Feature Spaces & Distance Measures

The KDD Process[1]

Knowledge Discovery in Databases

Selection

Selection is the way of selecting the attributes that are necessary to solve the current problem. For an example; the addresses of customers may not be of interest when discovering patterns in the selection of food items at a grocery store.

Preprocessing

Databases are notoriously "noisy" or may contain inaccurate or missing data. During the preprocessing stage the data is cleaned and normalized.

Transformation

During the transformation phase, attributes are converted to useful types. For an example converting nominal values to integers. New or "derived" values are defined.

Data Mining

At this point the data is subjected to one or several data mining methods such as classification, regression, or clustering.

Interpretation & Evaluation

The final step is the interpretation and documentation of the results from the previous steps. The results should be translated into a form understandable to the user. A commonly used way to to that is visualization. When the results are understood, it should be critically reviewed and and conflicts with previously believed or extracted knowledge resolved.

Feature Spaces

Feature spaces refer to the n-dimensions where the variable reside without including the target variable. An example could be:

Target

- Y = Thickness of car tires after some testing period.

Variables

- X_1 = distance travelled in the test.
- X_2 = time duration of test.
- X_3 = amount of chemical C in tires.

The feature space is therefore R^3.

[1] https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/KDD3.htm

Nominal (categories)

It is possible to tell whether to two values are equal - but no direction (better, worse,...) or meaningful distance.

Example: sex, eye, color, healthy/sick, amino acids etc.

Ordinal

There is an relation (better, worse,...), but no uniform distance.

Examples: grade, quality label, age class (20-29, 30-39,...), color (?).

Metric

Differences and relations between are meaningful. Values can be discrete or continuous.

Examples: weight, selling counts, age

Frequency

For a feature X, we have a sample x_1, x_2, ... , x_n

For each value a, we have f(a) as the number of occurrences in the sample.

Aggregations of Feature Vectors

Centroid

A centroid is the mean position (average) of all the points in a geometric shape. In a general metric space (that is, not a vector space), where we only have pairwise distances, it might not be possible to compute a centroid.

Medoid

The medoid is similar to the centroid. The most significant difference is that the medoid is part of the data set. This means that it has to be a data point already existing which are as close to the centroid as possible.

In a general metric space, the medoid is the object with the smallest average distance to all other objects.

Features for Images

When looking for features in images, there is generally three things to look after:

- Distribution of colors
- Textures
- Shapes (contours)

Color Histogram

A color histogram represents the distribution of colors over the pixels in an image. It is important that the same color space are used when comparing colors in images (RGB, HSV, HSL,...).

Also maybe do some normalization, so that the images have the same size.

The number of representants in a color histogram is the number of different colors. By lowering the amount of colors in the image, the number of representants are lowered. This may be a good idea, as it makes two images with different shades of the same color more likely to be classified as the same (of course, only do this if this is needed).

Features for Texts

Bag of Words

In the BOW algorithm, texts can be seen as sets of words or vectors of terms. A term can be a

- single word.
- phrase or part of a sentence.

Typically texts are transformed into a vector of term frequencies. However, term frequencies have some problems. For an example, many words are totally pointless for distinguishing texts (the, a, it, this, ...). Different versions of the same word can appear in the text (learn, learning, ...). The feature space often also becomes very high dimensional due to the number of words. Most words from a dictionary also doesn't appear at all in any of the texts compared.

Data Mining Models

There are several models used in data mining. Each are used in different cases and such also give different kinds of results.

Classification

Classification is used to classify new data based on previously known classifications of similar data.

Classification algorithms (not limited to):

- K-Nearest Neighbor
- Decision Trees
- Support Vector Machines

Clustering

Clustering is used when there is no substantial knowledge about some data. On example could be a lack of classifiers which would allow classification to be used instead. However, clustering is still descriptive, because it tells something about the data in its current state.

Clustering is therefore used to find data that shares similarities with other data and then cluster the data into groups.

Clustering algorithms (not limited to):

- K-Means
- DBSCAN
- OPTICS

Regression

Whereas classification and clustering is regarded as descriptive algorithms, regression is a predictive algorithm.

Regression is done by fitting a function on some data. The function is continuous and can therefore be used to predict things in the future.

Regression Algorithms (not limited to):

- Linear regression
- Polynomial regression

Distance Measures

Euclidean Norm

In euclidean mathematics, the euclidean norm is the calculated shortest path between two points in any n-dimensional space.

$$\|\boldsymbol{x}\| := \sqrt{x_1^2 + \cdots + x_n^2}.$$

Manhattan Norm

The Manhattan Norm is relating to the distance between two points on a rectangular grid.

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i| \,.$$

This norm is simply the sum of absolute values of the columns.

Chebyshev Norm

The Chebyshev Norm also known as the Chessbord Distance, is the shortest distance between two points on a grid where movement in vertical, horizontal and diagonal direction is allowed. The movement behavior is exactly the same as the king in a game of chess.

The Chebyshev distance between two vectors or points *p* and *q*, with standard coordinates p_i and q_i , respectively, is

$$D_{\text{Chebyshev}}(p, q) := \max_i(|p_i - q_i|).$$