

4장 분류 연습 - TO BE CONTINUED

분류

임요한

서울대학교

Aug, 2018

“분류” 수업에서 다룰 내용

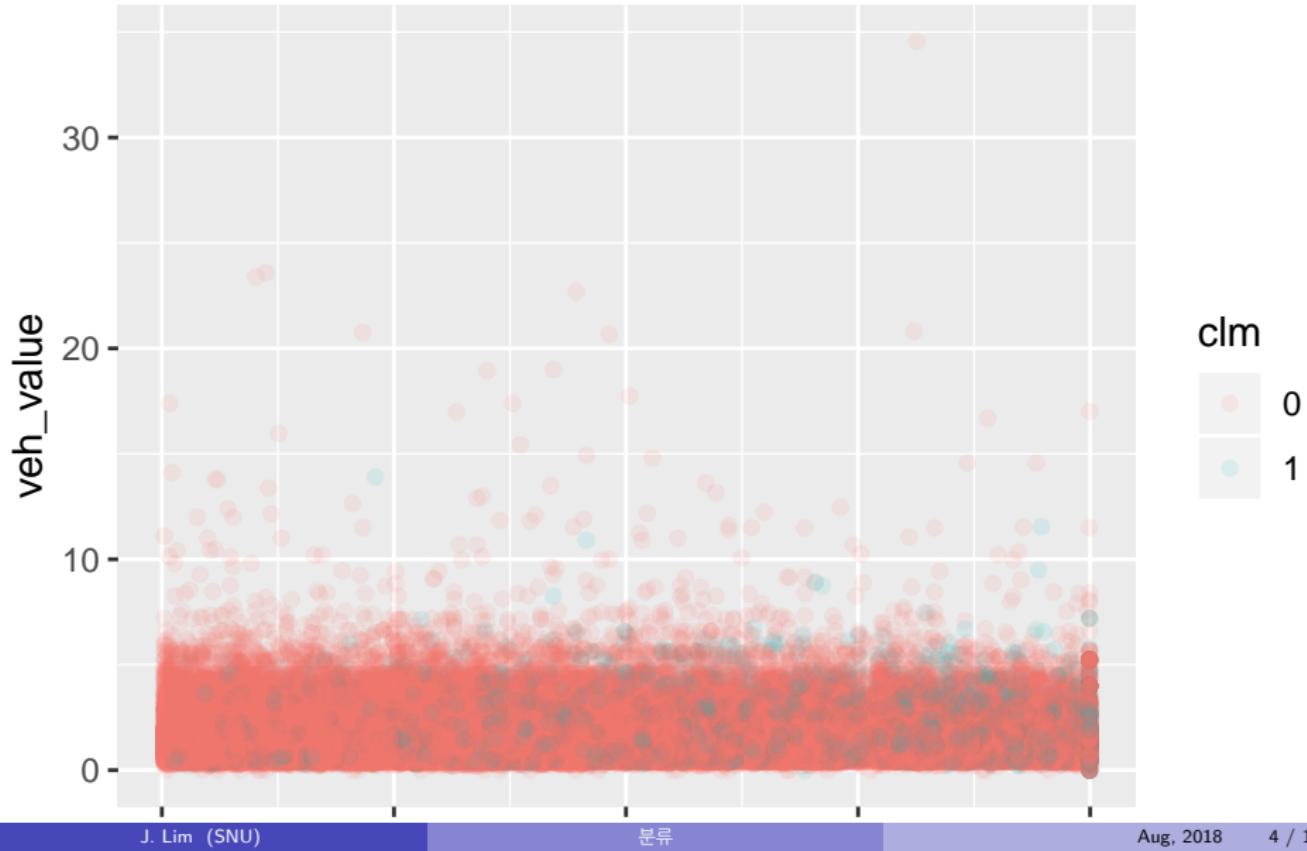
- 분류에 대한 소개
- 베이즈 분류기
- 최근접이웃방법
- 로지스틱 회귀
- 선형판별분석
- 혼동행렬, 오분류율, 민감도와 특이도, ROC 커브

보험자료

```
library(ggplot2)
ins=read.csv("insurance.csv",header=T)
ins=ins[,-c(3,4)]
ins$clm=as.factor(ins$clm)
head(ins)

##      clm exposure veh_body veh_age gender area agecat veh_value
## 1 0 0.3039014 HBACK 3 F C 2 1.06
## 2 0 0.6488706 HBACK 2 F A 4 1.03
## 3 0 0.5694730 UTE 2 F E 2 3.26
## 4 0 0.3175907 STNWG 2 F D 2 4.14
## 5 0 0.6488706 HBACK 4 F C 2 0.72
## 6 0 0.8542094 HDTOP 3 M C 4 2.01
```

```
ggplot(data = ins, aes(x =exposure, y = veh_value)) +  
  geom_point(alpha = 0.1, aes(color=clm))
```



알즈하이머자료

- diagnosis: “Impaired”, “Control”
- predictors: 130 variables
- 333 obs : 250 train samples, 83 test samples

```
load("AlzheimerDisease.RData")
head(predictors,n=1)

## ACE_CD143_Angiotensin_Converti ACTH_Adrenocorticotrophic_Hormon
## 1 2.0031 -1.386294
## Adiponectin Alpha_1_Antichymotrypsin Alpha_1_Antitrypsin
## 1 -5.360193 1.740466 -12.63136
## Alpha_1_Microglobulin Alpha_2_Macroglobulin Angiopoietin_2_ANG_
## 1 -2.577022 -72.65029 1.06471
## Angiotensinogen Apolipoprotein_A_IV Apolipoprotein_A1 Apolipop
## 1 2.510547 -1.427116 -7.402052 -0
## Apolipoprotein_B Apolipoprotein_CI Apolipoprotein_CIII Apolipop
## 1 -4.624044 -1.272966 -2.312635
## Apolipoprotein_E Apolipoprotein_H B_Lymphocyte_Chemoattractant_
## 1 3.754521 -0.1573491 2.2969
## BMP_6 Beta_2_Microglobulin Betacellulin C_Reactive_Protein
## 1 -2.200744 0.6931472 34 -4.074542
## CD40 CD5L Calbindin Calcitonin CgA Clusterin_A
## 1 -0.7964147 0.09531018 33.21363 1.386294 397.6536 3.55
## Complement_3 Complement_Factor_H Connective_Tissue_Growth_Facto
## 1 -10.36305 3.573725 0.530628
```

K-NN R code -1

```
library(class)
train.X=predictors[1:250,1:129]
test.X=predictors[251:333,1:129]
train.Y=as.numeric(diagnosis[1:250])
test.Y=as.numeric(diagnosis[251:333])
```

K-NN R code -2

```
set.seed(1)
knn.pred=knn(train=train.X,test=test.X,cl=train.Y,k=5,prob=TRUE)
table(knn.pred,test.Y)
```

```
##           test.Y
## knn.pred  1  2
##           1  3 12
##           2 12 56
```

K-NN 추가설명

- 함수 knn은 class 패키지 안에 있다.
- 다른 통계함수들은 모형을 적합한 후에, 적합한 결과를 이용해서 예측을 2 단계로 따로 하는데, knn은 한번에 예측을 한다.
- 인수 : train은 훈련자료의 설명변수를 행렬이나 데이터프레임으로, test는 예측을 할 시험자료의 설명변수값들을 행렬이나 데이터프레임으로 ci은 훈련자료의 반응변수로 팩터이다. k는 예측에 이용하는 이웃의 개수이다.
- 결과는 팩터이다. 확률은 attributes를 이용해 얻을 수 있다.
- 위의 코드는 $K = 1$ 인 경우 아래는 $K = 3$ 인 경우이다.
- set.seed를 쓴 이유 : knn은 tie가 있는 경우 랜덤하게 tie를 깨는데, 이를 고정시키기 위해 썼다.

K-NN - 강의록 주식예제

강의록 주식 예제 따라하기

- Credit default자료
- predictor: 1-23
- default: 24
- 표본수=30000: 25000 = training, 5000=testing

K-NN - 크레딧 디폴트

```
creditdefault=read.csv("creditdefault.csv",header=T)
creditdefault[,3]=as.factor(creditdefault[,3])
train.X=creditdefault[1:25000,1:23]
test.X=creditdefault[25001:30000,1:23]
train.Y=creditdefault[1:25000,24]
test.Y=creditdefault[25001:30000,24]
test=creditdefault[25001:30000,]
```

```
X=train.X  
head(X,n=1)
```

$X[1], X[2], X[4], X[5]$ 변수를 가지고 KNN을 적용하여 보자.