# K-Nearest Neighbors(K-NN) 알고리즘을 통한 KOSPI200 선물지수 예측효과 연구

김명현, 이세호, 신동훈, 대한경영학회지, 2015.10 (21 pages)

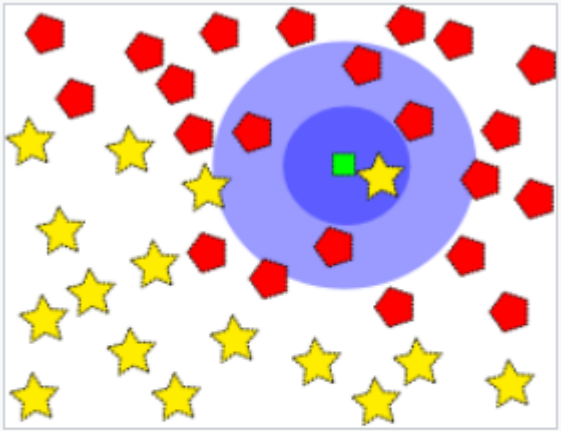
KAIG 세미나 (190305)

# 요약

- 본 논문에서는 머신러닝의 패턴분석기법 중 하나인 K-nearest neighbors(K-NN) 알고리즘을 KOSPI200 선물지수에 적용하여 기술적 분석의 예측력을 검증함.
- K-NN 알고리즘을 선택한 이유
  - 기계 학습의 방법 중 가장 간단한 방법으로 분류되며 모형 위험(Modeling Risk)를 최소화 할 수 있다는 장점 존재
  - 비정상성의 동학을 갖는 가격 레벨에서 분석을 진행할 수 있기 때문에 실제 시장참여자들의 투자패턴을 그대로 적용하는데 용이함
- 기존의 중요 기술적 지표들에 K-NN 알고리즘을 결합할 경우, 기술적 지표 자체를 이용한 투자전략보다 뛰어난 거래결과를 보임을 확인함

# Nearest Neighbors Algorithm

- **최근접 이웃 탐색**(영어: nearest neighbor search)은 가장 가까운 (또는 가장 근접한) 점을 찾기 위한 최적화 문제이다.
- 최근접 이웃 탐색문제는 최적화 문제의 하나로, n개의 데이터가 주어져있을 때, 어떠한 요청에 대한 응답으로 n개의 데이터 중 가장 비슷한 것을 고르는 문제이다. 이 때, 데이터는 $R^d$ 공간 위 점으로 표현된다. 가장 비슷한 점은 보통 유클리드 거리가 최소인 점을 말하지만, 때에 따라서는 특정 목적 함수를 최소화 시키는 점이기도 하다.

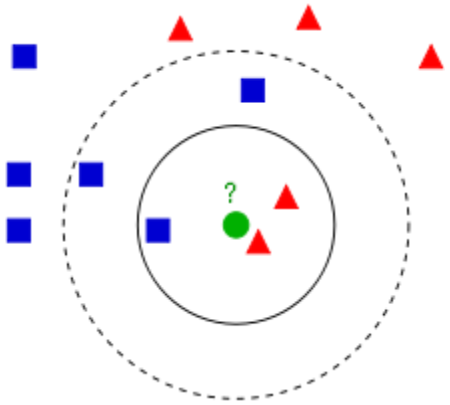


2차원 평면에서 주어진 테이터에서 가장 근접한 점을 찾는다. 이 그림의 경우에는 카테고리가 지정되어 있는 점에서 가장 비슷한 것을 고르는 문제이다.

KAIG

# K-Nearest Neighbors(K-NN)

- 패턴 인식에서, ***k*-최근접 이웃 알고리즘**(또는 줄여서 ***k*-NN**)은 분류나 회귀에 사용되는 비모수 방식이다.
- 두 경우 모두 입력이 특징 공간 내 k개의 가장 가까운 훈련 데이터로 구성되어 있으며, 출력은 *k*-NN이 분류로 사용되었는지 또는 회귀로 사용되었는지에 따라 다르다.
- *k*-NN 알고리즘은 가장 간단한 기계 학습 알고리즘에 속한다.




예시

검증 표본(초록색 원)은 첫 번째 파랑 네모의 항목이나 빨강 삼각형의 두 번째 항목으로 분류되어야 한다. 만약 "k = 3" (실선으로 그려진 원)이면 두 번째 항목으로 할당되어야 한다. 왜냐하면 2개의 삼각형과 1개의 사각형만이 안쪽 원 안에 있기 때문이다. 만약 "k = 5" (점선으로 그려진 원)이면 첫 번째 항목으로 분류되어야 한다. (바깥쪽 원 안에 있는 3개의 사각형 vs. 2개의 삼각형).

# K-Nearest Neighbors(K-NN)

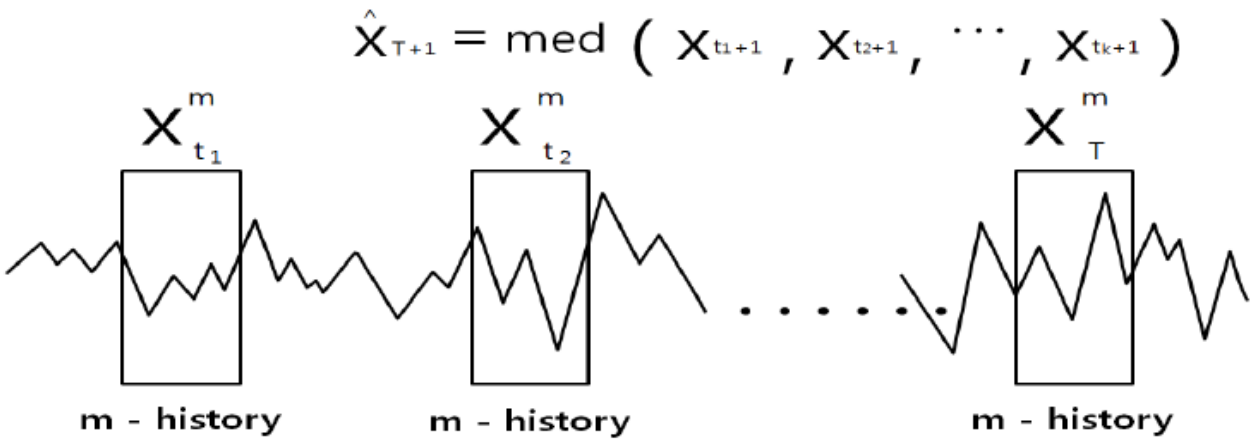



*Figure 1.* **Nearest Neighbor Prediction Technique:** Using non-overlapping m-period K sample data among all financial time series data, it predicts most present observation as the median value of the K sample data.

NN 예측 기술: 모든 차트 데이터 내의 m 주기의 K 샘플 데이터를 사용하여 가장 최근의 관측 데이터가 들어왔을 때, 데이터의 적합한 Value를 예측함

# K-nearest neighbors (KNN) 알고리즘과 Pseudocode

Input: D, set of K training objects, and test object $z = (x', y')$

Process:

1) Compute $d(x', x)$, the distance between $z$ and every object, $(x, y) \in D$

2) Select $D_z \subseteq D$, set of K closest training objects to $z$

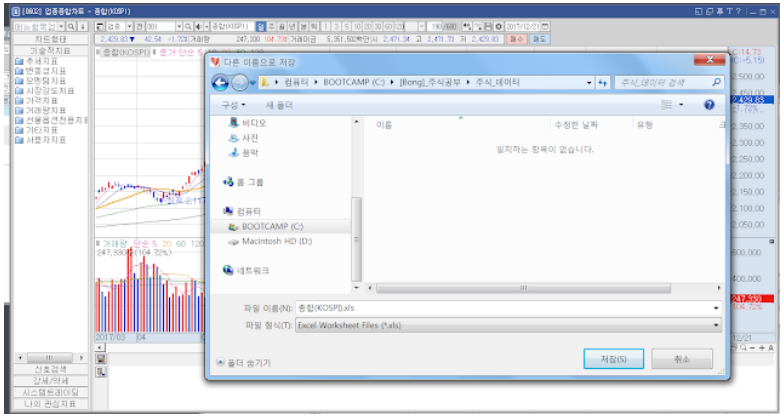
Output: $y' = \arg\max_v \sum_{(x_i, y_i) \in D_z} I(v = y_i)$

# K-nearest neighbors (KNN) 알고리즘과 Pseudocode

```
Function KNN
Input: A finite set D of points to be classified
       A finite set T of points
       A function c: T-> {1, ···, m}
       A natural number k
Output: A function r: D-> {1, ···, m}
Begin
For each x in D do
 Let U <- {}
 For each t in T
 Add the pair (d(x, t), c(t)) to U
 Sort the pairs in U using the first components
 Count the class labels from the first k elements from U
 Let r(x) be the class with the highest number of occurrence
End For each
Return r
End
```

# 사용 데이터

- 본 논문은 거래일 기준 2010년 12월 29일부터 2014년 12월 19일까지 총 1452개 의 KOSPI200 선물 최근 월물 거래가격을 이용해서 분석한다

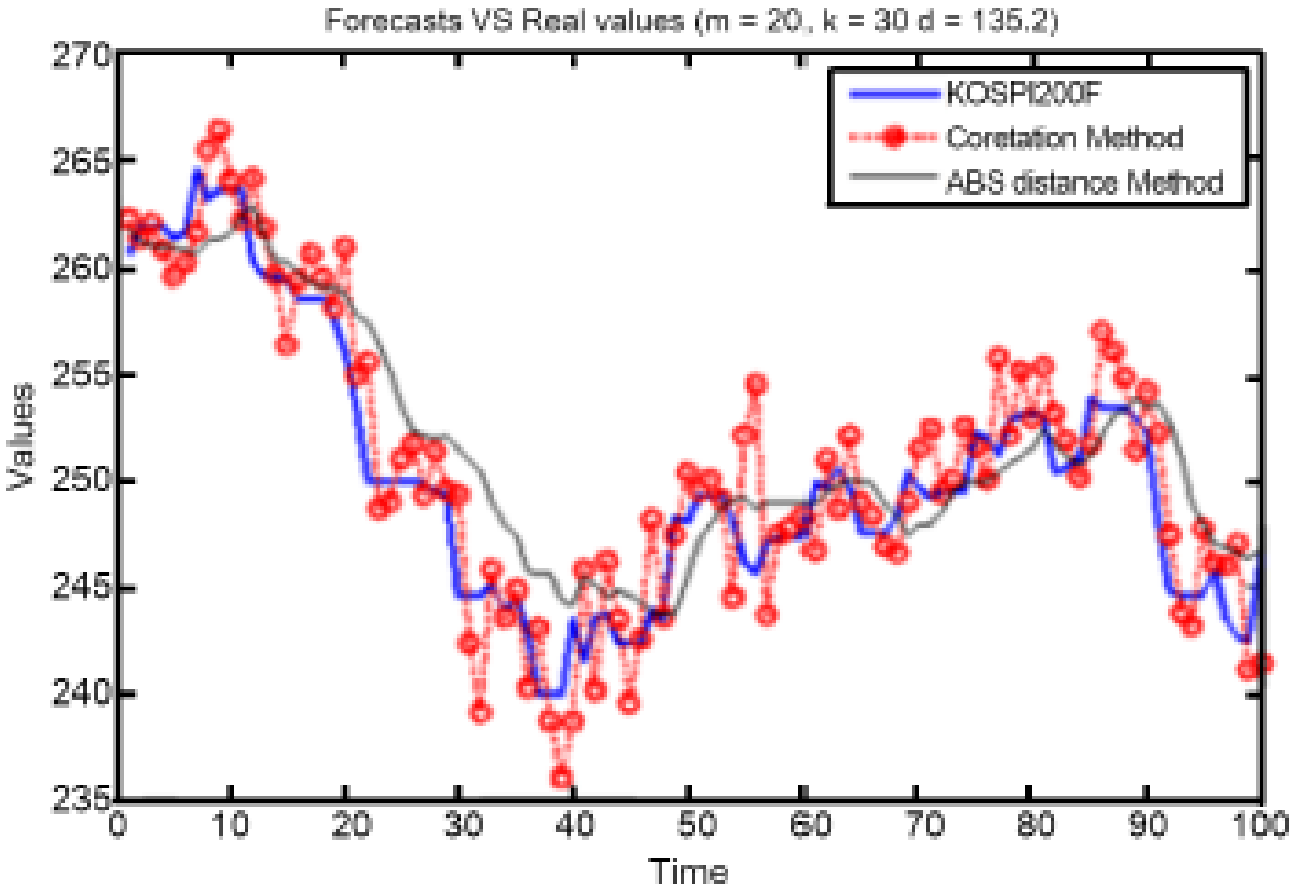


| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 날짜 | 시가 | 고가 | 저가 | 종가 | 종가 단순 5 | 10 | 20 | 60 |
| 2 | 2017/12/21 | 2471.34 | 2471.73 | 2429.83 | 2429.83 | 2468.936 | 2469.12 | 2483.518 | 2487.8 |
| 3 | 2017/12/20 | 2473.82 | 2484.5 | 2470.62 | 2472.37 | 2476.866 | 2472.335 | 2488.884 | 2487.507 |
| 4 | 2017/12/19 | 2487.83 | 2498.67 | 2470.49 | 2478.53 | 2478.502 | 2472.535 | 2492.291 | 2486.568 |
| 5 | 2017/12/18 | 2488.83 | 2488.83 | 2477.29 | 2481.88 | 2474.996 | 2475.694 | 2494.899 | 2485.563 |
| 6 | 2017/12/15 | 2488.39 | 2495.08 | 2475.99 | 2482.07 | 2472.918 | 2477.673 | 2497.189 | 2483.966 |
| 7 | 2017/12/14 | 2486.57 | 2514.61 | 2469.48 | 2469.48 | 2469.304 | 2477.007 | 2499.785 | 2482.226 |
| 8 | 2017/12/13 | 2462.42 | 2481.92 | 2457.7 | 2480.55 | 2467.804 | 2477.696 | 2503.05 | 2480.404 |
| 9 | 2017/12/12 | 2471.23 | 2471.43 | 2455.39 | 2461 | 2466.568 | 2480.931 | 2504.935 | 2478.486 |
| 10 | 2017/12/11 | 2467.81 | 2471.99 | 2457.98 | 2471.49 | 2476.392 | 2486.25 | 2508.217 | 2476.787 |
| 11 | 2017/12/08 | 2470.09 | 2473.17 | 2459.02 | 2464 | 2482.428 | 2489.882 | 2511.16 | 2474.658 |
| 12 | 2017/12/07 | 2479.55 | 2480.85 | 2452.4 | 2461.98 | 2484.71 | 2497.915 | 2515.108 | 2472.694 |
| 13 | 2017/12/06 | 2510.19 | 2510.47 | 2474.29 | 2474.37 | 2487.588 | 2505.432 | 2519.537 | 2470.325 |
| 14 | 2017/12/05 | 2490.73 | 2513.68 | 2487.15 | 2510.12 | 2495.294 | 2512.046 | 2523.439 | 2467.862 |
| 15 | 2017/12/04 | 2486.72 | 2501.67 | 2476.29 | 2501.67 | 2496.108 | 2514.104 | 2525.205 | 2464.855 |

# 결론

- 절대거리와 상관계수 두 방법론의 K-NN 알고리즘을 이용, 과거 가격패턴으로부터 의미 있는 예측성과를 검증함.
- 데이터의 특성에 따라 단일 시계열 방법론을 적용함
- 통계적 예측 오차를 측정하는 오차제곱 합을 제곱근한 RMSE 기준으로 분석, 상관계수 방법론은 실현된 KOSPI200 선물지수 값 근처를 중심으로 변동(Variations)을 보이고 있지만 실현된 값을 평균적으로 잘 예측함을 밝혀냄
- 절대거리 방법론의 경우, 선물지수 하락기에 실현된 값보다 큰 값을 지속적으로 예측하고, 지수 횡보기에는 변동을 보이는 상관계수 방법론보다 안정적인 예측력을 보이고 있음을 발견

# 결론


Forecasts VS Real values (m = 20, k = 30 d = 135.2)
KOSPI200F
Coretation Method
ABS distance Method
Values
Time
270
265
260
255
250
245
240
235
0
10
20
30
40
50
60
70
80
90
100

# 차후

- K-NN 목적함수 조사
- 실제 금융 데이터 수집 후 선물옵션차트와 주가-거래량 차트간 차이점 분석
- 주식 거래의 기술적 분석에 영향을 주는 요소 분석
- 본 논문은 머신러닝 패턴분석 기법을 활용해 기술적 분석에 활용되는 요소들의 예측성과를 측정하여 유효성을 검증함. 대표적으로 미결제약정과 차익/비차익 프로그램 순매수인데, 이들 외 다른 유효한 요소가 있을 지 고민 해봄.

- + FeedBack