

스플라인을 이용한 스코어 카드

최민성¹⁾ 구자용²⁾ 최대우³⁾

요약

신용위험 관리에서 필수적인 방법론이 스코어 카드이며 이를 작성하는 데에 있어서 널리 쓰이는 방법 중의 하나가 로지스틱 회귀분석이다. 본 논문에서는 로지스틱 회귀 방법에 기초한 스플라인 방법론을 소개하고자 한다. 최종 스코어 카드는 연속형 변수를 범주형 변수화 하므로 조각 선형 스플라인을 채택하였다. 모의 실험을 통하여 제안된 방법의 성능을 규명하였다.

주요용어: 기저함수 방법론, 신용위험, 판별분석, 함수추정

1. 서론

과거의 신용상태가 가까운 미래까지 지속된다는 인간 행동 속성상의 가정하에 과거의 신용과 관련된 정보들을 수집하여 통계적 예측 모형을 구축한다면 이를 이용하여 미래의 신용상태를 예측 할 수 있게 된다(이명식 1999). 본 논문에서는 로지스틱 회귀방법에 기반을 둔 스플라인 방법론을 소개하고자 한다. 이 방법에서는 각 고객의 고유 특성을 나타내는 설명 변수들 중 연속형 변수의 변동성을 예측하고 적합하기 위하여 조각 선형 스플라인 방법을 채택하였으며 이산형 변수의 그룹화를 통하여 각 범주의 수준을 분리 통합하였다. 특히 연속형 변수의 범주형 변수화를 통하여 불량 고객에 대한 거절 이유 설명이 쉽고 분석 보고서를 이용한 배점 조정이 가능하다.

2. 상수 스플라인 모형

본 절에서는 연속형 입력변수는 매듭점이 고정된 스플라인을 사용하고 이산형 입력변수는 가변수를 사용하는 판별 스플라인 기계(Classification Spline Machine, CSM)에 대해 설명하고자 한다.

CSM은 명목형 출력변수 $Y \in \{0, 1\}$ 와 이를 설명하는 데 쓰이는 입력변수 벡터 X 간의 관계를 스플라인 기저(spline)를 이용하여 모형화한다. 기저를 이용한 모형화 기법은 구자용(2000)에 보다 자세히 설명되어 있다.

1) (402-751) 인천시 남구 용현동 253 번지 인하대학교 통계학과, 석사수료

E-mail: mschoi@stat.inha.ac.kr

2) (402-751) 인천시 남구 용현동 253 번지 인하대학교 통계학과, 교수

E-mail: jykoo@stat.inha.ac.kr

3) (449-791) 경기도 용인시 모현면 왕산리 산 89 한국외국어대학교 자연과학대학 정보통계학과, 부교수

E-mail: dachoi@stat.hufs.ac.kr

신용위험관리에서 스코어 카드는 해석이 용이해야 하므로 GAMs(Hastie & Tibshirani, 1990)를 이용하고자 한다. 스플라인 기저함수 B_1, \dots, B_J 가 주어졌을 때

$$f(\mathbf{x}; \boldsymbol{\beta}) = \sum_{j=1}^J \beta_j B_j(\mathbf{x}) \quad (2.1)$$

라 하면 조건부 확률은

$$P(Y = 1 | \mathbf{X} = \mathbf{x}; \boldsymbol{\beta}) = \frac{\exp(f(\mathbf{x}; \boldsymbol{\beta}))}{1 + \exp(f(\mathbf{x}; \boldsymbol{\beta}))} \quad (2.2)$$

로 모형화 하고자 한다. CSM을 이용한 스코어 작성시 사용되는 기저함수는

$$1, (z_j - t)_+^0, 1_A(c_k) \quad (2.3)$$

로 구성되는데, z_j 는 연속형 입력 변수 중의 하나이며 c_k 는 이산형 입력변수 중의 하나이며, t 는 매듭점을 나타내며, A 는 c_k 의 범주들로 구성된 수준집합의 부분집합이다. 즉, CSM에서는 사용되는 기저 상수 기저 1, 조각 상수 스플라인 기저 $(z_j - t)_+^0$, 이산형 기저 $1_A(c_k)$ 가 있다. 이들 기저들은 기저 추가와 기저 삭제로 구성되는 단계적 기저 선택 알고리즘에 의해 결정된다. 본 논문에서 사용한 적합법은 Kooperberg, Bose & stone(1997), 구자용, 박현진 및 최대우(2000)등에서도 사용되었다.

3. 모의 실험

이 절에서는 독일 신용 데이터에 CSM의 적용 결과를 설명하고자 한다. 독일 신용데이터는 독일의 특정 신용 카드 회사의 자료로써 기존 고객의 신용도에 관련된 변수로 이루어져 있다. 각 고객의 연체 여부를 나타내는 반응 변수는

$$\begin{cases} Y = 0 : \text{우량고객} \\ Y = 1 : \text{불량고객} \end{cases} \quad (3.1)$$

로 정의한다.

독일 신용 데이터에서 CSM에 의해 선택된 기저를 정리하면 다음과 같다.

변수	기저의 종류	계수값	t 및 A
상수	상수	-2.72250	-
계좌 잔액	이산형	1.67286	A = {A11, A12}
연체 이력	이산형	1.56791	A = {A30, A31}
대출 기간	상수 스플라인	1.15387	t = 12
나이	상수 스플라인	-0.83100	t = 26
대출 목적	이산형	0.70332	A = {A40, A410, A45, A46}

위의 표에서는 상수 기저를 제외하고 나머지 선택된 기저 중에 중요 순서대로 나열된 것이며 이는 CSM에서 자동적으로 적합되어지며 선택된 기저에 대한 그림을 자동적으로 출력한다. 그림 3.1은 선택된 두개의 연속형 기저의 부분 적합 그림이고 그림 3.2는 이산형 기저를 나타낸다.

그림 3.1: CSM으로 적합된 연속형 기저

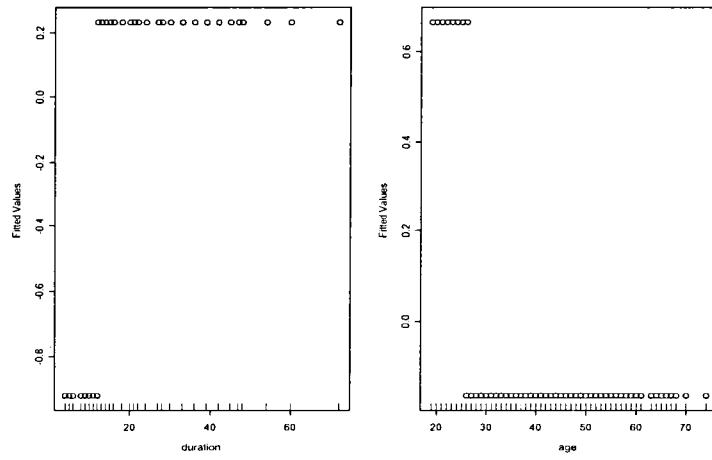
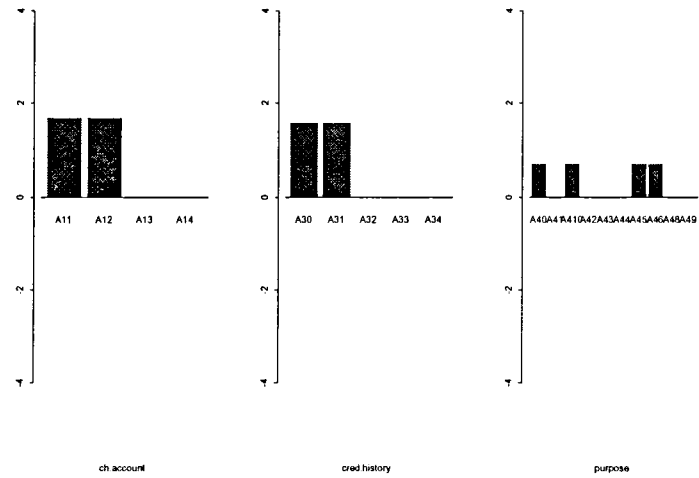


그림 3.2: CSM으로 적합된 이산형 기저



참고 문헌

- [1] 구자용, 박헌진 및 최대우(2000). 데이터 마이닝에서의 폴리클래스. 응용통계연구 제13권 2호 489-504.
- [2] 구자용(2000). 기저함수 방법론. 학술대회 초청강연.
- [3] 이명식(1999). 금융마케팅에서 고객평점제도의 효과성. 한국마케팅저널 제1권 제2호 56-76.
- [4] Kooperberg, C., Bose, S. & Stone, C. J.(1997). Polychotomous Regression. *J. Amer. Statist. Assoc.* **92** 117-127.
- [5] Hastie T. J. & Tibshirani, R. J.(1990). *Generalized Additive Models*. Chapman and Hall, London.