**1. Overview of data preparation**

**2. Understanding data**

**Part I: Data Integration**

**3. Feature derivation**

(1) One-degree feature derivation: quantity, category, time, space, text, etc.

(2) Two-degree feature derivation: multiplication, Cartesian product

(3) Multiple-degree feature derivation: generated by polynomial kernel functions

**4. Dealing with the problem of imbalanced categories**

(1) Under-sampling: EasyEnsemble

(2) Over-sampling: SMOTE

(3) Hybrid sampling

**Part II: Data Transformation**

**5. Normalization**

(1) Normalization of quantitative variables: standardization, min-max normalization, power transformation, whitening transformation, samples normalization

(2) Normalization of qualitative variables: one-hot encoding, ordinal encoding

**6. Discretization**

(1) Unsupervised discretization: width discretization, frequency discretization, mean-SD discretization, $K$-means-based discretization

(3) Supervised discretization: Chimerge, CAIM, MDLP-based discretization

**Part III: Data Cleansing**

**7. Cleaning dirty data:** duplicates, invalid values, errors

**8. Dealing with outliers**

(1) Distribution-based methods: 3-sigma principle, box plot, elliptic envelope

(2) Distance/density-based method: LOF

(3) Model-based methods: isolation forest, one-class SVM

**9. Dealing with missing data**

(1) Cold-deck imputation

(2) Imputation with estimates: KNN imputation, regression imputation, miss forest

(3) Imputation from sampling: random regression imputation, MICE, PMM

**Part IV: Data Reduction**

**10. Feature selection**

(1) Filter: unsupervised filter, supervised filter (based on multiple testing, mutual information, MIC)

(2) Wrapper: feature selection from single model, SFS, RFE

(3) Embed

**11. Feature extraction**

(1) Projection methods: truncated SVD, PCA, ICA, kernel PCA

(2) Manifold learning: MDS, Isomap, LLE, t-SNE