Chapter 4. 분류

선형회귀: 반응변수 Y가 양적이라고 가정. 하지만 많은 경우 반응변수는 범주형(categorical, 질적)이다.

분류: 관측치에 대한 질적 반응변수를 예측하는 것 (관측치를 범주, class에 할당하는 것에 관련되기 때문), 분류에 사용되는 방법들은 보통 질적 변수의 각 범주에 대한 확률을 예측 후 이것을 분류의 근거로 삼음

4장에서는 로지스틱회귀(logistic regression), 선형판별분석(linear discriminant analysis)을 다룬다

4.1 분류의 개요

분류문제의 사례

1. 응급실에 오는 환자가 3가지 의료 상태 중 어느 것의 증상을 가지고 있는가?
2. 사용자의 IP 주소, 과거 거래이력을 바탕으로 한 현지 진행되는 거래의 사기성 여부
3. 다수 환자들에 대한 DNA 염기서열 데이터를 기반으로 어떤 DNA 변이가 유해성을 띄는지에 대한 여부

회귀에서와 마찬가지로 분류에서도 훈련 관측치 셋이 있음(분류기 구성에 사용, 분류기는 train data와 더불어 test data에 대해서도 잘 동작해야 함)

4.2 왜 선형회귀를 사용하지 않는가?

환자의 증상을 근거로 응급 환자의 의료상태 예측하기(1번)

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

뇌졸중, 약물 과다복용, 간질발작의 진단을 양적 반응 변수 Y로 코딩할 수 있다. 이렇게 코딩하여 선형회귀모델을 최소제곱법을 사용하여 적합할 수 있지만, 이 코딩은 ordering 의 의미를 포함하고 있다. (결과에 순서)

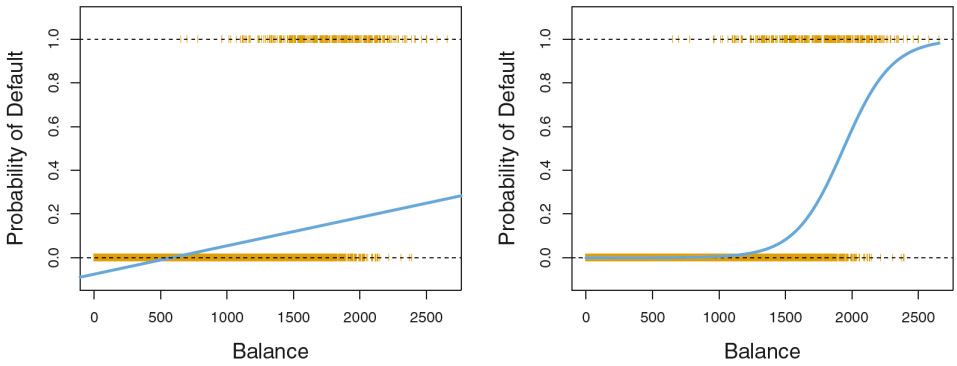
overdose가 stroke와 seizure 사이에 있고, 서로 차이가 동일함. 코딩을 다르게 하는 것도 가능하지만 다른 선형 모델이 만들어지면 검정 관측치에 대한 다른 예측을 하는 선형 모델이 만들어 질 것.

만약에 반응변수 값들이 약함, 적당함, 심함과 같이 순서를 가지고 정도의 차이가 같다면 앞선 사례처럼 코딩하는 것이 맞다. 하지만 일반적으로 3개 이상의 수준을 가지는 질적 반응변수를 선형회귀를 위해 양적으로 변경하는 자연스러운 방법은 없다

Binary인 경우, 가변수를 사용하여 반응변수를 코딩할 수 있다.

$$Y = \begin{cases} 0 & \text{if } \texttt{stroke}; \\ 1 & \text{if } \texttt{drug overdose}. \end{cases}$$

그 다음에 선형회귀를 적합하여 $\hat{Y}$ > 0.5 이면 약물 과다복용, 그렇지않으면 뇌졸중으로 예측할 수 있다. 위 코딩을 반대로 뒤집어도 결과는 동일하다.



선형회귀를 사용한 default의 추정확률과 로지스틱 회귀를 사용한 default에 대한 예측확률

선형회귀를 사용하여 얻은 $X\hat{\beta}$은 실제로 $Pr(\text{약물과다복용}|X)$의 추정치라는 것을 보여줄 수 있다. 하지만 선형회귀의 경우 일부는 [0,1] 범위 밖에 놓일 수 있어 확률로 해석하기 어렵다.

4.3 로지스틱 회귀

로지스틱 회귀는 반응변수 Y를 직접 모델링하지 않고 Y가 특정 범주에 속하는 확률을 모델링한다. (Default 자료에서 로지스틱 회귀는 연체확률을 모델링한다.)

$$Pr(default = Yes|balance)$$

이 값을 줄여서 $p(balance)$ 라 하고 범위는 0과 1 사이이다. $p(balance) > 0.5$인 사람은 $default = Yes$라고 예측할 수 있다. 임계치를 조정하여 사용할 수 있다.

$p(X) = Pr(Y = 1|X)$와 X 사이의 관계를 어떻게 모델링해야 하는가?

$p(X) = β_0 + β_1X$

balance를 사용하여 $default = Yes$를 예측하는 데 선형적인 방식을 사용할 때 문제점은 카드 대금이 0에 가까우면 예측확률이 음수가 되고 카드대금이 아주 큰 경우는 예측 확률이 1보다 크다. (카드대금과 상관없이 연체확률은 0과 1 사이어야 함)

이 문제를 해결하기 위해 모든 X값에 대해 0과 1 사이의 값을 제공하는 함수를 사용하여 $p(X)$를 모델링해야 한다. 로지스틱 회귀에서는 로지스틱 함수를 사용한다.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{4.2}$$

모델의 적합을 위해 최대가능도(maximum likelihood)라는 방법을 사용한다. 로지스틱 회귀모델의 경우 카드대금이 작을 경우 연체확률은 0, 카드 대금이 클 경우에는 1에 가까워지는 모습을 보인다. 로지스틱 함수를 통해 X 값에 관계없이 합리적인 예측값을 얻을 수 있다. 위의 식은 아래와 같이 표현할 수 있다.

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}. \tag{4.3}$$

$p(X)/[1-p(X)]$는 공산(odds)라고 하며 하며 항상 0과 ∞ 사이의 값을 가진다. 공산이 0에 가까우면 연체 확률이 매우 낮고, ∞에 가까우면 연체확률이 매우 높다.

양변에 로그를 취하면 다음을 얻을 수 있다

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X. \tag{4.4}$$

좌변은 로그 공산(log-odds), 로짓(logit)이라고 한다. 로지스틱 회귀모델은 X에 선형적인 로짓을 가진다. 선형회귀 모델에서 β_1은 X의 한 유닛 증가와 연관된 Y의 평균 변화를 제공한다. 반대로 로지스틱 회귀모델에서 X의 한 유닛 증가는 로그 공산을 β_1만큼 변화시킨다. (공산에 $e\beta_1$를 곱하는 것과 같다)

하지만 $p(X)$와 X 사이의 관계가 직선이 아니므로 β_1은 X의 한 유닛 증가와 관련된 $p(X)$의 변화와 일치하지 않는다.

회귀계수의 추정

β_0과 β_1은 알려져 있지 않고 training data를 기반으로 추정한다. 앞선 장에서는 최소제곱법을 사용하여 계수를 추정하였다. 로지스틱 회귀모델을 적합하는 데 일반적으로 최대가능도(maximum likelihood)를 사용한다. 예측한 개인에 대한 연체확률 $\hat{p}(x_i)$이 관측된 사람들의 연체상태와 가능한 가깝게 계수를 추정하려고 한다. 계수의 추정치를 $p(X)$에 대입하면 연체를 한 사람들에 대해서는 1에 가깝고 연체하지 않았던 사람들에 대해서는 0에 가까운 값을 제공한다.

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1-p(x_{i'})). \tag{4.5}$$

계수의 추정치는 이 가능도 함수를 *최대화*하도록 선택된다.

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

예측하기

계수들이 추정되면 카드 대금에 대해 Default 확률을 계산하는 것은 간단하다.

또한 가변수 기법을 사용하여 질적 설명변수들을 로지스틱 회귀모델과 함께 사용할 수 있다. 학생인지 아닌지 여부를 가변수로 코딩.

다중로지스틱 회귀

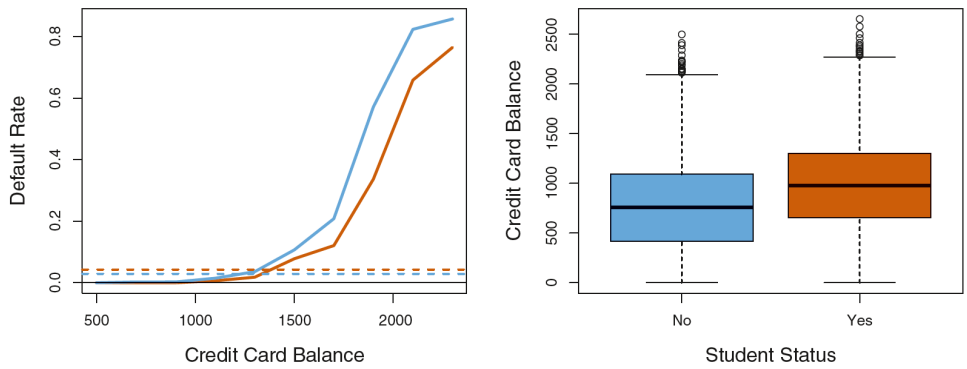
다수의 설명변수들을 사용하여 이진 반응변수 값 예측하기

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \qquad (4.6)$$

where $X = (X_1, \ldots, X_p)$ are p predictors. Equation 4.6 can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}. \qquad (4.7)$$

최대가능도를 사용하여 계수들을 추정한다.



표에 따른 연체확률에 상관하는 것이(감소, 증가) 다른 역설적 상황을 보여주는 그래프. 수평파선은 전체 연체율 실선은 대금의 함수로 나타낸 연체율.

오른쪽 박스 플랏은 학생과 학생이 아닌 사람의 카드 대금에 대한 것. 개별 학생은 학생 아닌 사람보다 연체율 낮음, 그러나 전체적으로는 지불해야 할 카드대금이 더 높음. 따라서 학생들이 아닌 사람들보다 연체율이 더 높음.

설명변수들이 서로 관련되어 있을 수 있을 때 단일 설명변수만 포함해서 회귀를 할 경우의 문제점을 보여준다.

confounding: 설명변수들 사이에 상관관계가 있을 때 (두 개 이상의 인자의 효과가 함께 나타나고 그것을 분리할 수 없을 때 그 인자들을 교락되어있다고 한다)

반응변수의 클래스가 2개보다 많은 로지스틱 회귀

클래스가 2개인 로지스틱 회귀 모델들은 다중 클래스 모델로 확장이 되지만 실제로는 판별분석(discriminant analysis) 방법을 다중클래스 분류에 일반적으로 사용된다. (LDA와 QDA)

4.4 선형판별 분석(LDA)

LDA 결정 규칙은 x의 선형 결합에만 의존

로지스틱 회귀는 로지스틱 함수를 사용하여 두개의 반응변수 클래스에 대해 Y의 조건부분포(conditional distribution) $Pr(Y=k|X=x)$를 모델링한다. 덜 직접적인 기법의 대안을 고려해보자.

대안적인 기법에서는 반응변수 Y의 각 클래스에서 X의 분포를 모델링, 다음에 베이즈 정리를 사용하여 $Pr(Y=k|X=x)$에 대한 추정치를 얻는다. (정규분포라고 가정할 경우 모델은 로지스틱 회귀와 형태가 비슷)

- 클래스들이 잘 분리될 때 로지스틱 회귀모델에 대한 모수 추정치는 아주 불안정하지만 선형판별 분석은 이런 문제가 없다
- 만약 n이 작고 각 클래스에서 설명변수 X의 분포가 근사적으로 정규분포이면 선형판별모델이 로지스틱 회귀모델보다 안정적
- 선형판별분석은 반응변수 클래스 수가 2보다 클때 일반적으로 사용

분류를 위한 베이즈 정리의 사용

질적인 반응변수 Y가 K개의 다르고 순서가 없는 값을 가질 때,

π_k(무작위로 선택된 관측치가 k번째 클래스에서 나올 전체 또는 사전 확률),

베이즈 분류기는 $p_k(X)$(관측치에 대해 주어진 설명변수값에 대해 그 관측치가 k번째 클래스에 속하는 확률)가 가장 큰 클래스로 관측치로 분류하며 모든 분류기 중에서 오차율이 가장 낮습니다.

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}. \tag{4.10}$$

선형판별분석 $(p = 1)$

$p = 1$(설명변수가 하나만 있다고 가정) $p_k(x)$를 추정할 수 있도록 $f_k(x)$에 대한 추정치를 얻고자 한다. $f_k(x)$가 정규분포라고 했을 때 1차원 경우의 정규밀도함수의 형태.

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right), \tag{4.11}$$

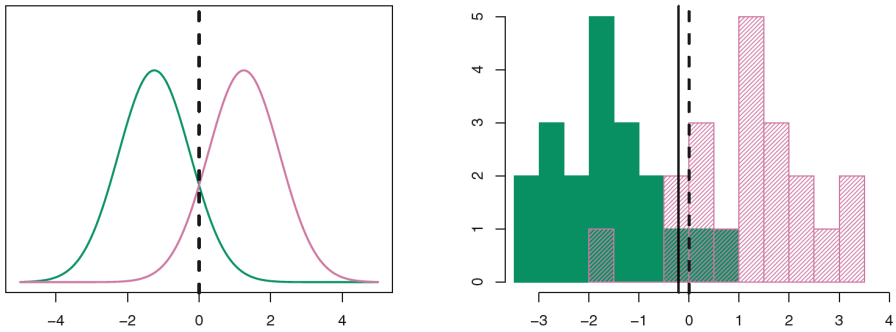
모든 K개 클래스에 대해 공통의 분산이 있다고 하고 (4.11)을 (4.10)에 대입

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}. \tag{4.12}$$

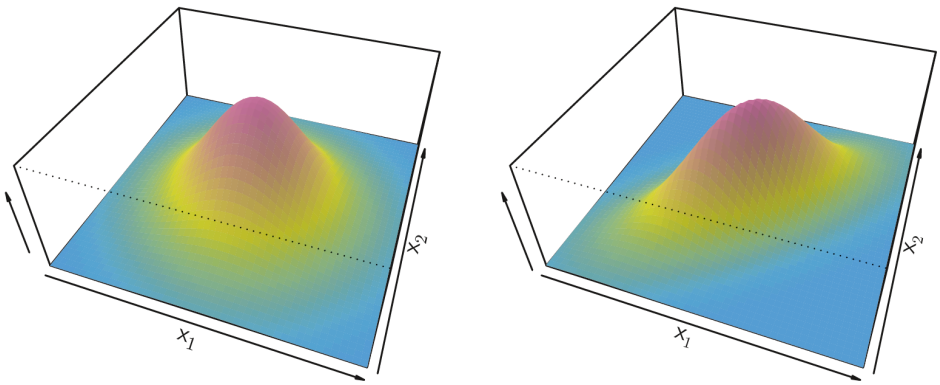
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{4.13}$$

4.12의 식에 로그를 취하고 항을 정리한 식이다. 베이즈 분류기는 4.12가 최대가 되는 클래스에 관측치 $X = x$를 할당한다. 베이즈 분류기는 이 식을 최대로 하는 클래스에 관측치를 할당하는 것과 동일하다고 보면 된다.

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}. \tag{4.14}$$



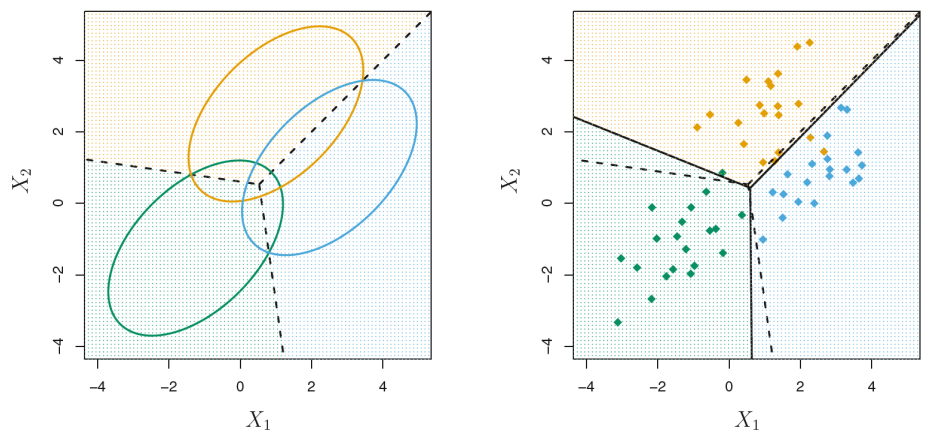
왼쪽은 2차원 정규밀도함수. 수직 파선은 베이즈 결정경계. 오른쪽은 두 클래스에서 20개 관측치 각각 추출하여 히스토그램으로 나타낸 것. 실선은 훈련데이터로부터 추정한 LDA 결정경계,



($p = 2$인 두개의 다변량 가우스 밀도함수) 설명변수들 사이에 상관성이 있거나 동일하지 않은 분산을 가지면 오른쪽처럼 왜곡될 수 있다.

선형판별분석 $(p > 1)$

특정 평균벡터와 공통의 공분산행렬을 가지는다변량가우스분포(다변량정규분포)를 따른다고 가정



그림은 세개의 클라스를 가진 예이다. 각 클래스의 관측치들은 클래스 특정 평균벡터와 공통 공분산행렬을 갖는 $p = 2$인 다변량가우스분포로부터 선택됩니다. 파선은 베이즈 결정 경계

다변량가우스분포는 각 설명변수가 1차원 정규분포를 따른다는 과정. 베이즈 분류기는 관측치 $X = x$를 다음 식이 최대가 되는 클래스에 할당한다.

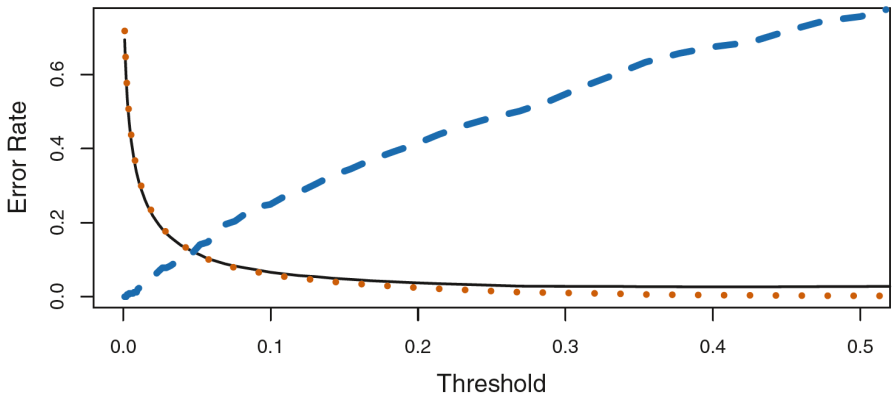
$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k \qquad (4.19)$$

이진 분류기에서 오류를 범할 수 있는데, 어느 것이 발생하는 지에 관심이 있고, 밑의 혼동 행렬은 이러한 정보를 나타낸다.

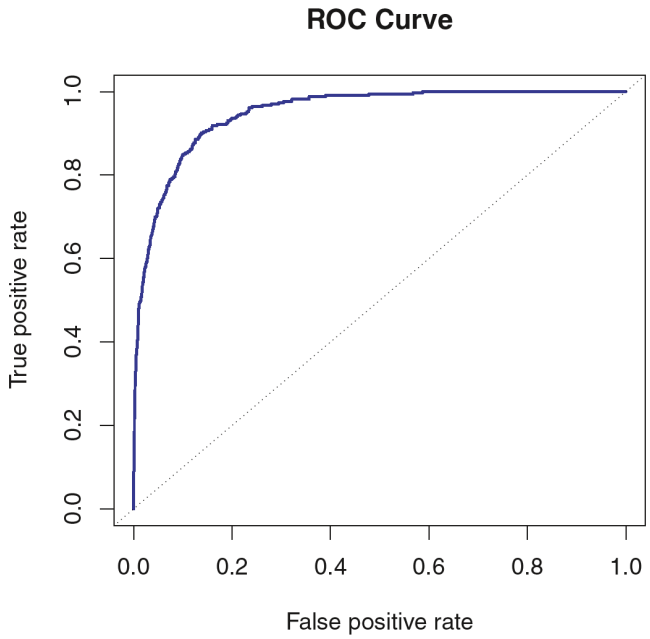
		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

실제로 오류가 났을 때 어떤 오류가 났고, 분류기의 성능을 개선하려면 어떤 decision을 내려야 하는지를 파악하기 위해 *혼동행렬*을 사용한다. 이 경우에는 민감도는 식별되는 실제 연체자의 비율이고, 특이도는 실제 식별되는 비연체자의 비율입니다.

여기서 LDA가 낮은 민감도(24.3%)를 가진다는 것을 확인 할 수 있는데, 이렇게 낮은 민감도를 가지는 이유는 LDA는 총오류율이 가장 낮은 베이즈 분류기에 근접하고자 하기 때문. 임계치를 조정하여 해결하고자 함



임계치를 낮추면 민감도는 감소하지만 총오류율은 약간 증가한다. trade-off가 발생하는 것을 고려하여 선택해야 한다.



ROC 곡선은 모든 가능한 임계치에 대해 2가지 오류를 동시에 나타내는그래프. 분류기의 전체적 성능은 곡선 아래의 면적이 클수록 더 좋은 분류기이다.

4.5 분류방법의 비교

앞서 로지스틱 회귀, LDA(선형판별분석)와 QDA에 대해 살펴보았고, 이전에 공부를 하면서 KNN에 대해서도 다루었다.

로지스틱 회귀와 LDA는 동기는 다르지만 밀접하게 연관되어 있다.

모든 상황에서 다른 것들보다 월등히 나은 방법은 없다

- 결정 경계가 실제로 선형일 때: LDA와 로지스틱 회귀 기법이 좋은 성능을 낸다
- 경계가 비선형적일 때: QDA가 더 좋은 결과를 낼 수 있다
- 훨씬 더 복잡한 비선형 결정경계의 경우: KNN과 같은 비모수적 기법이 더 나을 수 있다. (평활 수준을 주의 깊게 설정해야)