가우시안 구적법을 이용한 온라인 베이지안 로지스틱 회귀 모형에 관한 연구

김동완

2016년 3월 5일

차 례

제 1 장

서론

1.1 개요

모수에 대한 학습(learning)이나 추론(inference)를 진행할 때 데이터의 크기가 작거나 데이터의 수집으로부터 예측까지 시간적 여유가 있을 경우 데이터 전체를 한꺼번에 활용하는 일괄 처리(batch processing) 방식을 사용한다. 반면 데이터를 한꺼번에 처리하기에 그 크기가 지나치게 크거나 스트리밍(streaming)으로 유입되는 데이터에 대해서 실시간으로 예측을 처리해야 하는 경우 온라인 학습(online learning)을 사용해야 할 필요성이 있다.(Opper, 1999) 특히 온라인 학습에 있어서 베이지안 방법론을 이용한 접근은 Opper (1996) 가 제안하였다. 본 논문에서는 추정된 밀도 필터링(Assumed-density filtering, ADF)방법을 이용한 베이지안 온라인 학습 기법에 대해서 고찰해 보도록 한다.

제 2 장

가우시안 구적법을 이용한 온라인 베이지안 로지스틱 회귀 모형

2.1 로지스틱 회귀 모형

일반화 선형 모형은 크게 i) 반응변수 Y와 그것의 확률 분포, ii) 설명변수와 회귀계수의 선형식(systematic component), $\eta = \theta^T x$, iii) 연결함수(link function), $g(\cdot)$ 로 구성된다. 이 3가지 구성요소는 아래와 같이 '반응변수 Y의 기댓값 μ와 선형식 η의 관계가 연결함수로 표현되는 형태로 결합된다.(Agresti, 1996)

$$E(Y) = \mu = g^{-1}(\theta^T x)$$

로지스틱 회귀 모형은 일반화 선형 모형의 한 가지 형태로서 반응변수 Y가 이항분포를 따르고 연결함수가

$$g(\mu) = logit(\mu) = log\left(\frac{\mu}{1-\mu}\right)$$

위와 같은 log-odds(logit)인 경우를 말한다.

결국 반응변수 Y는 성공확률 π가 $g^{-1}(\theta^T x)$인 이항분포로서, 아래와 같이 π를 회귀계수 혹은 가중치(weight)와 설명변수의 선형결합을 인자로 하는 로지스틱(logistic, sigmoid) 함수로 표현할 수 있다.

$$\pi_i = P(y_i = 1) = \frac{1}{1 + exp(-\theta_i^T x_i)}$$

$$\begin{aligned} logit(E[Y]) &= logit(P(Y=1)) && (2.1)\\ &= logit(\pi(x)) && (2.2)\\ &= log\left(\frac{\pi(x)}{1-\pi(x)}\right) && (2.3)\\ &= \theta^T x && (2.4) \end{aligned}$$

2.2 추정된 밀도 필터링(Assumed-density filtering)

추정된 밀도 필터링(Assumed-density filtering, ADF)는 베이지안 네트워크 혹은 여타의 통계 모형에서 사후분포를 근사적으로 계산하는 방법으로서 통계학에는 Lauritzen (1992)에서 제안된 바 있다. 또한 분야에 따라 "추정된 밀도 필터링(Assumed-density filtering)", "온라인 베이지안 학습(On-line Bayesian learning)", "적률 대응(Moment matching)", "약한 주변화(Weak marginalization)"이라 부르기도 한다. (Minka, 2013)

ADF에서는 사후분포를 가우시안과 같은 특정 분포로 근사하는 방법으로서 예측-갱신-투영(predict-update-project)과정을 반복한다. 예측(predict) 과정에서는 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 구하고, 갱신(update) 과정에서는 앞서 구한 사전분포와 사후예측분포를 이용하여 θ에 대한 사후분포, $\hat{p}(\theta_t)$를 구한다. 마지막으로 이 사후 분포가 다루기

쉬운 형태가 아닌 경우가 빈번하기 때문에 다루기 쉬운 분포로 투영(project)하는 과정을 거치게 된다.

- 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:

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

근사 사후분포를 구할 때 위와 같이 쿨백-라이블러 발산값(Kullback-Leibler divergence)을 최소화하는 함수 $q(\theta_t)$를 구하는데 이는 (다루기 어려운)사후분포$\hat{p}(\theta_t)$를 다루기 쉬운 분포 공간으로 투영(project)하는 것이라 생각할 수 있다. 그런데 투영하려는 분포 q가 지수족에 속할 경우 단순히 적률대응(moment matching)만으로 $q(\theta_t)$를 구할수 있다.(Murphy, 2012)

2.3 일반화 선형 모형에서의 가우시안 근사

편의를 위해 설명변수와 회귀계수의 선형식(systematic component)을 $s_t = \theta_t^T x_t$라 하자. 만약 θ_t에 대한 1단계 사후예측분포, $q_{t|t-1}(\theta_t)$가 $\prod_i N(\theta_{t,i}; \mu_{t|t-1,i}, \sigma^2_{t|t-1,i})$

라면 s_t의 사후 예측분포, $q_{t|t-1}(s_t)$ 는 아래와 같다.

$$
\begin{aligned}
q_{t|t-1}(s_t) &\equiv N(s_t; m_{t|t-1}, v_{t|t-1}) && (2.5)\\
m_{t|t-1} &= \sum_{i=1}^{N} x_{t,i}\mu_{t|t-1,i} && (2.6)\\
v_{t|t-1} &= \sum_{i=1}^{N} x_{t,i}^2\sigma_{t|t-1,i}^2 && (2.7)
\end{aligned}
$$

이때 s_t의 사후분포, $q_t(s_t)$는 아래와 같다.

$$
\begin{aligned}
q_t(s_t) &\equiv N(s_t; m_t, v_t) && (2.8)\\
m_t &= \int s_t \frac{1}{z_t} f(y_t|s_t) q_{t|t-1}(s_t) ds_t && (2.9)\\
v_t &= \int s_t^2 \frac{1}{z_t} f(y_t|s_t) q_{t|t-1}(s_t) ds_t - m_t^2 && (2.10)\\
z_t &= \int f(y_t|s_t) q_{t|t-1}(s_t) ds_t && (2.11)\\
f(y_t|s_t) &\equiv Ber(y_t; \pi = sigmoid(s_t)) && (2.12)\\
&= \pi^{y_t}(1-\pi)^{(1-y_t)}, \quad y_t \in \{0,1\} && (2.13)\\
&= \left(\frac{1}{1+exp(-s_t)}\right)^{y_t} \left(\frac{exp(-s_t)}{1+exp(-s_t)}\right)^{(1-y_t)}, \quad y_t \in \{0,1\}
\end{aligned}
$$

위의 적분식을 계산하기 위하여 가우시안 구적법(Gaussian quadrature)을 사용할 수 있다. 가우시안 구적법을 이용하면 어떤 다항식과 알려진 함수 $W(x)$의 곱에 대한 계산을 아래와 같이 다항식 함수값의 가중합으로 근사할 수 있다.

$$\int_a^b W(x)f(x)dx \approx \sum_{j=1}^{N} w_j f(x_j)$$

특히 $W(x) = e^{-x^2}$와 어떤 함수의 곱을 적분하는 경우, $\int_{-\infty}^{+\infty} e^{-x^2} f(x)dx$, 가우스-에르미트 구적법(Gauss-Hermite quadrature)을 사용할 수 있다. χ'를 결정점(sample point)이라 하고 ω'를 가중치(weight)라고 할때, $\chi = \chi'\sqrt{2}\sigma_{s_t} +$

μ_{s_t}와 $\omega_i = \frac{\omega'}{\sqrt{\pi}}$로 변수변환할 수 있다. 이를 이용하여 앞서 2.9, 2.10, 2.11의 적분을 아래와 같이 근사할 수 있다.(Zoeter, 2007)

$$
\begin{aligned}
q_t(s_t) &= N(s_t; \tilde{m}_t, \tilde{v}_t) && (2.14)\\
\tilde{m}_t &= \frac{1}{\tilde{z}_t} \sum_i \chi_i f(y_t; \chi_i) \omega_i && (2.15)\\
\tilde{v}_t &= \frac{1}{\tilde{z}_t} \sum_i \chi_i^2 f(y_t; \chi_i) \omega_i - \tilde{m}_t^2 && (2.16)\\
\tilde{z}_t &= \sum_i f(y_t; \chi_i) \omega_i && (2.17)
\end{aligned}
$$

이렇게 구한 s_t의 사후분포를 이용하여 θ의 근사 사후분포, $q(\theta_t)$를 구할 수 있다. $q_{t|t-1}(s_{t-1})$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 아래와 같다.(Murphy, 2012)

$$
\begin{aligned}
q(\theta_t, i) &\sim N(\theta_{t,i}; \mu_{t,i}, \sigma_{t,i}^2) && (2.18)\\
\mu_{t,i} &= \mu_{t|t-1,i} + a_i \delta_m && (2.19)\\
\sigma_{t,i}^2 &= \sigma_{t|t-1,i}^2 + a_i^2 \delta_v && (2.20)\\
a_i &\triangleq \frac{x_{t,i} \sigma_{t|t-1,i}^2}{\sum_j x_{t,j}^2 \sigma_{t|t-1,i}^2} && (2.21)
\end{aligned}
$$

제 3 장

모의 실험

제 4 장

사례 연구

4.1 자료 설명

'온라인 광고'는 '개시자'(광고 대행)가 웹사이트에 이미지나 텍스트 혹은 복합된 형태의 광고물을 개시하고 '광고주'가 이에 대한 댓가를 지불하는 형태로 이루어진다. 광고에 대한 비용 책정의 방법은 크게 i) 광고 노출 횟수에 따른 과금(cost-per-impression, CPM), ii) 광고 클릭으로 광고주의 웹사이트에 방문한 횟수에 따른 과금(cost-per-click, CPC), iii) 광고 클릭 후 광고주의 웹사이트에서 구매 등의 특정 행위를 한 횟수에 따른 과금(cost-per-conversion, CPM) 으로 나뉜다. 광고주는 세가지 방법 중 고객이 실제 매출에 영향을 줄 수 있는 경우를 직접적으로 반영하는 CPC 혹은 CPM를 선호한다. 따라서 광고에 앞서 고객의 광고 클릭 혹은 이후 행위에 영향을 주는 많은 요인들에 따른 광고 클릭률을 예측하는 것이 중요한 문제일 수 밖에 없다.(Chapelle et al., 2013)

실제 온라인 광고의 클릭률 예측에 사용되는 데이터는 그 건수나 변수의 갯수가 상당히 크기 때문에 일괄처리(batch) 방식으로 처리하기 어렵기 때문

에 '온라인 학습'이 필요한 경우라고 할 수 있다. 사례 분석을 위해 Criteo[1]에서 'Kaggle 대회'[2]를 위해 공개한 4천 5백만건 상당의 온라인 광고 데이터[3]를 사용하였다.

4.2 광고 클릭률 온라인 예측

[1] www.criteo.com, 2005년 설립된 온라인 광고 회사

[2] www.kaggle.com 에서 진행하는 데이터 예측 분석 경연 대회

[3] http://labs.criteo.com/downloads/2014-kaggle-display-advertising-challenge-dataset/

제 5 장

결론

참고 문헌

Agresti, A. (1996). *An introduction to categorical data analysis*. Wiley, 2nd edition.

Chapelle, O., Manavoglu, E., and Rosales, R. (2013). Simple and scalable response prediction for display advertising. Technical Report 212.

Lauritzen, S. L. U. o. A. (1992). Propagation of Probabilities, Means and Variances in Mixed Graphical Association Models.

Minka, T. P. (2013). Expectation Propagation for approximate Bayesian inference. *Statistics*, 17(2):362–369.

Murphy, K. P. (2012). *Machine learning*. The MIT Press.

Opper, M. (1996). On-line versus Off-line Learning from Random Examples: General Results. *Physical Review Letters*, 77:4671–4674.

Opper, M. (1999). A Bayesian approach to on-line learning. *On-line learning in neural networks*, pages 363—-378.

Zoeter, O. (2007). Bayesian Generalized Linear Models in a Terabyte World.