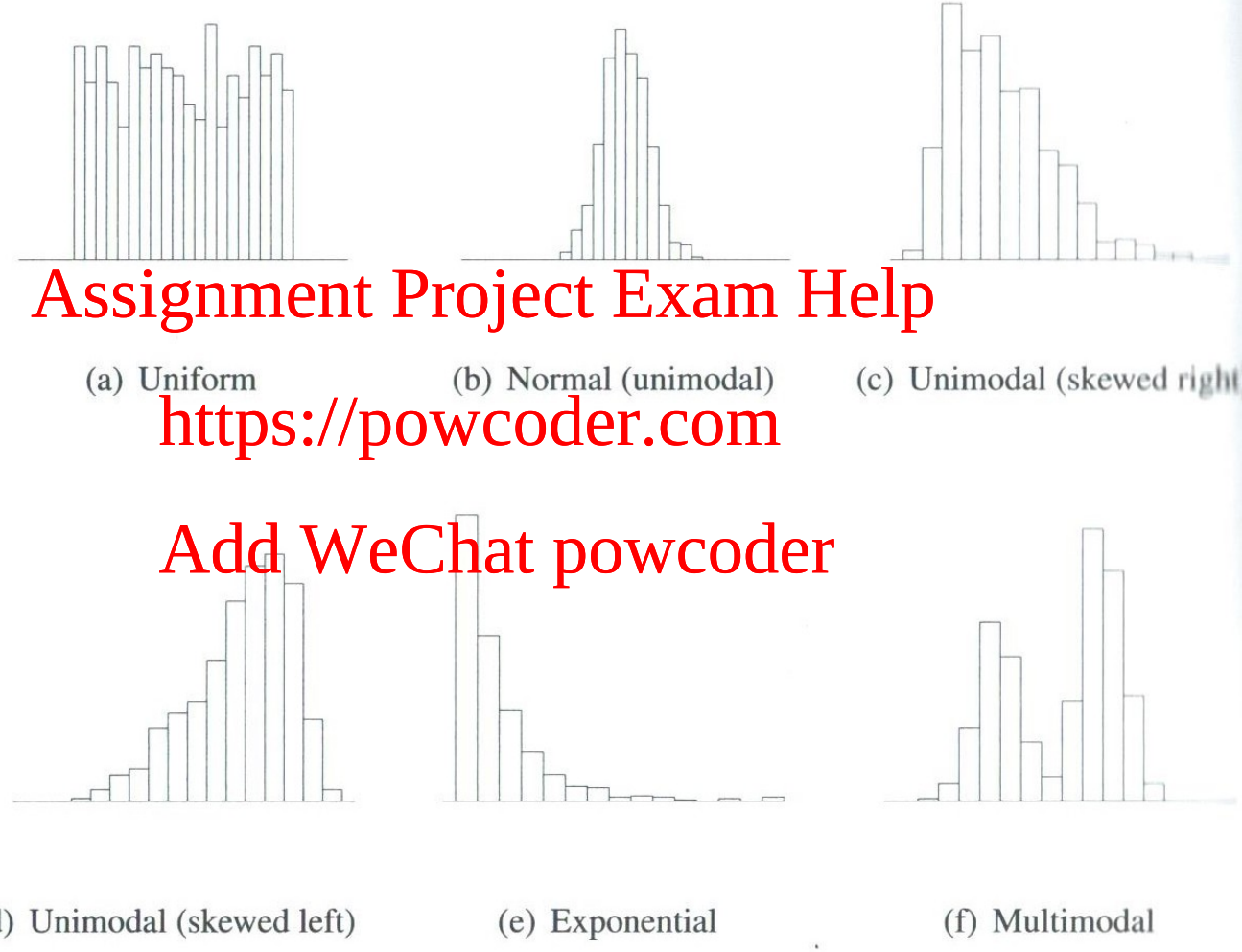
Figure 3.1

Visualizations of the continuous and categorical features in the motor insurance claims fraud detection ABT in Table 3.2[58].

What do the bar plots tell us?

- They show us how many values each feature can take.
- They also show us whether there are any dominant values for each feature and the extent of this dominance.

Histograms for 6 different data sets representing 6 well-known distributions



(a) Uniform (b) Normal (unimodal) (c) Unimodal (skewed right)

(d) Unimodal (skewed left) (e) Exponential (f) Multimodal

Figure 3.2

Histograms for six different sets of data, each of which exhibit well-known, common characteristics.

What do histograms tell us? I

- Features listing an ID number often have a **uniform** distribution.
  - Finding features with a uniform distribution is not that useful for learning.
- Naturally occurring phenomena often exhibit a **normal** distribution.
  - Finding features with a normal distribution is a good thing as many learning methods work well with such distributions
- Features representing salaries can often exhibit a **unimodal distribution with a right skew**: most people have salaries falling around a central tendency, but some individuals have very high salaries. These distributions are also said to have a **long tail**.
  - Such distributions are more difficult to learn than distributions without a skew

What do histograms tell us? II

- Features such as the number of times a person has made an insurance claim tend to follow an exponential distribution: the likelihood of low values occurring is high, but it diminishes rapidly for high values. Outliers are likely in exponential distributions, and these are problematic for learning systems.
- Features following a multimodal distribution have two or more very commonly occurring ranges of values that are clearly separated. The heights of a randomly selected sample of men and women, for example, is likely to follow a bimodal distribution. Multimodal distribution cause problems because the measures of central tendency and variation break down. But if the data can be separated into its different mode, then many problems will be solved.

Identifying Data Quality Issues

- Data quality issues refer to the presence of anything unusual about the data. The most common data quality issues are:
  - Missing values
  - Irregular cardinality
  - Outliers

- Data quality issues may occur in
  - Invalid data
    - These issues should be corrected before further processing
  - Valid data
    - Corrective steps are usually not taken, except if necessary for further processing.

Missing values

- When missing values are observed, the first step is to determine why this is happening.
  - Errors in data integration or the generation of values for derived fields ➔ Corrections must be made
  - Missing values in valid data: sensitive data omitted in certain cases, recording instruments didn't work for a certain period of time, etc. ➔ No correction can be made
- If the proportion of missing values is very high (e.g., greater than 60%), it is, sometimes, a good idea to remove the feature.

Handling Missing Values

- There are three general ways of handling missing values:
  - Dropping the feature. However, this shouldn't be done if only a few values are missing, as useful information could be lost.
  - Using a missing indicator value. This is sometimes informative. E.g., if a feature is missing due to the reluctance of a person to provide sensitive information represents useful knowledge about the person.
  - Imputation refers to replacing the missing value by a plausible estimated value. That shouldn't be done if too many values are missing (e.g., greater than 30% of missing values). One approach is to replace the missing values with a measure of central tendency for that feature (e.g., mean or median), but there are more advanced methods as well.

Irregular Cardinality

- Irregular cardinality occurs when the cardinality for a feature doesn't match what we expect. These should be inspected carefully as they may indicate an error that needs to be corrected. In particular, there are several issues to worry about:
  - Cardinality of 1.
  - Categorical features labeled as continuous.
  - Categorical features with higher cardinality than expected.
  - Categorical features with a very high number of values

Handling Irregular Cardinality

- Cardinality of 1. If this is due to an error, the error should be corrected. Otherwise, the feature should be removed since it will not offer any useful information
- Categorical features mistakenly labeled as continuous. Continuous features with low cardinality are often in this category, though not always (e.g., a feature indicating the number of children has low cardinality but is genuinely continuous). On the other hand, a gender feature that uses values 0 and 1 to indicate male or female is not naturally continuous. Such features should be recognized and treated as categorical.
- Categorical features with high cardinality. Such features are sometimes indicative of invalid data. E,g, a gender feature with 6 values could arise from the use of values: male, female, m, f, M, F. Such a feature should be cleaned up.
- Categorical features with a very large number of values. (E.g., larger than 50). These will cause learning systems difficulties. So they should be investigated and perhaps simplified or removed.

Outliers I

- Outliers are values that lie far away from the central tendency of a feature. There are two kinds of outliers: invalid and valid ones.
- Invalid outliers are caused by one of several issues: a data entry error (e.g., an entry of 100,000 instead of 1,000) or a defective measurement instrument.
- Valid outliers are correct values that are very different from the rest (e.g., a billionaire's salary versus other 'regular' salaries).

Outliers II

- There are two main approaches to detecting outliers:
  - Examine the minimum and maximum values for a feature. This could identify implausible values and, thus, invalid outliers (e.g., a value of -12 as minimum age value)
  - Compare the gaps between the median, minimum, maximum, 1rst quartile and 3rd quartile values. For example, if the gap between the 3rd quartile and maximum value is larger than the gap between the median and 3rd quartile, this indicates an unusual maximum value. However, it is likely that this represents a valid outlier. Since many learning systems do not handle outliers well, this is information worth noting. BTW, these kinds of outliers are easy to spot using box plots or distribution plots.

Handling Outliers

- Using a Clamping transformation. The idea here is to set the outlier values to lower or upper thresholds established manually using domain knowledge or calculated automatically from the data (e.g. 1st quartile value minus 1.5 * inter-quartile range for the lower value and 3rd quartile + 1.5 * inter-quartile range for the upper value. 1.5 is arbitrary and can be changed).
- Mean-based thresholds. These can be calculated as mean +/- 2 * standard-deviation for the upper and lower thresholds, respectively. Again, 2 is arbitrary).
- Modifying outlier values is controversial: some people argue that the transformation may remove the most interesting aspect of the data. On the other hand, since outliers hamper the performance of many learning system it is sometimes necessary to apply such transformations.

Advanced Data Exploration

- These are methods that examine the relationship between pairs of features:
  - If two features are highly correlated, we can reduce the dimensionality of the data set by removing one of them.
- For two continuous features, we can use scatter plots and matrices of scatter plots.
- For two categorical features, we can use the small multiples visualization.
- For a categorical and a continuous feature, we can use, once again, the small multiples visualization.
- In addition, to visualizing the relationships, we can also calculate formal measures such as the covariance and correlation of the two features.
- (See the in-class discussion of all these methods)
- We will discuss more complex feature relations when we explore the topic of feature selection.

Data Preparation

- In order to make the data more compatible with certain types of learning systems, data representations can be modified. Three such techniques are:
  - Normalization
  - Binning
  - Sampling

Normalization

- Having continuous features that cover different ranges can cause difficulty for some machine learning algorithms.
- There are two simple methods to normalization:
  - Range normalization, which performs a linear scaling of the original values of the continuous feature into a given range.
  - Transform the values into Standard scores. A standard score measures how many standard deviations a feature value is from the mean for that feature.
  - (See the description of these methods in class)
- Range normalization has the disadvantage of being sensitive to outliers while standard scores assume that the data is normally distributed.

Binning

- Binning involves converting a continuous feature into a categorical one. To perform binning, we define a series of ranges called bins that correspond to the new categorical features. Two popular approaches to binning are:
  - Equal Width binning
  - Equal Frequency binning
  - In both cases we need to specify how many bins we intend to use. This number is difficult to set: too few bins doesn't differentiate enough between the features. Too many means that there will be too few instances in each bin.

Sampling

- Some data sets are so large that we don't use the entire data set, but instead, focus on a sample of the data. Here are some sampling techniques:
  - Top sampling (only take the top x% instances) [not recommended since the original set may have been ordered in some way].
  - Random sampling
  - Stratified sampling [ensures that the relative frequencies of the values of some features are maintained in the sample]
  - Under sampling/Over sampling
- We will discuss sampling issues further in the context of the class imbalance problem.