Pattern Project

Saturday, December 14, 2024 2:45 PM

Step 0 → Finding dataset

- at least 3 classes or clusters

UCI, Kaggle, ...

Step 1 → Load data

- Missing values → Delete row or column
  - → replace with mean, median, mode...
- Dataset info → number of samples, Number of features
  - → feature type (Binary, continuous, ...)
  - → Duplicate samples → remove
  - → Outlier detection → remove or replace
  - → Standardization, Normalization.

Step 2 → Data preprocess

- Feature correlation → corr with target
  - → corr with other features
- Feature Selection / Feature Ranking
  - → Filter methods → Correlation
    - → Univariate Feature selection
      - Man whitney U test
      - chi square
      - Shapiro-wilk test
  - → wrappe methods → Forward elimination
    - → Backward elimination
    - → Recursive feature elimination
  - → Embedded methods → L1-regularization
    - → Feature importance
- Dimention reduction → can be used for reducing feature numbers.
  - PCA
  - → reduce features to 2 or 3 for better

- Dimention reduction → Can be used for reducing feature numbers.
  - PCA → reduce features to 2 or 3 for better visualization.
  - ICA
  - t-SNE
  - LDA
  - Autoencoders (NN based)

step3 → classifiers

- tree-based methods → Random forest
  → Decision tree
- Support vector machines → linear
  → Non-linear
- instance-based learning models → k-NN, k-means
- probabilistic classifier → Logistic regression
  → Naive bayes

Step4 → Performance evaluation

- model performane evaluation
  - dataset spits → train, test, validation
  - evaluation method → cross validation
    → k-fold
    → leave one out ...
  - Confusion matrix
  - Evaluation metrics → Accuracy
    → recall
    → precision
    → f1-score
    → ...
  - ROC Curve, AUC
  - overfit, underfit

- ROC curve, AUC
- overfit, underfit
- hyperparameter tuning

- Model comparison → compare performance between different classifiers.

Step5 → Conclusion

- what was the best performance.
- the most important features.

Situational step → feature extraction

- EEG, fMRI
- graph
- Connectivity matrix.