SK Planet Code Sprint 2013 Round 2

Sungjoo Ha
@shurain_

September 27th, 2013

1 About Me

하성주 (shurain)

- 서울대학교 컴퓨터공학부 B.S.
- 서울대학교 컴퓨터공학부 Ph.D candidate
- SK Planet Code Sprint 2013 Round 2 우승
- Optimization
- Parallel processing
- Machine learning
- @shurain_
- `blog.shurain.net`

2 Introduction

문제 소개

코드스프린트 Round2 문제에서는 2013년 4월부터 6월까지 3개월간의 T map 경부고속도로 교통정보를 제공하고 개발자는 이를 분석하여 7월 16일 24시간의 교통정보를 예측합니다.

- 총 2×126 구간 (상행/하행)
- 5분 단위 데이터
- 평균 시속 예측

데이터 형식

순서	데이터	예시	설명
1	날짜	20130510	2013년 05월 10일
2	시각	2345	23시 45분
3	상행/하행	U	U:상행, D:하행
4	구간 index	24	구간 1 ~ 126까지
5	구간 시작점	건천IC	해당 구간의 시작점 이름
6	구간 끝점	경주터널	해당 구간의 끝점 이름
7	구간 거리	3824	3,824 미터
8	구간 속도	77.78	시속 77.78 km

채점 방식

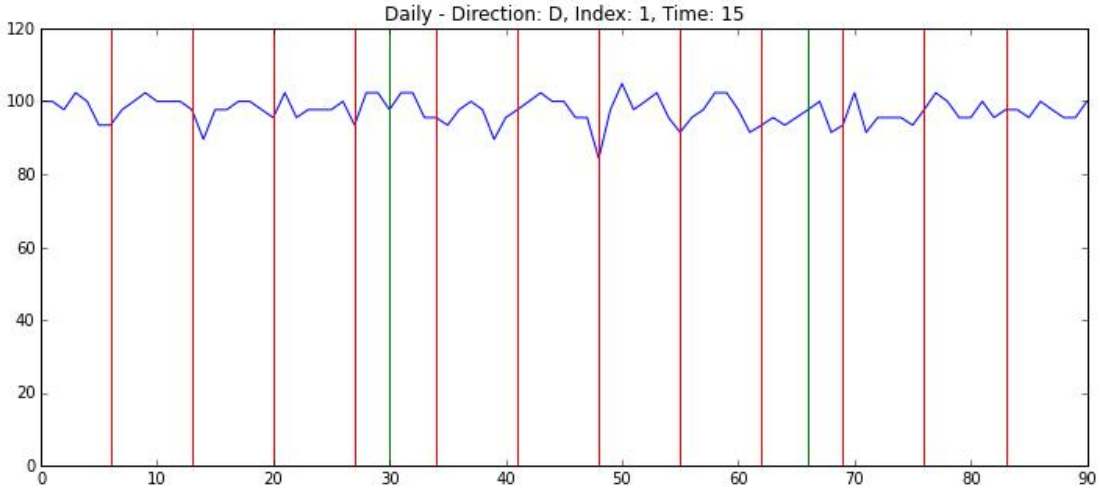
2013년7월 16일 경부고속도로의 실제 교통정보 데이터를 사용하여, 실제 속도와 제출한 속도의 차이를 계산 후 모든 구간에 차이를 합하여 가장 적은 값이 나온 제출자를 선정하게 됩니다.

- Mean Absolute Error (MAE)

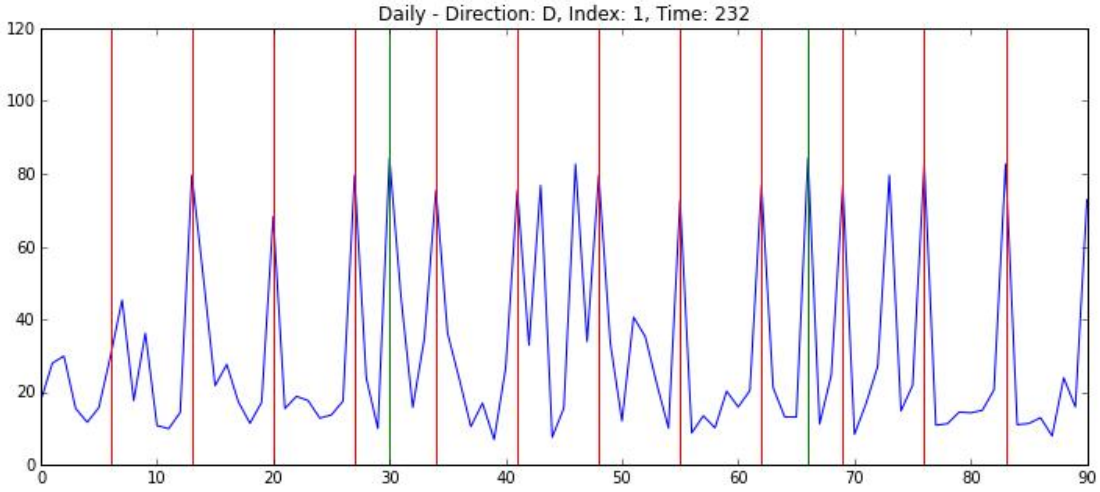
우승에 이르는 전략

- 중앙값을 썼더니 우승
- 모든 예측 문제에 접근하는 자세
  - 데이터를 보고
  - 그럴싸한 모델을 만든다
  - 실험을 통해 검증을 거치고 더 나은 모델을 만든다

데이터 분석
4월 1번 도로 하행선 01:15



데이터 분석
4월 1번 도로 하행선 19:20



가설

- 주말과 주중 데이터는 서로 다른 규칙을 따른다
- 주중 데이터는 서로 같은 규칙을 따른다

모델 생성

- 데이터가 있으니
- 기계 학습 기법에 다 넣고

- 돌린다
- PROFIT???

인과 관계

> 통계 모델에 데이터를 집어 넣고 숫자를 뽑아내서 그 결과가 현실 세계의 올바른 표상이라고 무작정 받아들일 수 있으면 참 좋을 것이다. 하지만 대부분의 경우, **인과 관계를 잘 고려하지 않으면 그럴 가망이 없는 결과만 얻을 뿐이다.**
>
> - Nate Silver

생각하기

- 지금 도로 상황이 보름 전의 도로 상황의 영향을 받는다?
- 말도 안 된다
- 아무리 좋은 기계 학습 기법을 사용해도 보름 전의 도로 상황으로 현재의 도로 상황을 예측하는 것은 힘들 것이다

통계적 접근
모델

- 하루 치 데이터를 생성하는 규칙이 있다고 가정하자
  - 이 규칙을 알면 데이터를 그대로 생성하면 된다

통계적 접근
모델

- 하루 치 데이터가 통째로 생성된다고 가정하자
  - 앞뒤 시간은 서로 영향을 주고 받으니까
  - 하지만 이것은 어려운 문제
    - 전체적인 움직임과 국소적인 움직임을 모두 고려해야 한다
    - 어렵다

통계적 접근
모델

- 각 데이터는 모두 시간에 대해 독립적으로 생성된다고 가정하자
  - 거짓말
  - 하지만 이제 쉬운 문제가 된다
  - 이제 변수 하나씩만 예측을 72,576 번 하면 된다

통계적 접근
예제

- 내가 갖고 있는 하루하루 데이터가 하나하나 정답이라고 가정하자
- 내가 어떤 값을 제출하면 위의 데이터에서 평균적으로 가장 좋은 결과를 얻을 수 있을까?

$$\operatorname*{argmin}_{\hat{X}} \sum_i |\hat{X} - X_i|$$

- 중앙값이 이에 대한 최적해라는 이론이 있다
- An optimality property of median

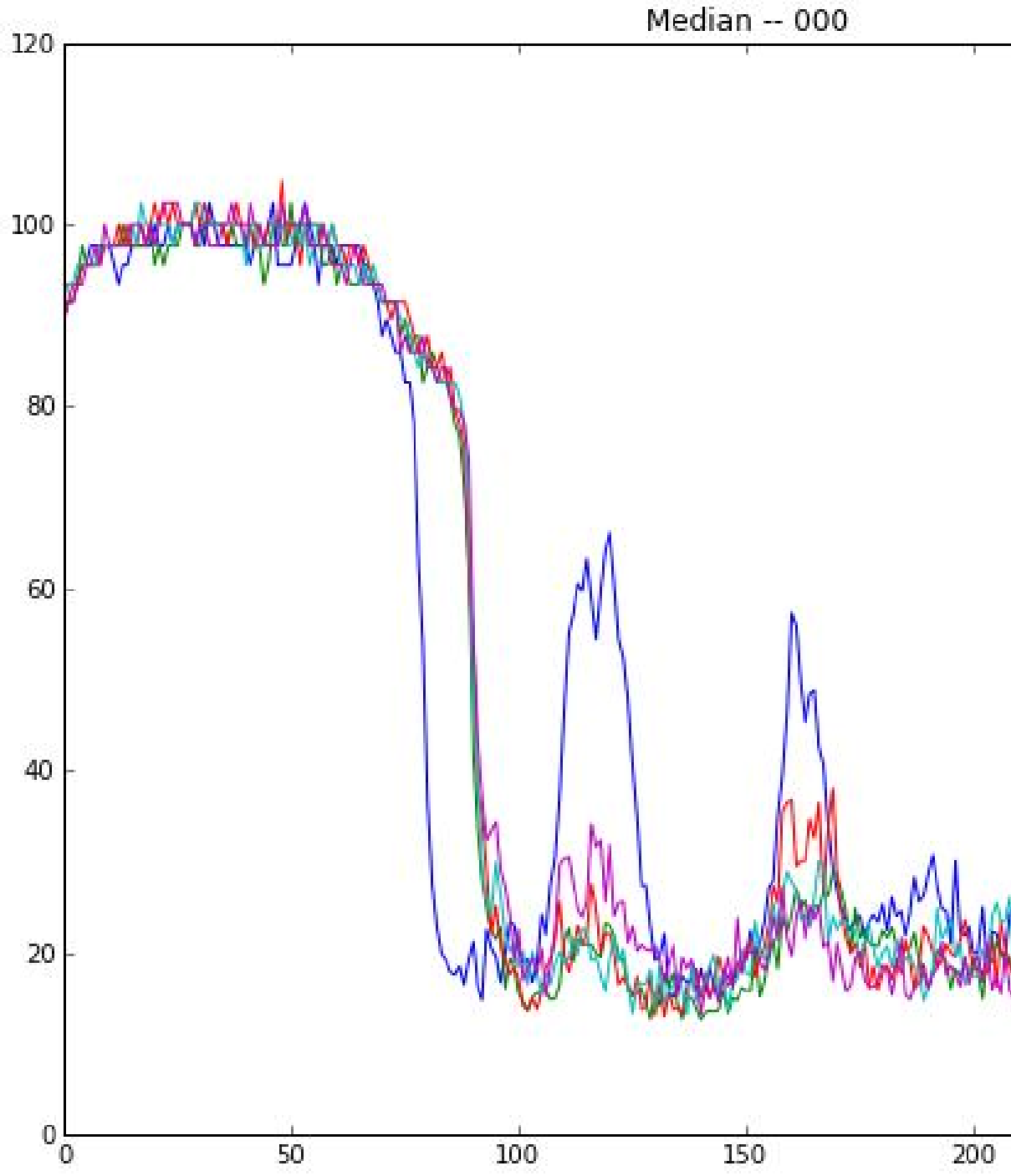
통계적 접근
결론

- 주어진 채점 기준에 (**MAE**) 가장 적절한 값은 모든 데이터의 중앙값 (median)
- 앞서 언급한 **가정을 모두 적용**하면 이 데이터만 사용해서 얻을 수 있는 이론적 최적값
  - 멋져 보이는 기계 학습 기법보다 더 좋은 결과가 보장됨

Result

순위	이름	제출일시	성적
1	Sungjoo Ha	2013-07-14 23:39	4.7263

Future Work
요일별 중앙값 분포의 차이



Future Work

더 정교한 모델

- 추가 데이터 사용
  - 날씨, ...
  - 더 많은 과거 데이터
- "진짜" 베이지안 접근
  - 데이터 생성 모델링
  - Prior belief 설정
  - Markov Chain Monte Carlo
  - Posterior 샘플로부터 손실 함수의 기대값 최적화
  - 블로그 글 참조
- 모델 앙상블

Reference

- The Signal and the Noise
- http://blog.shurain.net/2013/07/code-sprint-2013-round-2.html
- http://blog.shurain.net/2013/08/code-sprint-2013-round-2-post-mortem.html
- http://en.wikipedia.org/wiki/Median

Thank You!

bit.ly/codesprint2013r2
https://github.com/shurain/codesprint2013