Communications of the Korean Statistical Society Vol. 15, No. 2, 2008, pp. 283–292

베이즈 리스크를 이용한 커널형 분류에서 평활모수의 선택†

김래상[1], 강기훈[2]

요 약

커널밀도함수의 추정을 이용한 분류 문제에서 평활모수(smoothing parameter, bandwidth)의 선택은 핵심적으로 중요한 역할을 한다. 본 논문에서는 분류에서 베이즈 리스크를 최적화하기 위한 평활모수의 선택이 각 개별 확률밀도함수를 추정하기 위한 최적의 평활모수와 어떤 관계가 있는지 살펴보았다. 실제 상황에서 사용할 수 있는 평활모수의 선택 방법으로 붓스트랩(bootstrap)과 교차확인법(cross-validation)을 이용하는 것을 비교한 결과, 붓스트랩 방법은 Hall과 Kang (2005)에서 밝혀진 이론적인 성질에 부합하는 반면 교차확인법은 그렇지 못함을 확인하였다. 또한, 각 방법으로 정한 평활모수를 사용하여 오분류율을 조사해 본 결과에서도 붓스트랩 방법이 우월함을 알 수 있었다.

주요용어: 교차확인법; 붓스트랩; 비모수적 분류.

1. 서론

비모수적 확률밀도함수 추정의 대표적인 커널밀도함수 추정량(kernel density estimator)은 모집단의 확률밀도함수 f로 부터 얻은 크기 n인 랜덤표본을 $X_1, X_2, \ldots, X_n$이라 할 때, 다음과 같이 정의된다.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

여기서, $K_h(\cdot) = K(\cdot/h)/h$이며, K는 커널함수로 일반적으로 비음(nonnegative) 조건과 $\int K(x)dx = 1$을 만족한다. 평활모수 h는 추정된 함수의 부드러운 정도를 결정하는 것으로 적절한 조건 하에서 표본크기에 대한 최적의 차수는 $n^{-1/5}$임이 잘 알려져 있다. 자세한 내용은 Wand와 Jones (1995)의 2.5절을 참고하기 바란다.

커널밀도함수의 추정에서 평활모수를 선택하기 위한 방법에 관해서는 이미 많은 연구가 이루어졌다. 대표적인 것들로는 Park과 Marron (1990), Sheather와 Jones (1991), Faraway와 Jhun (1990), Kim 등 (1994) 등이 있고, 여러가지 방법들에 대해 모의실험을

† 본 연구는 과학기술부/한국과학재단 우수연구센터육성사업의 지원으로 수행되었음 (R11-2000-073-00000).

1) (130-791) 서울시 동대문구 이문동 한국외국어대학교 대학원 통계학과, 석사.
E-mail: rskim78@hotmail.com
2) (449-791) 경기도 용인시 모현면 한국외국어대학교 정보통계학과, 부교수.
교신저자: khkang@hufs.ac.kr

통한 비교는 Park과 Turlach (1992), Jones 등 (1992) 등이 있다. 평활모수의 선택에 관한 다양한 연구 결과는 Wand와 Jones (1995)의 3장에 정리되어 있다.

본 연구는 두 모집단에 대한 분류 문제에서 각 모집단의 커널밀도함수 추정량을 이용하여 우도 확률을 계산하고, 이를 베이즈 리스크의 추정에 끼워 넣고 최적화를 시도하는 방법을 다루었다. 이를 위해 두 모집단의 분포함수를 F, G라 하고 각각에 대응되는 확률밀도함수를 f, g라 하자. 그리고, 관심이 있는 구간 $\mathcal{I}$에 있는 분류되지 않은 새로운 자료 x가 F로부터 나온 자료라 할 수 있는 사전 확률을 $0 < p < 1$라고 하자. 그러면 분류를 위한 기준으로 $\Delta(x) \equiv p\,f(x) - (1-p)\,g(x)$의 부호에 따른 방법을 생각할 수 있다. 즉, $\Delta(x)$의 부호가 양(음)이면 $F\,(G)$에서 온 것으로 분류하는 방법이다. 이 방법은 자료를 분류할 때 사용되는 분류기(classifier)를 $\mathcal{A}$라 할 때 Hall과 Kang (2005)에서 언급된 바와 같이 다음과 같은 베이즈 리스크를 최소화하는 것이다.

$$\begin{aligned} err_{\mathcal{A}}(f, g|\mathcal{I}) = p \int_{\mathcal{I}} & P(\mathcal{A}\text{에 의해 } x\text{가 } g\text{로 분류된 경우}) f(x)dx \\ & + (1-p) \int_{\mathcal{I}} P(\mathcal{A}\text{에 의해 } x\text{가 } f\text{로 분류된 경우}) g(x)dx. \end{aligned} \tag{1.1}$$

모집단 F, G로부터 각각 훈련용으로 뽑은 표본집합들을 $\mathcal{X} = \{X_1, \ldots, X_m\}$, $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$라 할 때, 표본에 기초한 새로운 관측값 x의 분류 기준으로는 다음의 통계량을 사용한다.

$$\hat{\Delta}(x) = p\hat{f}(x) - (1-p)\hat{g}(x). \tag{1.2}$$

여기서 사용된 커널밀도함수 추정량 $\hat{f}$과 $\hat{g}$는 다음과 같이 주어진다.

$$\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} K_{h_1}(x - X_i), \quad \hat{g}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_2}(x - Y_i). \tag{1.3}$$

즉, 식 (1.2)를 계산하여 양의 값이면 x가 F분포로부터 나온 자료로 결정하고 음의 값이면 G분포에서 나온 것으로 결정한다. 그리고, $\hat{\Delta}(x)$값이 0이면 F, G 둘 중 어느 분포에서 나온 것으로 분류해도 무방하다.

Hall과 Kang (2005)은 식 (1.1)과 (1.2)에 기초하여 분류 문제에서 이론적으로 최적인 평활모수의 차수를 규명하고 이를 각각의 개별 확률밀도함수의 추정을 위한 최적의 평활모수와 비교하고 있다. 또한, 실질적인 평활모수의 선택을 위해서는 붓스트랩 방법을 이용할 것을 제안하고 있다. 본 논문에서는 함수추정 분야에서 평활모수의 선택기준으로 흔히 쓰이는 교차확인법을 이 문제에 적용하였을 때의 결과를 살펴보고 붓스트랩을 이용한 경우와 비교해 보고자 한다. 이를 위해 Hall과 Kang (2005)에서 사용된 4가지 경우에 분류를 위한 최적의 평활모수를 교차확인법을 사용하여 구하고 그 차수를 확률밀도함수를 추정하기 위한 최적 평활모수의 차수와 비교하였다. 또한, 붓스트랩과 교차확인법을 이용하여 선택된 평활모수를 이용하여 검정표본에 적용한 각각의 오분류율도 비교하였다.

2. 평활모수의 선택

2.1. 붓스트랩(Bootstrap)

Hall과 Kang (2005)은 확률밀도함수의 추정에 기초한 식 (1.1)의 베이즈 리스크를 붓스트랩 방법을 이용하여 추정하였는데, 간단히 소개하기로 하자. 붓스트랩 샘플을 형성하기 위하여 새로운 평활모수 h_3과 h_4를 각각 사용하여 식 (1.3)의 한 버전으로 $\tilde{f}$와 $\tilde{g}$를 구한다. 여기서 h_3과 h_4는 확률밀도함수의 두 번 미분 또는 네 번 미분한 것을 일치성 있게 추정할 수 있도록 차수가 $n^{-\sigma}$, $0 < \sigma < 1/9$인 것을 사용해야 한다. 이에 관한 자세한 이론적인 내용은 Hall과 Kang (2005)의 4장을, 그리고 실질적인 선택의 한 방법은 6장을 참고하기 바란다.

두 확률밀도함수의 추정량 $\tilde{f}$와 $\tilde{g}$가 형성되면 독립적으로 각각 크기 m인 표본 $\mathcal{X}^* = \{X_1^*, \ldots, X_m^*\}$와 크기 n인 표본 $\mathcal{Y}^* = \{Y_1^*, \ldots, Y_n^*\}$을 추출하여 붓스트랩 표본집단을 만든다. 이 표본과 커널 함수를 이용하여 붓스트랩 버전의 커널밀도함수 추정량을 다음과 같이 구하게 된다.

$$\hat{f}^*(x) = \frac{1}{m}\sum_{j=1}^{m} K_{h_1}(x - X_j^*), \quad \hat{g}^*(x) = \frac{1}{n}\sum_{j=1}^{n} K_{h_2}(x - Y_j^*).$$

위의 추정된 밀도함수를 이용하여 식 (1.2)의 붓스트랩 버전을 다음과 같이 구한다.

$$\hat{\Delta}^*(x) = p\hat{f}^*(x) - (1-p)\hat{g}^*(x). \tag{2.1}$$

다음으로 식 (2.1)을 이용하여 붓스트랩 버전의 $\widehat{err}_{\mathcal{A}}(h_1, h_2)$을 다음과 같이 계산한다.

$$\begin{aligned}\widehat{err}_{\mathcal{A}}(h_1, h_2) = p\int P\{\hat{\Delta}^*(x) < 0 | \mathcal{X} \cup \mathcal{Y}\}\tilde{f}(x)dx \\ +(1-p)\int P\{\hat{\Delta}^*(x) > 0 | \mathcal{X} \cup \mathcal{Y}\}\tilde{g}(x)dx.\end{aligned} \tag{2.2}$$

각 분포의 추정치들을 이용하여 $\widehat{err}_{\mathcal{A}}(h_1, h_2)$를 계산하고, 그 값이 최소가 되도록 하는 평활모수의 쌍 $(\hat{h}_1, \hat{h}_2)$을 선택하여 평활모수 (h_1, h_2)의 추정치로 사용한다. 식 (2.2)에서 $P\{\hat{\Delta}^*(x) < 0 | \mathcal{X} \cup \mathcal{Y}\}$을 구하기 위해서는 붓스트랩 표본을 반복추출해서 계산해야 한다.

2.2. 교차확인법(Cross-validation)

분류에서 교차확인법을 이용한 평활모수의 선택은 분류기를 계산할 때 필요한 각 커널밀도함수의 추정량에 교차확인법의 아이디어를 이용하는 것이다. 이를 위해 $\hat{f}_{-i}$과 $\hat{g}_{-i}$를 각각의 표본집합에서 i번째 표본을 빼고 남은 표본집합 $\mathcal{X}_i = \mathcal{X} \setminus \{X_i\}$와 $\mathcal{Y}_i = \mathcal{Y} \setminus \{Y_i\}$를 이용한 커널밀도함수 추정량을 나타낸다고 하자. 이는 각각의 평활모수 h_1, h_2를 이용하여 다음과 같이 정의된다.

$$\hat{f}_{-i}(X_i) = \frac{1}{m-1}\sum_{j=1, j\neq i}^{m} K_{h_1}(X_i - X_j), \quad \hat{g}_{-i}(Y_i) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} K_{h_2}(Y_i - Y_j).$$

교차확인법에 의해 구한 커널밀도함수 추정량을 이용하여 분류기의 교차확인법 버전 $\hat{\Delta}_{f,-i}(X_i)$, $\hat{\Delta}_{g,-i}(Y_i)$을 계산하는데 그 정의는 다음과 같다.

$$\hat{\Delta}_{f,-i}(X_i) = p\hat{f}_{-i}(X_i) - (1-p)\hat{g}(X_i), \quad \hat{\Delta}_{g,-i}(Y_i) = p\hat{f}(Y_i) - (1-p)\hat{g}_{-i}(Y_i).$$

이를 이용하여 베이즈 리스크의 교차확인법 버전을 다음과 같이 구할 수 있다.

$$\widetilde{err}_{\mathcal{A}}(h_1,h_2) = \frac{p}{m}\sum_{i=1}^{m} I\{\hat{\Delta}_{f,-i}(X_i) < 0, X_i \in \mathcal{I}\} + \frac{(1-p)}{n}\sum_{i=1}^{n} I\{\hat{\Delta}_{g,-i}(Y_i) > 0, Y_i \in \mathcal{I}\}. \tag{2.3}$$

교차확인법을 이용한 평활모수의 추정은 식 (2.3)의 $\widetilde{err}_{\mathcal{A}}(h_1,h_2)$를 계산하여, 그 값을 최소화하는 $(\hat{h}_1,\hat{h}_2)$ 쌍을 선택한다. 이러한 교차확인법은 직관적으로 쉽고 간단하게 적용될 수 있다는 장점 때문에 비모수적 함수추정 분야에서 자주 사용된다. 반면에 분류를 위한 밀도함수의 추정에서는 이론적으로 한 점에서 평균제곱오차를 정확하게 측정하지 못한다는 점 때문에 $err_{\mathcal{A}}(f,g)$를 최소화하는 최적의 (h_1,h_2)의 값을 제대로 추정하지 못한다는 단점이 있다. 이에 대한 이론적인 측면의 고찰은 Hall과 Kang (2005)의 7장을 참고하기 바란다.

3. 모의실험

3.1. 평활모수의 차수 비교

Hall과 Kang (2005)에 의하면 두 확률밀도함수 f와 g가 한 점 x에서 만나는 경우에 분류를 위한 베이즈 리스크를 최소로 하는 평활모수의 최적 차수(order)는 x에서 두 함수의 형태 관계에 의해 결정된다. 여기서 형태라 함은 요철(curvature)을 의미하는데, $f''(x)g''(x) < 0$인 경우에는 최적 평활모수의 차수가 $n^{-1/5}$인 데 비해 $f''(x)g''(x) > 0$인 경우에는 $n^{-1/9}$이 된다. 즉, 만나는 지점의 요철이 같은 경우에는 각 개별 확률밀도함수를 추정하기 위한 최적의 평활모수의 차수와 같고, 요철이 반대인 경우에는 보다 넓은 범위의 평활모수를 사용해야 한다는 것이다. 놀라운 것은 교차하는 한 점에서의 성질이 분류를 위한 전체 평활모수의 차수를 결정한다는 것이다.

이러한 사실이 실제 자료 분석에 얼마나 부합하는지 비교를 위해 모의실험을 시행하였다. 이를 위해 Hall과 Kang (2005)에서 사용된 분포들을 이용하였고, 비교를 위한 표본 또한 같은 것을 사용하였다. 즉, f를 표준정규분포로 하고 g를 4가지 다른 분포로 하여 표본의 크기 m과 n을 20부터 200까지 로그 등간격으로 10등분하여 변화시켜 가면서 비교하도록 하였다. 각 경우에 해당되는 분포는 다음에 주어져 있다.

경우 1 : 예 1 : $f \sim N(0,1)$, $g \sim N(-1.2, 0.6^2)$,

예 2 : $f \sim N(0,1)$, $g \sim \frac{1}{5}N(0,1) + \frac{1}{5}N\left(1, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}N\left(\frac{19}{12}, \left(\frac{5}{9}\right)^2\right)$,

경우 2: 예 1: $f \sim N(0,1),\ g \sim N(1,1)$,
예 2: $f \sim N(0,1),\ g \sim C(0,1)$.

여기서 경우 1의 예 1은 두 밀도함수의 교점이 $x = -0.515$이고 그 교점에서 $f''(x) = -0.255$, $g''(x) = 0.281$이다. 또, 예 2는 교점이 $x = 0.707$이고 그 교점에서 $f''(x) = -0.156$, $g''(x) = 0.327$이다. 즉, 경우 1은 $f''(x)g''(x) < 0$이므로 분류를 위한 평활모수의 이론적 최적 차수가 $n^{-1/5}$이 되어야 한다. 그리고 경우 2의 예 1에서는 교점이 $x = 0.5$이고 그 교점에서 $f''(x) = g''(x) = -0.264$이다. 경우 2의 예 2에서 $C(0,1)$은 표준코쉬분포이며, 두 분포의 교점은 $x = \pm 1.851$이고 그 교점에서 $f''(x) = 0.175$, $g''(x) = 0.068$이다. 이 경우에는 두 점에서 만나지만 완벽히 대칭이므로 이론적인 성질은 한 점에서 만나는 경우와 동일하게 된다. 즉, 경우 2의 두 예는 $f''(x)g''(x) > 0$이므로 이론적인 최적 평활모수의 차수가 $n^{-1/9}$이다.

식 (2.2)와 (2.3)을 각각 최소로 하는 평활모수를 추정하기 위하여 세밀하게 구간을 나누어 찾았다. 우선 Wand와 Jones (1995)의 3.2절에서 소개된 바와 같은 정규분포를 이용한 평활량 $\hat{h}_{NS1}$과 $\hat{h}_{NS2}$를 다음과 같이 계산한다.

$$\hat{h}_{NS1} = \left[\frac{8\pi^2 R(K)}{3\mu_2(K)^2 m}\right]^{\frac{1}{5}} \hat{\sigma}_X, \quad \hat{h}_{NS2} = \left[\frac{8\pi^2 R(K)}{3\mu_2(K)^2 n}\right]^{\frac{1}{5}} \hat{\sigma}_Y$$

여기서, $R(K) = \int K^2(x)dx$, $\mu_2(K) = \int x^2 K(x)dx$이고 $\hat{\sigma}$는 표본으로부터 구한 표준편차의 추정량이다. 이렇게 계산된 평활모수를 이용하여 $(\hat{h}_{NSi}/3, 3\hat{h}_{NSi})$, $i = 1, 2$, 구간을 만들고, 그 구간을 각각 51등분 하여 격자점을 생성하고 식 (2.2)와 (2.3)을 최소로 하는 최적의 평활모수를 찾아 분류를 위한 최적의 평활모수 추정치 $(\hat{h}_1, \hat{h}_2)$로 삼는다. 이러한 작업을 100번 반복하였고, 식 (2.2)와 (2.3)에서 p값은 사전정보가 없으므로 1/2을 사용하였다.

모의실험의 결과는 그림 3.1과 3.2에 나타내어져 있다. 이 그림들에서 x축은 표본크기의 로그값$(\log n)$을 나타내며 y축은 평활모수의 로그값을 부호를 바꾼 것$(-\log \hat{h}_j,\ j = 1, 2)$에 해당된다. 검은색 각 점들은 100번 반복에 따른 $(\log n, -\log \hat{h}_j)$의 평균치를 나타내고, 실선은 각 점을 적합시킨 회귀직선이며, 적합된 회귀직선의 중앙을 지나면서 기울기가 각각 1/5과 1/9인 직선이 각각 점선과 쇄선으로 나타내져 있다. 즉, 적당한 α에 대해 $h \sim n^{-\alpha}$이 성립하는지를 알아보는 방법으로 로그를 취하고 단순회귀를 실시하여 기울기를 살펴보는 것이다.

앞에서 언급한 바와 같이 경우 1에서는 $f''(x)g''(x) > 0$ 즉, f와 g가 교차하는 지점에서 요철이 둘다 오목하거나 둘다 볼록하므로 최적 평활모수의 차수가 $n^{-1/5}$이어야 한다. 그림 3.1에서 보면 상단 두 그림이 예 1에 대한 것이고 하단 두 개가 예 2에 대한 것이다. 각 단에서 왼쪽은 평활모수 h_1, 오른쪽은 h_2에 대한 것이다. 두 예에서 모두 적합된 회귀직선이 기울기가 1/5인 점선에 가까워야 하는데 오히려 쇄선에 가까운 형태로 나타나고 있어 이론적인 성질에 부합되지 않음을 알 수 있다. 반면, 같은 표본을 사용한 Hall과 Kang (2005)의 Figure 2에서 보면 붓스트랩 기법은 평활모수의 이론적인 최적 차수를 충실히 반영하고 있다. 그림 3.2에 제시된 경우 2에서는 f와 g가 교차하는 지

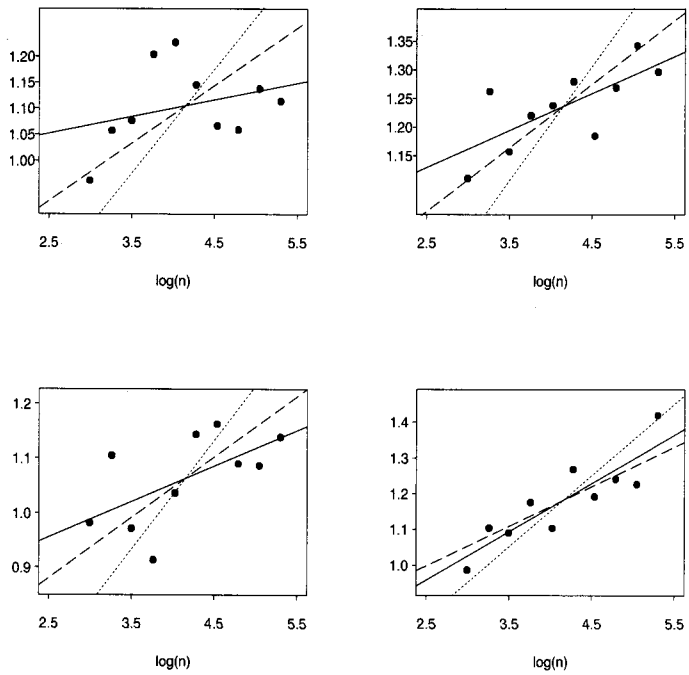

그림 3.1: 경우 1에서 표본 크기에 따른 평활모수의 차수 비교 (상단은 예1 하단은 예 2. 왼쪽은 평활모수 h_1, 오른쪽은 평활모수 h_2에 해당됨. 좌표점은 $(\log n, -\log \hat{h}_j)$의 평균값, 실선은 좌표점을 적합시킨 회귀직선, 점선은 적합된 회귀직선의 중앙을 통과하는 기울기 1/5인 직선, 쇄선은 기울기 1/9인 직선)

점에서 요철이 서로 모양이 다르므로 즉, $f''(x)g''(x) < 0$이므로 이론적으로 최적 평활모수의 차수가 $n^{-1/9}$이어야 한다. 그런데 결과를 보면 회귀직선이 기울기가 1/9인 직선에 가깝게 나타나지 않고 있다. 즉, 교차확인법에 의한 방법은 최적 평활모수의 차수를 제대로 찾아내지 못하고 있음을 확인할 수 있다. 경우 2에 대한 붓스트랩 방법의 결과는 Hall과 Kang (2005)의 Figure 3에 있고, 역시 평활모수의 최적 차수에 대한 이론적인 성질을 반영하고 있음을 알 수 있다.

3.2. 오분류율의 비교

2장에서 소개한 평활모수의 선택 방법이 실제 오분류율에는 어떤 영향을 미치는지 모의실험을 시행하였다. 즉, 훈련표본에서 분류를 위한 평활모수의 추정치를 이용하여 각 커널밀도함수를 추정하고, 새로운 검정표본을 생성하여 오분류된 개수를 세어 비율을 비교하였다. 붓스트랩 방법으로 평활모수를 추정하기 위한 식 (2.2)에서 붓스트랩 표

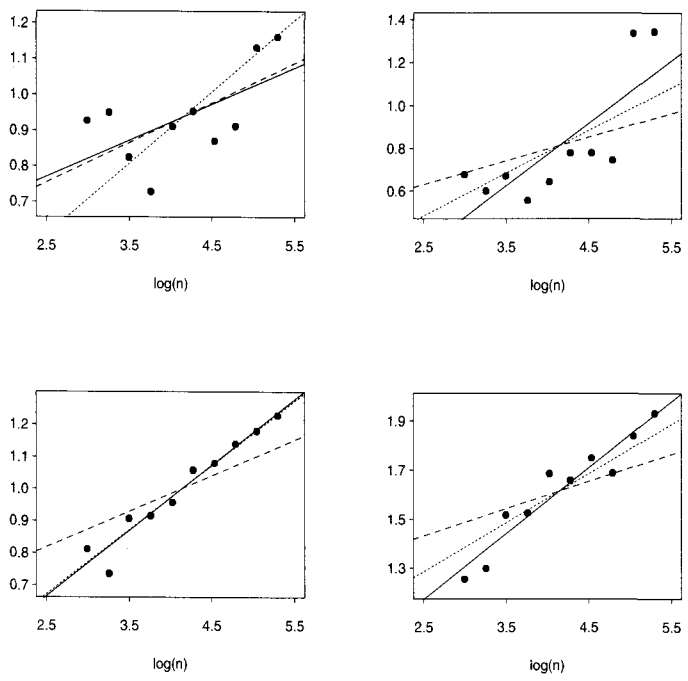

그림 3.2: 경우 2에서 표본 크기에 따른 평활모수의 차수 비교 (상단은 예 1 하단은 예 2. 왼쪽은 평활모수 h_1, 오른쪽은 평활모수 h_2에 해당됨.)

본은 100회 반복하여 계산하였다. 오분류 개수를 구하는 방법은 다음과 같은 분류기 $\hat{\Delta}$의 부호를 이용한다. 즉, 검정표본을 $\mathcal{T}\mathcal{X} = \{X_1^t, \ldots, X_m^t\}$와 $\mathcal{T}\mathcal{Y} = \{Y_1^t, \ldots, Y_n^t\}$라 하고,

$$\begin{aligned}\hat{\Delta}_f(X^t) &= p\hat{f}(X^t) - (1-p)\hat{g}(X^t) \\ &= \frac{p}{m}\sum_{i=1}^{m} K_{h_1}(X^t - X_i) - \frac{(1-p)}{n}\sum_{j=1}^{n} K_{h_2}(X^t - Y_j), \qquad (3.1)\\ \hat{\Delta}_g(Y^t) &= p\hat{f}(Y^t) - (1-p)\hat{g}(Y^t) \\ &= \frac{p}{m}\sum_{i=1}^{m} K_{h_1}(Y^t - X_i) - \frac{(1-p)}{n}\sum_{j=1}^{n} K_{h_2}(Y^t - Y_j), \qquad (3.2)\end{aligned}$$

을 계산한다. 오분류된 경우는 식 (3.1)이 0보다 작을 때와 식 (3.2)가 0보다 클 때이다. 이 두 가지 경우의 개수를 모두 더한 값을 전체 검정표본의 크기 $(m+n)$로 나누어 오분류율을 구할 수 있다.

2장에서 설명한 방법으로 붓스트랩과 교차확인법을 이용하여 최적 평활모수를 계산하고 검정표본을 생성하여 그 표본집합을 가지고 각각 오분류율을 계산한 결과를 표

표 3.1: 오분류율의 비교

분류상황	방법 \ 표본크기	25	50	100
경우 1: 예 1	Boot	0.220	0.218	0.217
	CV	0.255	0.229	0.222
경우 1: 예 2	Boot	0.234	0.233	0.231
	CV	0.324	0.318	0.292
경우 2: 예 1	Boot	0.325	0.316	0.313
	CV	0.368	0.334	0.329
경우 2: 예 2	Boot	0.365	0.361	0.358
	CV	0.400	0.391	0.387

3.1에 제시하였다. 표본의 크기는 25, 50 그리고 100인 세 경우를 고려하였고 오분류율은 100번 반복한 평균을 계산하였다. 표 3.1을 보면 모든 경우에서 붓스트랩을 이용한 결과가 교차확인법 보다 오분류율이 작게 나타나고 있음을 알 수 있다.

결론

비모수적 분류에서 베이즈 리스크의 최적화를 위한 커널밀도함수의 추정에서 평활모수의 선택에 관해 살펴보았다. 붓스트랩과 교차확인법을 비교한 결과 붓스트랩 방법이 평활모수의 이론적인 차수를 잘 반영함에 비해 평활모수의 선택 기준으로 흔히 사용되는 교차확인법은 전혀 부합하지 못하고 있음을 확인하였다. 또한, 각 방법을 이용한 오분류율의 비교에서도 붓스트랩 방법이 더 우수함을 알 수 있었기에 베이즈 리스크를 최적화하기 위한 분류를 위해서는 교차확인법을 이용한 평활모수의 선택은 피해야 할 것이다. 본 연구에서는 한 점에서 만나는 두 확률밀도함수의 경우만을 살펴보았지만 여러 분포에 대한 분류이거나 여러 점에서 만나는 두 분포의 경우에도 Hall과 Kang (2005)에서 규명된 이론적인 성질을 만족하려면 실질적으로 교차확인법보다 붓스트랩 방법을 이용해야 함을 예상할 수 있다.

참고문헌

Faraway, J. J. and Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association*, **85**, 1119–1122.

Hall, P. and Kang, K. H. (2005). Bandwidth choice for nonparametric classification. *The Annals of Statistics*, **33**, 284–306.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401–407.

Kim, W. C., Park, B. U. and Marron, J. S. (1994). Asymptotically best bandwidth selectors in kernel density estimation. *Statistics & Probability Letters*, **19**, 119–127.

Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66–72.

Park, B. U. and Turlach, B. A (1992). Practical performance of several data driven bandwidth selectors (with discussion). *Computational Statistics*, **7**, 251–270.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, Ser. B, **53**, 683–690.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall/CRC, London.

[2008년 1월 접수, 2008년 2월 채택]

On Practical Choice of Smoothing Parameter in Nonparametric Classification†

Rae-Sang Kim[1], Kee-Hoon Kang[2]

Abstract

Smoothing parameter or bandwidth plays a key role in nonparametric classification based on kernel density estimation. We consider choosing smoothing parameter in nonparametric classification, which optimize the Bayes risk. Hall and Kang (2005) clarified the theoretical properties of smoothing parameter in terms of minimizing Bayes risk and derived the optimal order of it. Bootstrap method was used in their exploring numerical properties. We compare cross-validation and bootstrap method numerically in terms of optimal order of bandwidth. Effects on misclassification rate are also examined. We confirm that bootstrap method is superior to cross-validation in both cases.

Keywords: Bayes risk; bootstrap; cross-validation.

† This research was supported by the SRC/ERC program of MOST/KOSEF (R11-2000-073-00000).

1) Master of Science, Department of Statistics, Hankuk University of Foreign Studies, 270 Imun-dong, Dongdaemoon-Gu, Seoul 130-791, Korea. E-mail: rskim78@hotmail.com

2) Associate Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin 449-791, Korea. Correspondence: khkang@hufs.ac.kr