R 실습

2015.09.23

심재웅

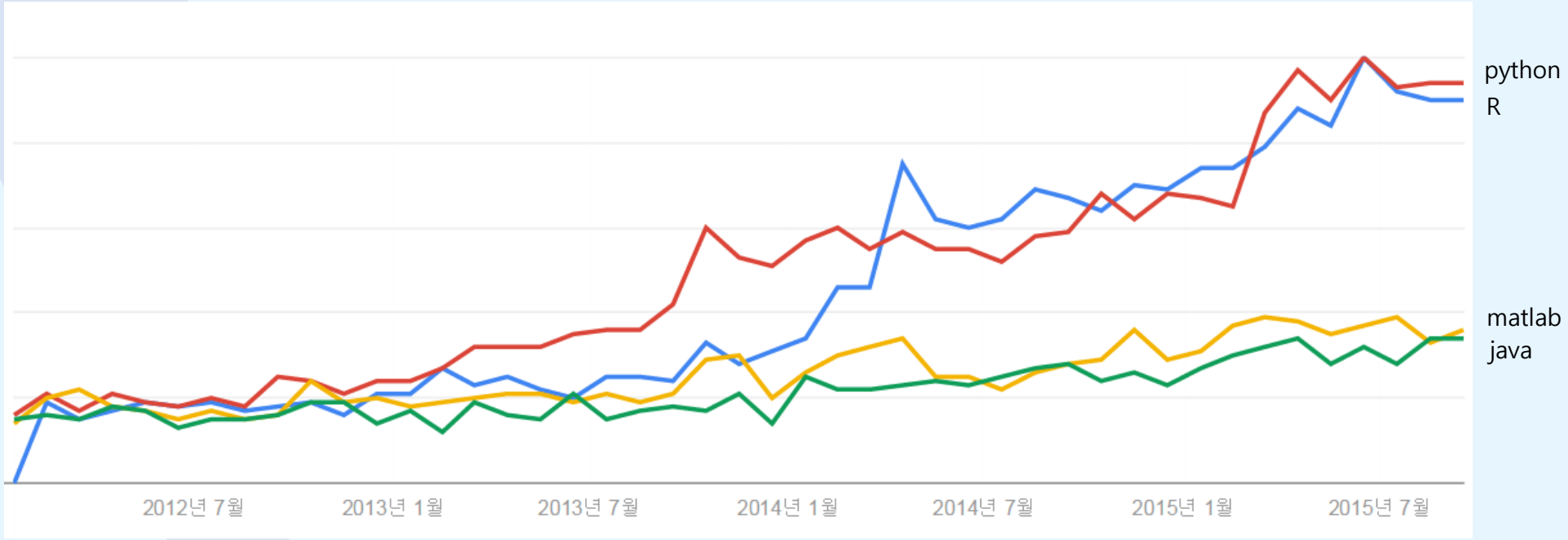
R ?

- "R is an integrated suite of software facilities for data manipulation, calculation and graphical display."

- an effective data handling and storage facility
- a suite of operators for calculations on arrays, in particular matrices
- a large, coherent, integrated collection of intermediate tools for data analysis
- graphical facilities for data analysis and display either directly at the computer or on hardcopy
- a well developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities

R for machine learning

- R은 비교적 배우기 쉬운 편이고, Statistics, Machine learning 관련 패키지의 공유가 활발한 편이라 많이 사용됨
- '분석' 레벨에서는R, Python, Matlab등이 많이 사용됨
- cf) Julia, Go, ...
- R vs Python : http://blog.datacamp.com/r-or-python-for-data-analysis/



구글 트렌드 검색 ' ** + machine learning'

R??

- Python, Matlab과 같이 스크립트 언어이기 때문에 비교적 문법이 쉬움
- 스크립트언어
  - 코드를 한 줄 씩 실행
  - 전체프로그램이 완성되지 않아도 코드의 일부분만을 실행시킬 수 있음

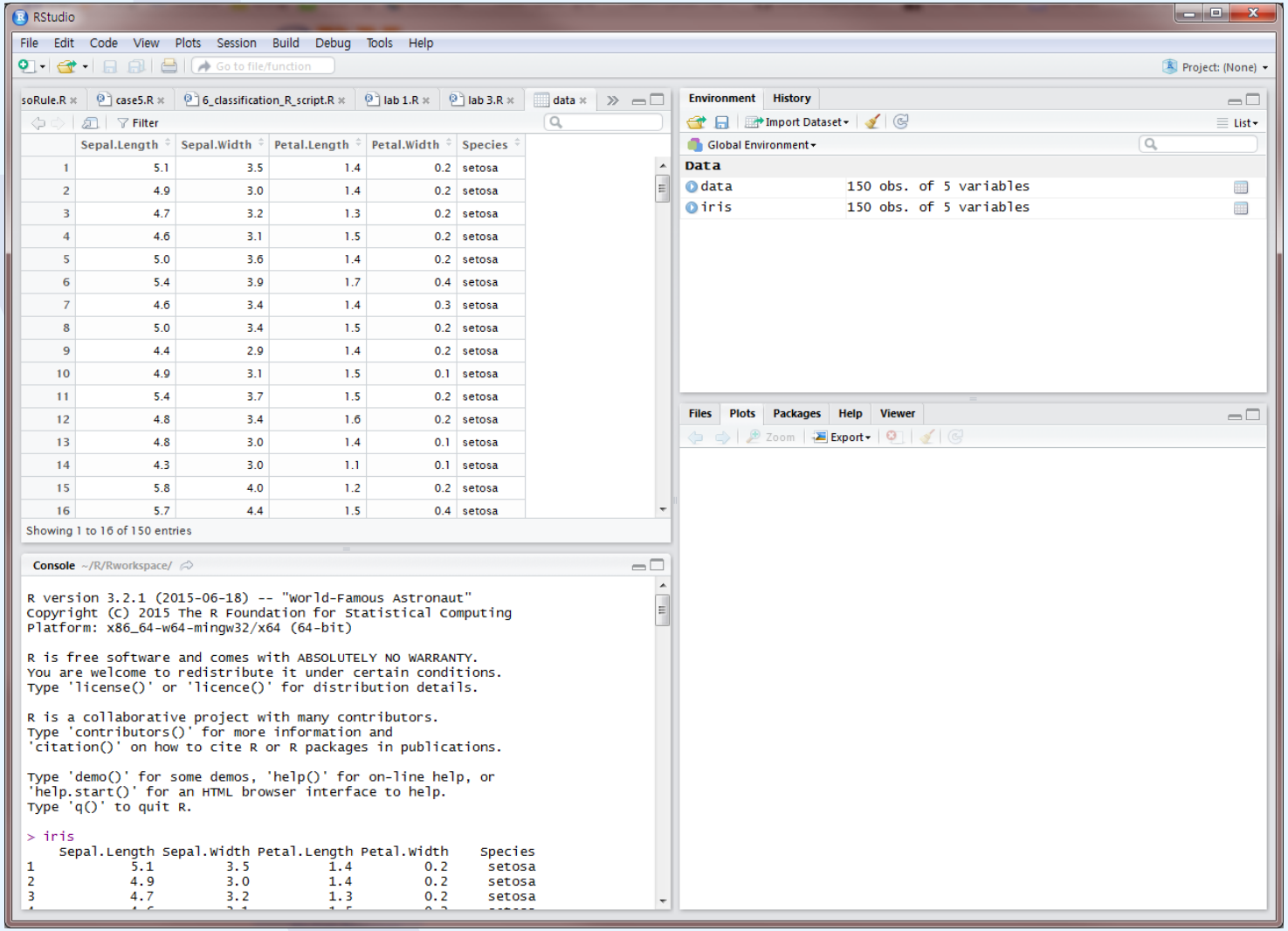
- Java에서처럼

```
public static void main(String args[]){
        System.out.println("Hello, world!");
}
```

- 필요 없음

Install : R + RStudio

- R을 쉽게 이용할 수 있도록 도와주는 IDE (RStudio: R = Eclipse : Java)



실습

- Data description
  - Bank marketing dataset (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing)
  - 은행에서 고객의 정기 예금 가입 여부를 예측하기 위한 목적으로 만든 데이터 셋
  - Input variables: 고객의 나이, 직업, 교육 수준, 잔고 등 16개
  - Output variable: 정기 예금 가입 여부

age	job	marital	education	default	balance	housing	loan
58	management	married	tertiary	no	2143	yes	no
44	Technician	single	secondary	no	29	yes	no
…	…	…	…	…	…	…	…

…

…

contact	day	month	duration	campaign	pdays	previous	poutcome	y
unknown	5	may	261	1	-1	0	unknown	no
unknown	5	may	151	1	-1	0	unknown	no
…	…	…	…	…	…	…	…	…

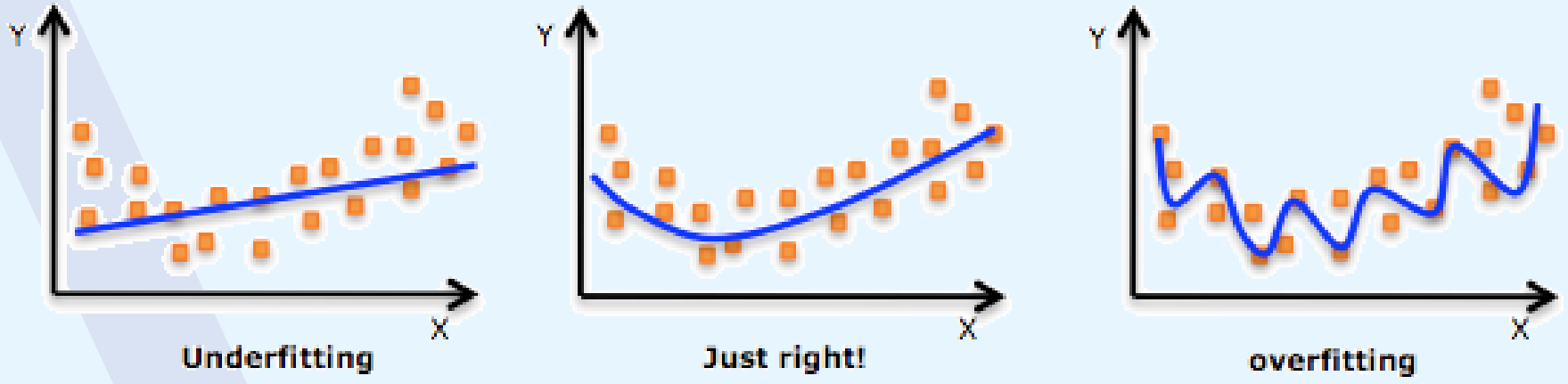
- Classification: 정기 예금 가입 고객 예측 문제

실습

- Data import
- Simple manipulation
- Exploration
- Classification modeling

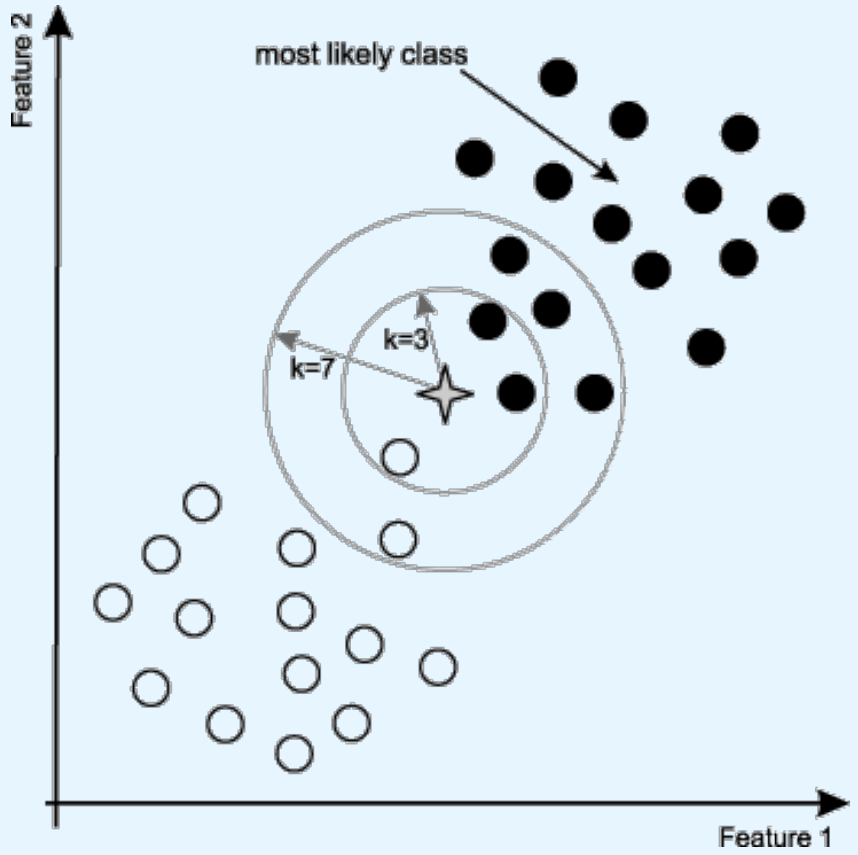
실습

- Training set & validation set
  - Classification, Regression은 무언가를 예측하기 위해 사용되는 기법
  - 모델을 통해 새로운 데이터를 예측할 수 없으면 의미가 없음
  - 새로운 데이터에 대해서도 적용될 수 있도록 일반화된 결과를 얻어야 함
- Validation set을 고려하지 않으면?
  - 데이터 셋에 대하여 일대일 함수처럼 매핑하면 정확도가 거의 100%에 근접
  - 그러나 Robustness가 떨어짐
- 따라서 7:3 (혹은 6:2:2) 정도로 데이터셋을 분리하고 학습 & 검증을 거쳐야 함



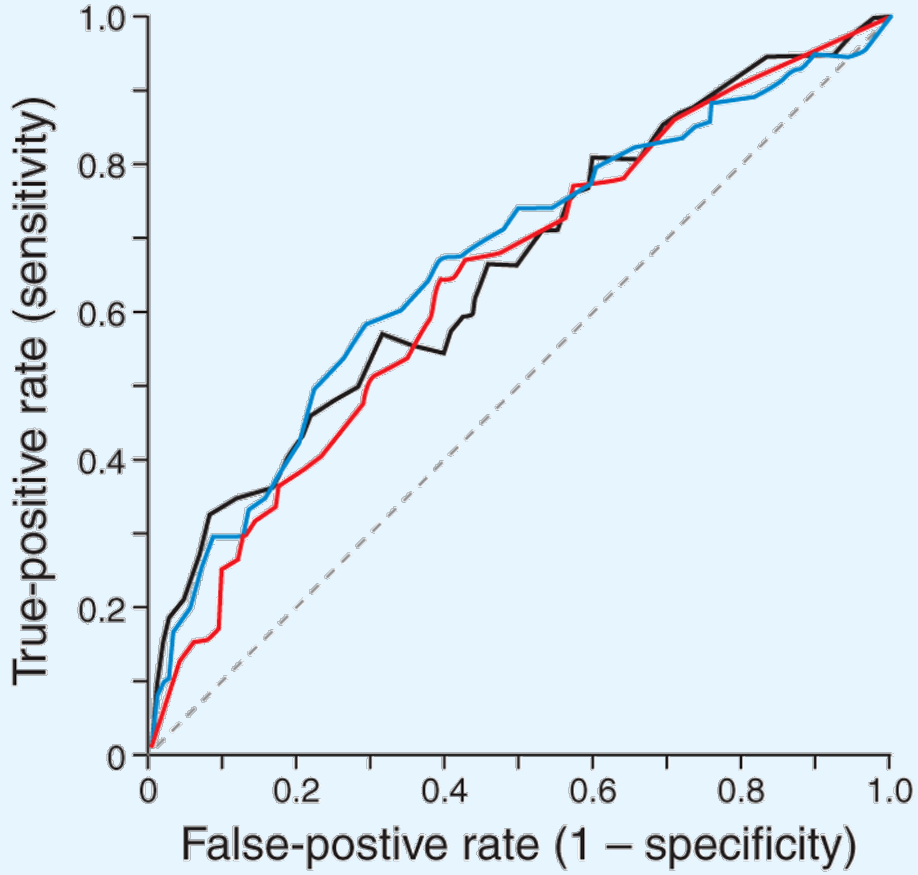
실습

- k-nearest neighbors classification
    - 어떤 object의 class는 주변의 이웃들과 비슷할 것이다. -> 주변의 이웃들의 class를 조사한 후 voting
    - 기본적으로는 이웃의 수 k만 정하면 됨



실습

- ROC curve
  - 예측 모델의 성능을 평가하기 위함
  - 그래프가 좌상향 쪽으로 치우쳐 있으면 성능이 좋은 것



How to use package

- 널리 알려져 있는 기법은 다른 누군가가 구현하여 패키지로 공개해 놓음
- R에서의 패키지는 Pilot study를 하기 좋음
- 좀 더 정확한 분석을 위해서는 논문을 통해 검증된 패키지를 사용하거나 직접 개발하는 것이 좋음
- 어쨌거나 일반적인 알고리즘을 적용하는 것은 어렵지 않음

- 예를 들어, k-nn의 경우,
- 1)"k-nn r package" 구글링
- 2)간편해 보이는 패키지 선택
- 3)첨부된 매뉴얼이나 예시 따라 하기
- 에러가나면,
- 1)에러 메시지를 구글에 검색
- 2)따라 하기

인터넷의 수많은 tutorial

- http://blog.datacamp.com/machine-learning-in-r/
- http://blog.datacamp.com/make-histogram-ggplot2/
- https://www.kaggle.com/c/titanic/details/new-getting-started-with-r
- https://www.kaggle.com/c/facial-keypoints-detection/details/getting-started-with-r
- ….

- Interactive Learning sites for R
  - Datacamp : http://www.datacamp.com
  - Code school : http://www.codeschool.com/courses/try-r
- Other websites
  - http://www.cyclismo.org/tutorial/R/index.html : R 기초 문법
  - http://www.rdatamining.com/ Data Mining 방법론 예제 위주
  - http://caret.r-forge.r-project.org/ 기본적인 DM/ML 알고리즘

결국

- 코드 짜는 것은 검색하면 다 나옴
- 개념을 정확히 알고 알맞은 알고리즘을 쓰는 것이 중요

모델의 종류

- Dimensionality reduction
  - PCA, forward/backward selection…
- Clustering
  - K-means, hierarchical clustering…
- Association rule mining
  - Apriori algorithm
- Classification
  - Logistic regression, KNN, naïve bayes, classification tree, neural network…
- Prediction
  - Linear regression, KNN, neural network, regression tree…

Unsupervised learning algorithm

Supervised learning algorithm

기본적인 분석 Flow

- Spotfire에 데이터 적재 후 탐색 및 시각화
  - 데이터가 어떻게 생겼나, 특성이 무엇인가, 어떤 분석을 할 것인가…

- 전처리
  - Spotfire에서 예쁜 table을 만들고 R로 올리는 것이 편함
  - Dummy variable, missing value imputation, outlier removal, normalization …
  - 전처리도 몇 가지로 바꿔가면서 반복 작업할 부분은 R에서 하는 것이 좋음

- R로 모델링 및 결과 확인
  - Data partition
  - Feature selection
  - Parameter 찾기
  - modeling

- 결과 해석

Other issues

- Novelty detection

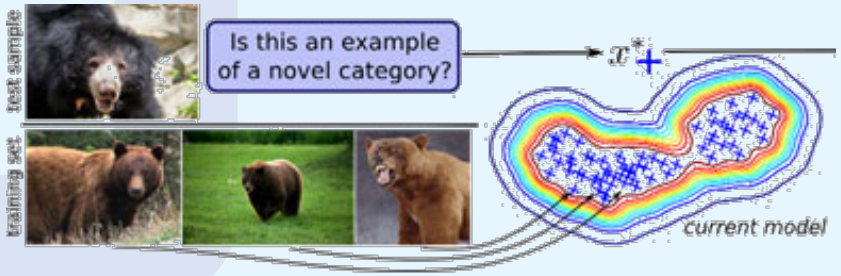
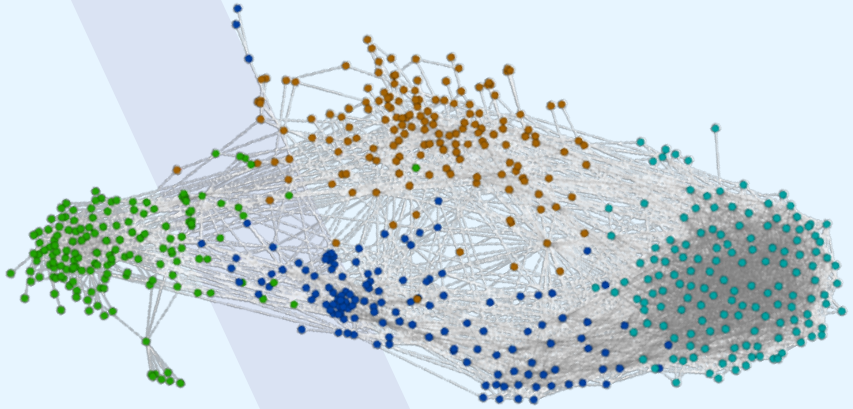


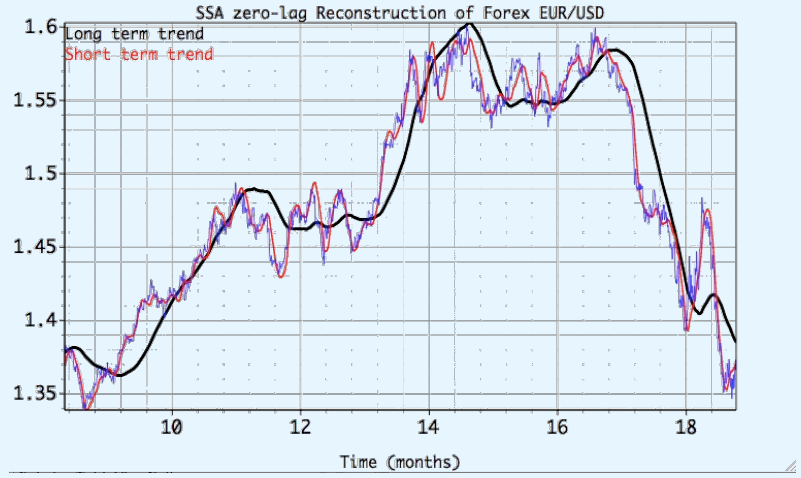
FIGURE 1 EXAMPLE OF NOVELTY DETECTION, BODESHEIM (2012).

- Text mining

- Network analysis



- Time series analysis



기타

- 프로젝트에 관해서
  - R로 안해도 됨
  - 수업에서 배운 알고리즘 외에도 사용해도 됨
  - Business implication 중요
    - 이 분석을 왜 하나?
    - 기대 효과는?

- R 과제
  - 책 18.3 Tayko software cataloger (영문판 379p)

Q & A