df

이경선 / 학생 / 통계학과

October 2022

1 Introduction

12.1 판별 분석

g 클래스 세트가 있고 각 경우에 대해 클래스를 알고 있다고 가정합니다(정확하게 가정). 그런 다음 클래스 정보를 사용하여 데이터 구조를 밝힐 수 있습니다. W가 클래스 평균에 중심을 둔 변수의 공분산 행렬인 클래스 내 공분산 행렬을 나타내고 B가 클래스 평균에 의한 예측의 클래스 간 공분산 행렬을 나타낸다고 하자. M을 클래스 평균의 $g \times p$ 행렬로 하고, G를 클래스 지시 변수의 $n \times g$ 행렬로 하자(따라서 i가 클래스 j에 할당된 경우에만 $g_{ij} = 1$). 그런 다음 예측은 GM이다. $\bar{x}$를 전체 샘플에 대한 변수의 평균으로 하자. 그러면 표본 공분산 행렬은

$$W = \frac{(X - GM)^T (X - GM)}{n - g}, =$$

g-1

B는 최대 $\min(p, g-1)$의 순위를 갖는다. Fisher(1936)는 클래스 내 분산에 대한 클래스 평균의 분리의 최대 비율을 갖는 변수의 선형 조합 xa를 찾는 선형 판별 분석을 도입했습니다. 즉, $a^T BBa / a^T WaW$ 이것을 계산하려면, 그룹 내 상관 행렬로 동일성을 가질 수 있도록 변수의 구형화(305페이지 참조) xS를 선택합니다. 재조정된 변수에서 문제는 $\|a\| = 1$에 따라 $a^T Ba$를 최대화하는 것이며, PCA에서 보았듯이, 이것은 a를 가장 큰 고유값에 해당하는 B의 고유 벡터로 가져감으로써 해결된다. 선형 조합 a는 부호 변경까지 고유합니다(여러 고유값이 없는 경우). 프로그램에서 반환되는 a의 정확한 배수는 클래스 내 분산 행렬의 정의에 따라 달라집니다. 우리는 $n-g$의 전통적인 제수를 사용하지만 n과 $n-1$의 제수가

사용되었다.
주성분의 경우, 다음으로 큰 고유값에 해당하는 선형 성분을 추가로 취할 수 있다. 최대 $r = \min(p, g-1)$ 양의 고유값이 있을 것이다. 고유값은 선형 조합으로 설명되는 클래스 간 분산의 비율로, 사용할 개수를 선택하는 데 도움이 될 수 있습니다. 해당 변환 변수를 선형 판별 변수 또는 표준 변수라고 합니다. 데이터를 처음 몇 개의 선형 판별기에 표시하는 것이 종종 유용합니다(그림 12.1). 그룹 내 공분산이 동일성이어야 하므로 동일한 규모의 그림을 선택했습니다. (그림(ir) 사용).아이다)는 이 그림을 색상 없이 제공할 것이다.) 선형 판별은 일반적으로 데이터 세트에서 평균 0을 갖도록 중심화된다.

——————————— 이는 그룹 간 분산의 99.65%가 첫 번째 판별 축에 있음을 보여준다. 사용. 플롯(ir.lda, dimension = 1) 그림(ir.lda, 유형 = "밀도", 치수 = 1) 첫 번째 선형 판별에서 그룹의 분포를 조사합니다.
우리가 설명한 접근 방식은 브라이언(1951년)에 이어 전통적인 접근 방식이지만, 그것만이 아니다. (12.1)에서 B의 정의는 데이터 세트의 크기에 따라 그룹에 가중치를 부여한다. Rao(1948)는 그룹 평균의 가중되지 않은 공분산 행렬을 사용했고, 우리 소프트웨어는 클래스의 이전 확률로 가중된 공분산 행렬을 사용한다.
정규 모집단에 대한 차별 차별에 대한 대안적인 접근법은 확률 모델을 통한 것이다. π_c가 클래스의 이전 확률을 나타내고, $p(x \mid c)$가 각 클래스에 대한 관찰 분포 밀도를 나타낸다고 하자. 그렇다면 x를 관찰한 후 등급의 후방 분포는 다음과 같다.

$$p(c \mid x) =_c p(x \mid c)p(x) \propto \pi_c p(x \mid c)$$

그리고 예상 오류 수를 가장 적게 만드는 할당 규칙이 최대 $p(c \mid x)$로 클래스를 선택한다는 것을 보여주는 것은 매우 간단하다. 이를 베이즈 규칙이라고 한다. (우리는 섹션 12.2에서 더 일반적인 버전을 고려한다.)

이제 클래스 c에 대한 분포가 평균 μ_c 및 공분산 Σ_c로 다변량 정규 분포라고 가정한다. 그러면 베이즈 규칙이 최소화됩니다.

$$filename Q_c = -2\log p(x \mid c) - 2\log \pi_c = (x - {}_c)\,\Sigma_c^{-1}\,(x - \mu_c)^T + \log|\Sigma_c| - 2\log \pi_c$$

$$(12.3) Xm@t_{\mathrm{J}} LXX|xDpX\text{ø}tp, X|xDh_{\text{¸}}\backslash\text{‘}_{\text{¸}}.Pt\Gamma\backslash$$

Q_c 간의 차이는 x의 2차 함수이므로 이 방법을 2차 판별 분석이라고 하며 결정

영역의 경계는 x 공간의 2차 표면이다. 이것은 우리의 함수 qda에 의해 구현된다.

또한 클래스에 공통 공분산 행렬 Σ가 있다고 가정한다.

Q_c의 차이는 x의 선형 함수이며, 우리는 $-Q_c/2$ 또는 를 최대화할 수 있다.

$$L_c = x\Sigma^{-1}\mu_c\Sigma^{-1}\mu_c^T/2 + \log \pi_c$$

(12.3) 또는 (12.4)를 사용하려면 μ_c 및 Σ_c 또는 Σ를 추정해야 한다. 명확한 추정치가 사용되며, 각 클래스 내의 표본 평균과 공분산 행렬이 사용되며, Σ의 경우 W가 사용된다.

이것이 피셔의 선형 판별과 어떤 관련이 있나요? 후자는 클래스 내 샘플 분산을 단위로 하는 선형 판별 변수인 새로운 변수를 제공하며, 그룹 평균 간의 차이는 전적으로 첫 번째 r 변수에 있다. 따라서 이러한 변수에서 마할라노비스 거리 $(widehat\Gamma X)Sigma = W$)는 다음과 같습니다.

$$\|x - \mu_c\|^2$$

그리고 벡터의 첫 번째 r 성분만 c에 의존한다. 마찬가지로, 이 변수들에 대해서

$$L_c = x\mu_c^T - \|\mu_c\|^2/2 + \log \pi_c$$

r차원에서도 작업할 수 있습니다. 클래스가 두 개뿐인 경우, 단일 선형 판별자가 있고,

$$L_2 - L_1 = x\left(\mu_2 - \mu_1\right)^T + const$$

이것은 계수 $(\mu_2-$를 갖는 선형 판별의 아핀 함수이다. $\mu_1)^T$가 단위 길이로 조정되었습니다.

선형 판별 분석은 $g = 2$에 대한 로지스틱 회귀 분석인 $p(c \mid x)$와 $g > 2$에 대한 다항식 로그 선형 모델을 사용합니다. 그러나 7장의 방법과는 달리 사용되는 매개 변수 추정 방법이 다르다. 선형 판별 분석은 모집단이 실제로 그룹 내 공분산 행렬이 동일한 다변량 정규 분포를 따르지만, 그 우수성은 취약하기 때문에 분류에

일반적으로 7장의 방법이 선호된다.

게 데이터 세트

알 수 없는 색상 형태(종)의 미래의 렙토그라푸스 게의 성별을 예측하는 규칙을 구성할 수 있습니까? BD는 남성과 여성에 대해 다르게 측정되므로 분석에서 생략하는 것이 신중해 보였습니다. 우선, 우리는 양식 간의 차이를 무시한다. 매우 비정규적인 모집단에 대한 선형 판별 분석은 본질적으로 $\mathrm{x_{,}}| > .\mathrm{CL}^3\mathrm{RW}^{-2}\mathrm{CW}^{-1}$ 는 차원 중립 양입니다. 파란색 양식에 대해 6개의 오류가 발생했습니다.

특히 그룹 내 분포가 공동 정규성에 가까워 보이므로 색상 형식을 고려하는 것이 타당하다(그림 4.13(96페이지) 및 11.2(306페이지) 참조). 처음 두 선형 판별은 그룹 간 변동을 지배한다. 그림 12.2는 이러한 변수에 대한 데이터를 보여준다.

우리는 모든 결정 표면을 정확히 플롯에 나타낼 수 없다. 그러나 처음 두 개의 선형 판별을 데이터로 사용하면 매우 좋은 근사치를 얻을 수 있다. 그림 12.2를 참조하라.

독자는 이 문제에 대해 2차 변별을 시도하도록 초대받는다. 그룹의 공분산이 매우 유사하게 나타나기 때문에, 그것은 선형 판별보다 매우 약간 더 잘 수행된다.

다변량 위치 및 척도의 강력한 추정

우리는 $W(B$가 아닌)에 대한 더 강력한 추정치를 고려하고 싶을 수 있다. 다소 반직관적으로, 다변량 평균의 각 성분에 강력한 위치 추정기를 적용하는 것만으로는 충분하지 않으며(Rousseew and Leroy, 1987, p. 250), 평균과 분산의 추정을 동시에 고려하는 것이 더 쉽다.

다변량 분산은 특이치에 매우 민감합니다. 강력한 공분산 추정을 위한 두 가지 방법은 우리의 함수 cov.rob [1]와 S-PLUS 함수 cov.mve 및 cov.mcd(Rousseau, 1984; Rousseau and Leroy, 1987)와 covRobin 라이브러리 섹션 robust를 통해 사용할 수 있다. p 변수에 대한 n개의 관측치가 있다고 가정합니다. 최소 부피 타원체 방법은 최소 부피인 $h = \lfloor (n + p + 1)/2 \rfloor$ 점을 포함하는 타원체를 찾고, 최소 공분산 결정자 방법은 공분산이 최소 결정 인자를 갖는 h 점을 찾는다(따라서 이러한 점들의 평균에 대한 기존의 신뢰 타원체는 최소 부피를 갖는다). MCD는 높은 통계 효율 때문에 선호된다. 우리의 함수 cov.rob는 둘 다 구현한다.

MVE 또는 MCD에 대한 검색은 평균 및 분산 행렬(선택에 맞게 조정됨)이 초기 추정치를 제공하는 h 포인트를 제공합니다. 이것은 초기 공분산을 사용하여 초기 평균에서 마할라노비스 거리가 너무 크지 않은 점(특히 정규성에서 97.5% 포인트 내에서)을 선택하고 평균과 분산 행렬을 반환함으로써 정제된다.

대안적 접근 방식은 M-추정의 아이디어를 이 설정으로 확장하여 소수의 ν 자유도에 다변량 t_ν 분포를 맞추는 것이다. 이것은 우리의 함수 cov.trob에서 구현된다; 사용된 알고리즘의 이면에 있는 이론은 Kent, Tyler and Vardi (1994)와 Ripley (1996)에서 주어진다. 보통 cov.trob는 cov보다 빠르다.

강도, 하지만 후자의 극단적인 저항은 부족하다. 우리는 두 개 이상의 클래스에 선형 판별 분석을 사용할 수 있으며, 법의학 유리 데이터 세트 fgl로 이를 설명한다.

우리의 함수 lda는 최소 부피 타원체 추정치(그러나 그룹 중심의 강력한 추정은 없음) 또는 방법 = "t"를 설정하여 다변량 t_ν 분포를 사용하기 위한 인수 방법 = "mve"를 가지고 있다. 이는 그림 12.3에서 알 수 있듯이 fg1 포렌식 유리 데이터에 상당한 차이를 가져온다. 기본 $\nu = 5$를 사용한다.

12.2 분류 이론

패턴 인식 용어에서는 분류와 함께 주어진 예제를 훈련 세트라고 하며, 미래의 사례는 테스트 세트를 형성한다. 성공의 주요 척도는 오류(또는 잘못된 분류) 비율입니다. 훈련 세트를 다시 분류하여 편향된 추정치를 얻을 수 있지만, 전체 모집단에서 무작위로 선택한 테스트 세트의 오류율은 편향되지 않은 추정치가 될 것이다.

발생한 오류의 유형을 아는 것이 도움이 될 수 있습니다. 혼란 행렬은 클래스 j로 분류된 실제 클래스 i를 가진 경우의 수를 제공한다. 일부 문제에서 일부 오류는 다른 오류보다 더 나쁜 것으로 간주되므로 클래스 i의 경우를 클래스 j에 할당하는 데 비용 L_{ij}을 할당한다. 그러면 우리는 오류율보다는 평균 오류 비용에 관심이 있을 것이다.

평균 오류 비용이 $\sum_i L_{ic}(i \mid x)p(i \mid x)$를 최소화하는 클래스 c에 할당하는 Bayes 규칙에 의해 최소화된다는 것을 보여주는 것은 상당히 쉽다. 여기서 $p(i \mid x)$

는 x를 관찰한 후 클래스의 사후 분포이다. 모든 오류의 비용이 동일하면, 이 규칙은 가장 큰 사후 확률 $p(c \mid x)$를 갖는 클래스 c를 선택하는 것과 같다. 최소 평균 비용은 베이즈 위험으로 알려져 있다. 우리는 종종 그것에 대한 하한을 Ripley(1996, 페이지 196-7)의 방법으로 추정할 수 있다(347페이지의 예 참조).

우리는 12.1 섹션에서 정상 모집단에 대해 $p(c \mid x)$를 계산하는 방법과 동일한 오류 비용으로 베이즈 규칙을 추정하는 것이 선형 및 2차 판별 분석으로 이어지는 방법을 보았다. 우리 함수가 예측한 대로. lda and predict. qda 반환 사후 확률, 오류 비용이 있는 분류에도 사용할 수 있다.

사후 확률 $p(c \mid x)$도 직접 추정할 수 있다. 단 두 클래스의 경우 glm에 의해 적합한 로지스틱 회귀 분석을 사용하여 $p(1 \mid x)$를 모델링할 수 있다. 세 개 이상의 클래스에 대해 다중 로지스틱 모델이 필요합니다. 대리 로그 선형 포아송 GLM 모델(섹션 7.3)을 사용하여 이를 적합시킬 수 있지만 라이브러리 섹션 nnet에서 멀티콤 함수를 사용하는 것이 일반적으로 더 빠르고 쉽습니다.
분류 트리는 특별한 다중 로지스틱 모델에 의해 $p(c \mid x)$를 직접 모델링하는데, 여기서 오른쪽은 트리에 의해 케이스가 지정될 잎을 지정하는 단일 요인이다. 다시, 사후 확률은 예측 방법에 의해 주어지기 때문에 불평등한 오류 비용에 대한 베이즈 규칙을 추정하는 것이 쉽다.
예측 및 '플러그인' 규칙 마지막 몇 단락에서 우리는 중요한 점에 대해 이야기했다. 베이즈 규칙을 찾으려면 사후 확률 $p(c \mid x)$를 알아야 한다. 이것들은 알려지지 않았기 때문에 명시적이거나 암묵적인 매개 변수 계열 $p(c \mid x; \theta)$를 사용한다. 지금까지 고려된 방법에서 우리는 $p(c \mid x; \hat{\theta})$가 실제 사후 확률인 것처럼 행동한다. 여기서 $\hat{\theta}$는 훈련 세트 øø˘$\mathcal{X}$⊔.$\mathcal{T}$는 종종 적절한 가능성을 최대화함으로써 이루어진다. 이를 '플러그인' 규칙이라고 합니다. 그러나 $p(c \mid x)$의 '올바른' 추정치는 예측 추정치를 사용하기 위해 (Ripley, 1996, §2.4)이다.

$$\tilde{p}(c \mid x) = P(c = c \mid X = x,$$

(12.5)의 통합을 분석적으로 수행하는 것은 종종 가능하지 않지만 θ에 대한 적절한 '모호한' 선행으로 선형 및 2차 판별이 가능하다(Aitchison and Dunsmore, 1975; Geisser, 1993; Ripley, 1996). 이 추정치는 함수 1da 및 qda에 대한 예측 방법의 방법 = "구체적"으로 구현된다. 분산 행렬의 좋은 추정치를 위한 충분한

데이터가 있는 경우 특히 선형 판별의 경우 차이가 작은 경우가 많습니다. 없을 때, 모란과 머피(Murphy)는 인수 방법 = "disciased"에 의해 구현된 $\log p(x \mid c)$의 편향되지 않은 추정기를 사용하여 상당한 개선을 얻을 수 있다고 주장한다.

간단한 예: 쿠싱 증후군
우리는 Aitchison과 Dunsmore(1975, 표 11.1-3)에서 가져와 Ripley(1996)가 같은 목적으로 사용한 작은 예를 통해 이러한 방법을 설명한다. 이 자료는 부신에 의한 코르티솔의 과다 분비와 관련된 과민성 질환인 쿠싱 증후군 환자에 대한 진단 테스트에 관한 것이다. 이 데이터 세트는 a, b, c로 표현되는 세 가지 유형의 증후군을 가지고 있다. (이들은 '부종', '양쪽 과형성', '암종'을 인코딩하고 과잉 분비의 근본적인 원인을 나타낸다.) 이것은 조직병리학적으로만 결정될 수 있다.) 관찰은 스테로이드 대사물 테트라하이드로코르티존 및 임신 트리올의 비뇨기 배설률(mg/24 h)이며 로그 척도로 간주된다.

알 수 없는 유형(u 표시)의 환자가 6명인데, 이 중 한 명은 나중에 네 번째 유형으로 밝혀졌고, 다른 한 명은 결함이 있는 것으로 측정되었다.

그림 12.4는 1da에 의해 생성된 분류와 2차 판별 분석의 다양한 옵션을 보여준다. 이것은 m에 의해 생산되었다.

(함수 사전 그림은 스크립트에 나와 있습니다.) 우리는 이것들을 m에 의해 수행되는 로지스틱 식별과 대조할 수 있다.

(기능 쿠션도는 스크립트에 나와 있습니다.) 여기서와 같이 클래스가 분산 행렬이 상당히 다를 경우 선형 및 로지스틱 판별은 상당히 다른 답을 제공할 수 있습니다(그림 12.4 및 12.5 비교).

분류 트리의 경우 m을 사용할 수 있습니다.

이렇게 작은 데이터 세트를 사용하여 우리는 그림 12.5에 표시된 트리의 크기를 조정하려고 시도하지 않는다.
혼합물 판별 분석 (플러그인) 이론의 또 다른 응용은 혼합 판별 분석(Hastie and Tibshirani, 1996)으로 라이브러리 섹션 mda에서 구현된다. 이 값은 다변량 정규

혼합물 분포를 각 클래스에 적합시킨 다음 적용됩니다(12.2).

12.3 비모수 규칙

클래스 밀도 또는 로그 후면의 비모수 추정치에 기반한 많은 비모수 분류자가 있다. 라이브러리 섹션 클래스는 k-가장 가까운 이웃 분류기와 관련 방법(Devijver and Kittler, 1982; Ripley, 1996) 및 학습 벡터 양자화(Kohonen, 1990, 1995; Ripley, 1996)를 구현한다. 이것들은 모두 일부 참조 세트에서 가장 가까운 k 예를 찾고, 이러한 k 예제 클래스 중 다수표를 얻거나, 마찬가지로 k 예제 중 클래스의 비율로 사후 확률 $p(c \mid x)$를 추정하는 것에 기초한다.

방법은 기준 세트의 선택에 따라 다르다. k-가장 가까운 이웃 방법은 전체 훈련 세트 또는 편집된 하위 집합을 사용한다. 학습 벡터 양자화는 훈련 세트를 요약하기 위해 훈련 세트 예제가 아닌 공간에서 점을 선택하는 데 있어 K-평균과 유사하지만, K-평균과 달리 예제의 클래스를 고려한다.

이 방법들은 거의 항상 유클리드 거리로 '가장 가까운 것'을 측정한다. 쿠싱 증후군 데이터의 경우 기록된 공변량에 대해 유클리드 거리를 사용하는 대신 임의로 균등하게 스케일링합니다.

이 데이터 집합은 너무 작아서 라이브러리 섹션 클래스에서 편집 및 LVQ 메서드를 시도할 수 없습니다.

12.4 신경망

신경망은 8.10절에서 본 것처럼 다중 로지스틱 회귀 분석의 유연한 비선형 확장을 제공한다. 우리는 다음과 같은 코드로 쿠싱 증후군의 예를 고려할 수 있다. [2]
결과는 그림 12.7과 같다. 우리는 다른 실행이 다른 분류자를 제공하기 때문에 모든 경우에 가능성에 대한 다중 로컬 최대값이 있다는 것을 알 수 있다.
일단 패널티가 있으면 숨겨진 장치의 수를 선택하는 것이 중요하지 않은 경우가 많다(그림 12.7 참조). 예측 접근법의 정신은 로컬 최대값에 대해 예측된 $p(c \mid x)$를 평균화하는 것이다. 단순한 평균으로 종종 충분할 것이다: m
여기서 발생하는 국소 최대값은 두 가지 매우 다른 유형이 있으며, 일부 국소 최대값은 여러 번(최대 수렴 공차까지) 발생한다. 평균은 두 가지 유형의 분류기보다

더 좋습니다.

12.5 서포트 벡터 머신

SVM(지원 벡터 시스템)은 이 분야의 최신 메서드 집합입니다. 그들은 열정적으로 홍보되었지만, 시험 문제와 발표할 많은 등급의 분류자의 구성원을 선택하는 것에 대한 선택 효과에 대해서는 거의 고려하지 않았다. 원래의 아이디어는 Bosser et al. (1992); Cortes and Vapnik (1995); Vapnik (1995, 1998); Cristianini와 Shawe-Taylor (2000) 그리고 Hastie 등에 있다. (2001, 4.5, 12.2, 12.3)는 기본 이론을 제시한다.

$g = 2$ 클래스에 대한 방법은 설명하기에 꽤 간단하다. 로지스틱 회귀 분석은 한 쪽에 모든 클래스 1 점이 있고 다른 쪽에 모든 클래스 2 점이 있는 초평면이 있는 분리 가능한 경우에 정확히 적합합니다. 그러한 초평면이 하나만 있는 것은 우연일 것이며, 로지스틱 회귀를 적합시키는 것은 그룹 간의 '갭' 중간에 결정 표면 $p(2 \mid x) = 0.5$를 맞추는 경향이 있다. 지원 벡터 방법은 간격의 중간, 즉 최대 여백(초평면에서 가장 가까운 점까지의 거리)을 직접 찾으려고 시도한다. 이것은 표준 방법으로 해결할 수 있는 2차 프로그래밍 문제이다. [3] 이러한 초평면은 초평면에서 정확히 여백 거리에 있는 데이터 포인트인 지지 벡터를 가지고 있다. 그것은 전형적으로 매우 좋은 분류기가 될 것이다.

문제는 일반적으로 분리된 초평면이 존재하지 않는다는 것이다. 이 어려움은 두 가지 방법으로 해결된다. 첫째, 라그랑주 곱셈기 $C > 0$로 '잘 맞지 않는' 거리의 총합에 대한 제약에 따라 일부 점이 여백의 잘못된 쪽에 있도록 허용할 수 있다. 이것은 거리의 합이 다소 자의적으로 사용되기 때문에 여전히 2차 프로그래밍 문제이다.

둘째, 변수 집합은 원래 변수 집합의 비선형 함수를 취함으로써 크게 확장된다. 따라서 분류 초평면 $f(x) = x^T +_0 = 0$을 찾는 대신 $f(x) = h(x| > .\mathrm{M} \gg p$ 함수 h_i의 벡터에 대한 $^T + \deg_0 = 0$. 그렇다면 최적의 분리 초평면을 찾는 것은 해결과 같다.

$$\min_{0,} \sum_{i=1}^{n} \left[1 - y_i f\left(x_i\right)\right]_{+ \frac{1}{2C} \|} |^2$$

여기서 두 클래스에 대해 $y_i = \pm 1$. 이것은 가중치가 감쇠된 로지스틱 회귀 분석

(멀티놈에 의해 적합될 수 있음)과 다르지 않은 또 다른 벌칙 적합 문제이다(Hastie et al., 2001, p. 380). SVM의 주장된 장점은 지원 벡터만 찾으면 되기 때문에 함수 h 패밀리는 무한 차원에서도 클 수 있다는 것이다.

패키지 e1071에는 함수 svm에 R용 SVM이 구현되어 있습니다. [4] 기본값은 잘 되지 않지만 게 데이터를 조정한 후 21개의 지원 벡터로 좋은 판별을 얻을 수 있다. 여기서 비용은 C이고 감마는 h를 형성하는 데 사용되는 커널의 계수이다.

m 단위로 10배 교차 검증을 시도할 수 있습니다.

$g > 2$ 클래스에 대한 확장은 훨씬 덜 우아하며, 몇 가지 아이디어가 사용되었다. svm 함수는 분류기가 각 클래스 쌍을 비교하여 구축되는 Knerr e al.(1990)에 귀속되는 하나를 사용하며, 결과 $g(g-1)/2$ 분류기 중 과반수가 예측 클래스를 결정한다.

12.6 법의학 유리 예시

법의학 유리 데이터 세트 fgl는 9개의 측정으로 6개의 클래스에서 214개의 포인트를 가지고 있으며 분류 방법에 대한 상당히 엄격한 테스트를 제공한다. 이미 살펴본 바와 같이(그림 4.17(99,5페이지)201, 309페이지의 4, 309페이지의 11.5 및 337페이지의 12.3) 유리의 유형은 작고 잘 분리된 그룹을 형성하지 않으며 한계 분포는 정규 분포와 거리가 멀다. 일부 소규모 클래스(예: 9, 13 및 17)가 있으므로 2차 판별 분석을 사용할 수 없습니다.

우리는 모든 방법에 대해 동일한 랜덤 파티션을 사용하여 10배 교차 검증을 통해 성능을 평가한다.

로지스틱 회귀 분석은 적절한 벤치마크를 제공하며(흔히 그렇듯이), 이 예제에서는 선형 판별 분석이 똑같이 잘 수행됩니다.

우리는 가장 가까운 이웃 방법을 사용하여 베이즈 위험의 하한을 약 10로 추정할 수 있다(Ripley, 1996, p. 196-7).

9장에서 우리는 약 6개의 크기의 분류 트리를 이 데이터 세트에 맞출 수 있다는 것을 보았다. 그룹별로 4개에서 7개까지 다양한 크기의 트리 선택에 대해 교차 검증해야 합니다.

우리는 V배 교차 검증을 통해 신경망 모델을 테스트하기 위한 몇 가지 일반적인 함수를 작성했다. 먼저 입력 범위가 $[0, 1]$이 되도록 데이터 세트를 재조정한다.

완전히 지정된 신경망을 맞추는 것은 간단하다. 그러나 여러 핏에서 평균을 내고 내부 교차 검증을 통해 숨겨진 단위의 수와 무게 감소량을 선택하려고 합니다. 이를 위해 우리는 다른 문제에 맞게 쉽게 사용하거나 수정할 수 있는 상당히 일반적인 함수를 작성했다. (코드에 대한 내용은 스크립트를 참조하십시오.)

이것은 신경망에 1000번 적합하며, 따라서 상당히 느리다(PC에서 약 30분).

이 코드는 교차 검증된 오류율을 기반으로 신경망을 선택한다. 대안은 로그 점수를 사용하는 것인데, 이는 유효성 검사 세트에서 이탈도를 찾는 것과 같습니다. 예측 클래스가 맞으면 0을 세고 그렇지 않으면 1을 세는 대신 실제 클래스 c에 대해 $-\log p(c \mid x)$를 세는다. 우리는 CVn2에서 m선을 n으로 대체함으로써 이 변형을 쉽게 코딩할 수 있다.

지원 벡터 시스템

다른 랜덤 파티션이 사용되므로 foHowing은 더 빠르지만 위의 결과와 비교할 수 없습니다.

학습 벡터 양자화 k-가장 가까운 이웃 방법으로서의 LVQ의 경우 적절한 메트릭을 선택해야 한다. 다음 실험에서는 원래 변수에 유클리드 거리를 사용했지만, 재조정된 변수나 마하라노비스 거리도 시도할 수 있었다.

우리는 더 작은 c1ass의 대표자가 너무 적기 때문에 c1ass보다 훨씬 먼저 설정했다. 1vqinit의 우리의 초기화 코드는 대표자의 수를 선택하는 데 있어 Kohonen의 것을 따른다. 이 문제에서는 각 c1ass에서 4개씩 24개의 포인트가 선택된다.

초기화는 랜덤이므로 결과가 다를 수 있습니다.

12.7 교정 그림

$p(c \mid x)$에 적합한 모델이 발견된 한 가지 척도는 예측 확률이 잘 보정된다는 것이다. 즉, 확률 p로 예측되는 이벤트의 약 p의 일부가 실제로 발생한다는 것이다. 확률 예측의 교정 시험 방법은 기상 예측과 관련하여 개발되었다(Dawid, 1982, 1986). 포렌식 유리 예제의 경우 각 사례에 대해 각 클래스에 대해 하나씩 6개의 확률 예측을 만들고 있습니다. 그것들이 진정한 예측이라는 것을 확실히 하기 위해, 우리는 교차 검증 절차를 사용해야 한다. 코드를 약간 변경하면 다음과 같은 확률 예측이 가능하다.

이것들을 플롯해서 매끄럽게 할 수 있습니다.

x 축을 따라 점들의 분포가 이 예보다 훨씬 더 고르지 않을 수 있기 때문에 여기서 10ess와 같은 적응형 대역폭을 가진 평활 방법이 필요하다. 결과는 그림 12.9와 같다. 이 그림은 특히 1에 대한 확률 선량에서 예측에 대한 상당한 과신뢰를 보여줍니다. 실제로 0.9보다 큰 확률로 예측된 사건 중 22/64만 발생했다. (근본적인 원인은 일부 기본 das 분포의 다중 모달 특성이다.)

보정 그림이 직선적이지 않은 경우 가장 좋은 해결책은 더 나은 모형을 찾는 것입니다. 때때로 과신뢰는 경미하며 주로 예측 추정치보다는 플러그인 사용에 기인한다. 그런 다음 그림을 사용하여 확률을 조정할 수 있습니다(세 개 이상의 c1ass에 대해 하나로 합하려면 추가 조정이 필요할 수 있습니다).