빅데이터 분석을 위한 실시간 로지스틱 회귀모형에 관한 연구 : 베이지안 접근법을 중심으로

김동완

고려대학교 정책대학원
데이터 통계학과

2016년 6월 3일

Overview

1. 끓말
   - 개요
2. 온라인 최적화 알고리즘과 해싱을 이용한 가변수 코딩
   - 해싱을 이용한 가변수 코딩
   - 온라인 최적화 방법
3. 사례연구
   - 타이타닉 탑승자 데이터
   - 온라인 광고 데이터
   - 맺음말

새로운 형태의 데이터

- 실제 IT 분야에서 해결해야하는 문제
  - TrueSkill과 같이 다량의 플레이어 게임 메칭 데이터를 이용해 플레이어의 승률을 계산하여 최적의 게임 메칭 상대를 찾는 문제
  - 특정 Facebook 사용자의 timeline에 다양한 조건의 광고 중 어떤 광고를 노출 시켜야 광고 클릭 확률이 높을 것인지를 예측하는 문제
  - 매출의 대부분을 차지하는 고부가 가치 유저(High-Valued Player)가 게임에서 이탈할 확률을 계산하는 문제
  - 온라인 광고 퍼블리싱 상황에서 실시간으로 사이트마다 최적의 광고 선택 문제
- 이런 문제들의 특징
  - 유동적인 다 범주 변수
  - 多 샘플, 실시간 분석

새로운 형태의 데이터

- 접근 방법
  - 유동적인 다 범주 변수
    → 해싱을 이용한 가변수 코딩(Feature Hashing)
  - 多 샘플, 실시간 분석
    → 온라인 최적화
    - 확률적 경사 하강법(SGD) vs 추정된 밀도 필터링(ADF)
- 의문점
  - 기존의 배치 방식 대비 예측률
  - 대용량 데이터에 대한 분석 속도

해싱을 이용한 가변수 코딩

- 해시 함수
  - 해시 함수는 임의의 길이의 데이터를 고정된 길이의 데이터로 매핑하는 알고리즘.
  - 암호화에 많이 사용됨
- 종류
  - 암호화 해시
    : 복호화가 어려워야 함(MD5, SHA)
  - 비암호화 해시
    : 분포적 특성과 속도가 중요(Murmur, City, Spooky)
    $\rightarrow$ 데이터 분석에 사용

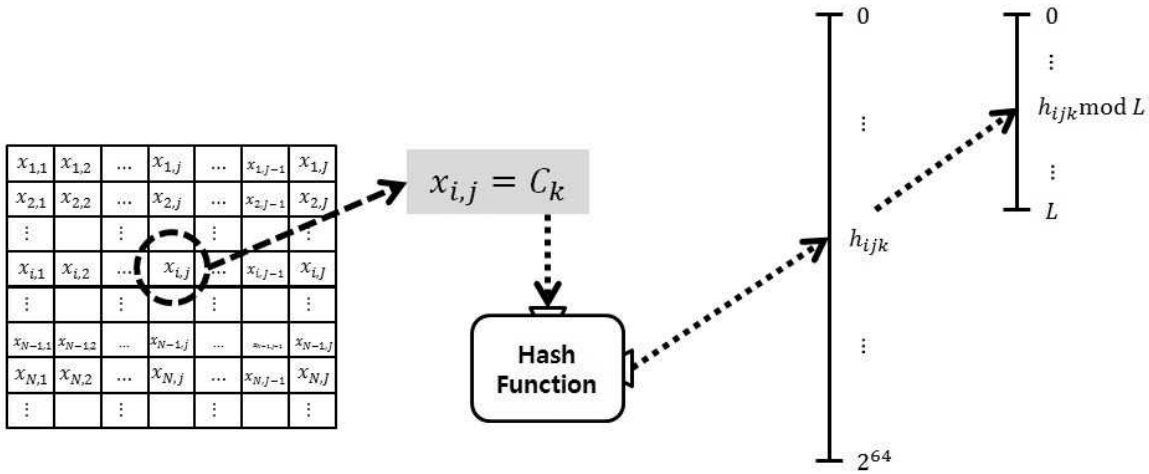
해싱을 이용한 가변수 코딩

- 해시 함수
  - 해시 함수는 임의의 길이의 데이터를 고정된 길이의 데이터로 매핑하는 알고리즘.
  - 암호화에 많이 사용됨
- 종류
  - 암호화 해시
    : 복호화가 어려워야 함(MD5, SHA)
  - 비암호화 해시
    : 분포적 특성과 속도가 중요(Murmur, City, Spooky)
    $\rightarrow$ 데이터 분석에 사용

해싱을 이용한 가변수 코딩

- 데이터가 두개의 범주형 변수 $V_1 \in \{c_1, c_2, c_3\}$과 $V_2 \in \{c_4, c_5, c_6\}$로만 구성되어 있다고 할 때,
  - 일반적인 가변수 코딩
    : n건의 데이터에서 i번째 데이터 벡터 $[x_{i1}, x_{i2}]$는
    $[b_{11}, b_{12}, b_{21}, b_{22}], b_{ij} \in \{0, 1\}$의 4자리 벡터로 가변수 코딩
  - 해싱을 가변수 코딩
    : 데이터 값을 각각에 대한 해시 함수$(H(\cdot))$의 결과 값 벡터
    $[h_{i1} = H(x_{i1}), h_{i2} = H(x_{i2})], h_{ij} \in \mathbb{R}_{>0}$로 코딩
    : 해시 함수의 최대 비트 수를 b라 할때, h_{ij}의 최대값은 2^b이고,
    실제로 필요한 예상되는 변수의 수가 L개로 제한된다면
    $2^b \ mod \ L$을 h_{ij} 대신 사용

해싱을 이용한 가변수 코딩



- 해싱을 이용한 가변수 코딩은 아래의 경우에도
  빈번한 가변수 코딩을 수행할 필요가 없어 효과적
  - 범주형 변수의 수와 각 범주를 구성하는 범주가 많을 경우
  - 범주의 수가 지속적으로 추가될 수 있는 경우
  - 실시간 분석이 요구되는 경우

확률적 경사 하강법(Stochastic Gradient Descent)

- 최적화(optimization) 알고리즘의 하나인 경사하강법(Gradient descent)에서는 전체 샘플 데이터를 스캔 할 때마다 회귀 계수 추정치를 갱신
- 비용함수(cost function)를 $J(w) = \frac{1}{2}\sum_i (target^{(i)} - output^{(i)})^2$ 라 할때 길이가 j인 계수 벡터(weight vector) w_i를 $w_{i+1} = w_i + \Delta w$로 갱신하는 매 반복에서 j개의 모수 추정치 (w)를 얼마만큼 줄일지 혹은 늘릴지 Δw 값을 결정

$$\begin{aligned} \Delta w_j &= -\alpha \frac{\delta J}{\delta w_j} && (1) \\ &= -\alpha \sum_i (target^{(i)} - output^{(i)})(-x^{(i)_j}) && (2) \\ &= \alpha \sum_i (target^{(i)} - output^{(i)})(x^{(i)_j}) && (3) \end{aligned}$$

- 전체 샘플 데이터를 사용하는 대신 샘플 하나 혹은 일부분만을 사용하여 w 값을 갱신해 가는 경사하강법을 확률적 경사 하강법이라 함

추정된 밀도 필터링(Assumed-density filtering)

- ADF에서는 사후분포를 가우시안과 같은 특정 분포로 근사하는 방법으로서 예측-갱신-투영(predict-update-project)과정을 반복
  - 예측(predict)
    : 모수 θ에 대한 $t-1$ 시점의 사전분포$q_{t-1}(\theta_{t-1})$와
    t시점의 관측치를 이용하여
    이후 시점 t에서의 θ에 대한 사후예측분포, $q_{t|t-1}(\theta_t)$를 구함
  - 갱신(update)
    : 앞서 구한 사전분포와 사후예측분포를 이용하여
    θ에 대한 사후분포, $\hat{p}(\theta_t)$를 구함
  - 투영(project)
    : 마지막으로 이 사후 분포가 다루기 쉬운 형태가 아닌 경우가
    빈번하기 때문에 다루기 쉬운 분포로 투영(project)

추정된 밀도 필터링(Assumed-density filtering)

- 근사 사전분포:

$$q_{t-1}(\theta_{t-1}) \approx p(\theta_{t-1}|y_{1:t-1})$$

- 1단계 사후예측분포:

$$q_{t|t-1}(\theta_t) = \int p(\theta_t|\theta_{t-1}) q_{t-1}(\theta_{t-1}) d\theta_{t-1}$$

- 사후분포:

$$\hat{p}(\theta_t) = \frac{1}{Z_t} p(y_t|\theta_t) q_{t|t-1}(\theta_t)$$

- 정규화 상수(normalizing constant):

$$Z_t = \int p(y_t|\theta_{t-1}) q_{t|t-1}(\theta_t) d\theta_t$$

- 근사 사후분포:

$$q(\theta_t) = \arg\min_{q \in Q} \mathrm{KL}(\hat{p}(\theta_t)||q(\theta_t))$$

추정된 밀도 필터링(Assumed-density filtering)

- 일반화 선형 모형에서의 가우시안 근사
  : 투영하려는 분포 q가 지수족 이므로 moment matching 대응 사용
  - 일반화 선형 모형의 systematic component를 $s_t = \theta_t^T x_t$라 하면 s_t의 사후 예측분포 $q_t(s_t)$는

$$
\begin{aligned}
q_t(s_t) &\equiv N(s_t; m_t, v_t) \\
m_t &= \int s_t \frac{1}{z_t} f(y_t|s_t) q_{t|t-1}(s_t) ds_t \\
v_t &= \int s_t^2 \frac{1}{z_t} f(y_t|s_t) q_{t|t-1}(s_t) ds_t - m_t^2 \\
z_t &= \int f(y_t|s_t) q_{t|t-1}(s_t) ds_t \\
f(y_t|s_t) &\equiv Ber(y_t; \pi = sigmoid(s_t)) \\
&= \pi^{y_t}(1-\pi)^{(1-y_t)}, \quad y_t \in \{0, 1\}
\end{aligned}
$$

추정된 밀도 필터링(Assumed-density filtering)

- 일반화 선형 모형에서의 가우시안 근사
  - 가우스-에르미트 구적법을 사용하여 s_t의 사후분포를 근사하면,

$$
\begin{aligned}
q_t(s_t) &= N(s_t; \tilde{m}_t, \tilde{v}_t) \\
\tilde{m}_t &= \frac{1}{\tilde{z}_t} \sum_i \chi_i f(y_t; \chi_i) \omega_i \\
\tilde{v}_t &= \frac{1}{\tilde{z}_t} \sum_i \chi_i^2 f(y_t; \chi_i) \omega_i - \tilde{m}_t^2 \\
\tilde{z}_t &= \sum_i f(y_t; \chi_i) \omega_i
\end{aligned}
$$

  - $q_{t|t-1}(s_{t-1})$을 $q_t(s_t)$로 갱신한 후 평균과 분산의 변화를 각각 $\sigma_m = m_t - m_{t|t-1}$과 $\sigma_v = v_t - v_{t|t-1}$라고 하면, t시점의 i번째 θ의 분포는

$$
q(\theta_t, i) \sim N(\theta_{t,i}; \mu_{t,i}, \sigma_{t,i}^2) \tag{4}
$$

$$
\mu_{t,i} = \mu_{t|t-1,i} + a_i \delta_m, \quad \sigma_{t,i}^2 = \sigma_{t|t-1,i}^2 + a_i^2 \delta_v, \quad a_i \triangleq \frac{x_{t,i} \sigma_{t|t-1,i}^2}{\sum_j x_{t,j}^2 \sigma_{t|t-1,i}^2}
$$

소규모 데이터를 이용한 분석

- 타이타닉 탑승자들의 여러 특성에 따른 생존 여부를 나타내는 889건의 데이터를 800건의 훈련 데이터와 나머지 89건을 테스트 데이터로 나누고, 아래 3개 방법을 이용한 예측 성능을 비교
  - 최대우도 추정(MLE)
  - 확률적 경사 하강법(SGD)
  - 추정된 밀도 필터링(ADF)

	TP	**FP**	**FN**	**TN**	**Accu**	**Prec**	**Recall**	**F1-Score**	*logloss*[1]
MLE	24	9	5	41	0.843	0.727	0.826	0.774	0.423
SGD	22	11	6	50	0.809	0.667	0.786	0.721	0.411
ADF	21	12	7	49	0.787	0.636	0.750	0.688	0.437

1) $logloss = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(p_i) + (1-y_i)\log(1-p_i)\right)$

대규모 데이터를 이용한 분석

- 사례 분석을 위해 Criteo에서 데이터 분석 경연을 위해 공개한 4천 5백만건 상당의 온라인 광고 데이터 사용
  - 데이터는 웹사이트 방문자가 해당 광고를 클릭 했으냐 혹은 하지 않았느냐를 나타내는 이항 반응변수와 광고의 특성을 나타내는 39개의 설명변수로 구성
  - 각 설명변수는 범주형으로서 각 변수의 범주는 500개 이상이고 점차 새로운 범주가 등장

Training data size: 10^3, **Test data size:** 10^3

	Accu	**Prec**	**Recall**	**F1-Score**	*logloss*[1]	**exec time(sec)**
SGD	0.716	0.380	0.526	0.441	0.570	0.308
ADF	0.702	0.343	0.437	0.384	0.570	0.461

대규모 데이터를 이용한 분석

Training data size: 10^5, **Test data size:** 10^5

	Accu	**Prec**	**Recall**	**F1-Score**	*logloss*[1]	**exec time(sec)**
SGD	0.710	0.460	0.674	0.547	0.564	22.602
ADF	0.699	0.449	0.690	0.544	0.575	30.981

Training data size: 10^7, **Test data size:** 10^7

	Accu	**Prec**	**Recall**	**F1-Score**	*logloss*[1]	**exec time(sec)**
SGD	0.712	0.460	0.707	0.557	0.560	2148.999
ADF	0.713	0.461	0.709	0.558	0.559	2756.530

- 데이터 크기가 작을 경우(10^3) SGD가 다소 나은 예측 성능을 보임.
- 데이터 크기가 커질수록 두 모형의 성능 차이는 줄어들고 ADF가 다소 나은 성능을 보임
- 실행 속도는 3가지 데이터 크기(10^3, 10^5, 10^7)에서 모두 SGD가 빠르나, 그 차이는 44%, 37%, 28%로 데이터 크기가 커질수록 줄어듬.

맺음말

- 多 범주의 多 변수 데이터 분석에 있어서 해싱을 이용한 가변수 코딩의 유용성 확인
- 온라인 방식의 적합 방법의 성능이 배치 방식에 비해 크게 떨어지지 않음.
- 대규모 데이터에 대해서 '근사 방법을 이용한 베이지안 방법론'이 충분히 좋은 속도와 성능을 보임
- 확률적 경사하강법에서는 모수별 step size의 초기값, 추정된 밀도 필터링에서는 모수별 초기 분산의 초기값에 예측 성능이 크게 영향을 받으나 데이터 크기가 커질수록 그 영향은 줄어듬.
- 변수 선택이나 변수간 교호작용 등 추후 고려해볼 것들이 있음.

감사합니다.