분야별 순서 & 함수

속성	분야	순서	패키지	ㅈ
정형 데이터마이닝	의사결정나무 (분류나무)	* 완전모델생성 (rpart : cp, minsplit 옵션), plotcp 그래프 - rpart(Survived~Pclass+Sex+Ticket+Fare+Cabin+Embarked, data=train.data, method="class", cp=0.0001, minsplit=5) * 교차검정 (printcp, which.min) xerror값이 가장 작은 행의 cp값을 찾는다 * 가지치기(prune) : prune(train.tree, cp=0.001466276) * predict(train.pr, newdata=test.data, "class")	rpart(rpart, prune)	
정형 데이터마이닝	로지스틱 회귀 나무	모델(lgm, summary, step) 5. 예측, 정분류율(predict, confusionMatrix) 6. ROC 모형(ROCR 패키지, prediction 함수 등) 로지스티 회귀 시 predict는 type="response"해야 확률 / / link는 log-odds값 ⇒ ifelse 이용해 0.5를 기준으로 0과 1을 분류	caret(confusionMatrix)	
정형 데이터마이닝	나이브 베이즈 분류	모델 (naiveBayes) 5. 데이터 검증(predict, type은 class와 raw) 6. 데이터 통합(data.frame 이용, 원본 데이터, class 데이터, raw 데이터) 7. 6번 class 데이터에 대한 정분류율 확인(confusionMatrix)	e1071(naiveBayes) caret(confusionMatrix)	
정형 데이터마이닝	랜덤포레스트	randomForest(종속~독립변수, data, ntree=50(기준몰라), mtry=sqrt(독립변수 수), importance=T) * (타이타닉 사례) randomForest(titan.formula, train.titan, ntree=500, mtry=3, nodesize=0.01*nrow(test.titan)) #랜덤포레스트모델$importance ⇒ 변수별 중요도 보여줌 (정확도, 노드 불순도 개선) # varImpPlot(랜덤포레스트모델) ⇒ 변수 중요도 보여줌 **randomForest 시 predict는 type="prob" or "class"** 경향보려면 prob!!! newdata에 사용할 x값을 지정할 때에는 주의 해야한다. data.frame의 colname을 회귀분석에서 사용한 설명변수 명과 동일하게 설정해야지 predict() command 사용시 오류가 발생하지 않는다.	randomForest(함수 동일)	
정형 데이터마이닝	다항로지스틱 회귀분석	# 로지스틱 회귀분석에서 분류 그룹이 세개 이상일 때 사용 # nnet 패키지의 multinom(종속~독립, data)	nnet	
정형 데이터마이닝	반복구문 사용	# for(i in k) result=numeric() k=1:20 for(i in k){ pred ← knn(train.df, test.df, class, k=i) t ← table(pred, test$credit.rating) result[i] ← (t[1,1]+t[2,2])/sum(t) } result which.max(result)		
통계 검정	T-test(1 sample)	* 원 샘플 티테스트 : 집단이 하나. (예시 : 파인트 무게는 진짜 320g인가?) * **정규성 검정(shapiro.test)** → 결과에 따라 t.test or wilcox.test *정규성 검정 통과하려면 피값이 0.05보다 커야 함.		http cee2
통계 검정	T-test(2 sample)	* 집단이 두개 이상이면 등분산/이분산 검증을 해야 함 1. 등분산 검정 : **var.test 함수 사용** ⇒ var.test(열값~그룹값, data) ⇒ p값이 나오는데 0.05보다 높으면 등분산, 낮으면 이분산 ⇒ 정규성, 등분산성을 만족시켜야 독립표본 T 검정을 할 수 있음. ⇒ 티테스트 옵션에서 var.equal=T, 이분산 시 var.equal=F 단측검정할 때 less와 greater 주의		
통계 검정	대응표본 T 검정	* 하나의 같은 그룹을 두번에 걸쳐 측정하고 차이가 있나 알아봄 1. 차이 변수를 하나 만든다 (후 데이터 - 전 데이터) 2. 통계분석 : t.test(post, pre, alternative = c("two.sided"), paired = T, conf.level = 0.95) - 데이터 넣는 자리에 사후, 사전 데이터 순으로 넣는다는 것 - paried = T ? : 두 개를 묶어서 볼 거냐		

속성	분야	순서	패키지	지
통계 검정	one way anova	* 집단이 세 개 이상일 때 집단의 평균 차이 검정 1. 등분산 검정 : bartlett.test(종속~독립, 데이터) 2. 아노바 분석 : aov(종속~독립, 데이터) 3. 사후검정 : TurkeyHSD(아노바 분석 결과) 4. 통계결과 그래프 그리기		
통계 검정	two way anova	* 두 변수 간의 교호작용까지 본다 1. aov 함수에서 모델을 입력할 때 다음과 같이 함. aov(치킨의 맛~온도+시간+온도:시간, 데이터) 2. 상호작용효과 그래프 : interaction.plot(온도, 시간, 치킨의 맛) ⇒ 그래프가 교차한다는 건 교호작용 효과가 있다는 것임. 3. 사후검증 - 1) 상호작용 효과가 있는 경우 : 온도, 시간 등 그룹을 나눠서 분석 2) 상호작용 효과가 없는 경우 : 그룹수에 따라 티테스트 혹은 아노바 (사실 아까 표 준용하면 결과는 이미 나와있음. 안 해줘도 됨)		
통계 검정	상관분석	cor(corrData, use="complete.obs", method=c("pearson")) # use 옵션 : 결측치 데이터 있으면 빼고 계산하라는 것 # method 옵션 : 피어슨 (비모수 방법론 설정 따로 있음-스피어만 등)		
통계 검정	교차분석(카이제곱분석)	* 종류 : 사전설계(실험군, 대조군 미리 나눈다), 사후설계(결과 보고 원인 추적) * 사전교차분석 1) 분할표 만든다 : n.table, prop.table, addmargins 2) 카이스퀘어 분석 : chisq.test(nTable) 3) 상대위험률 계산 * 사후교차분석 n.hand ← table(survey$W.Hnd) chisq.test(n.hand, p=c(0.2,0.8)) # 전체 응답자 중 왼손잡이 비율이 20%, 오른손잡이 비율이 80%		
통계 검정	다중회귀분석	1. 통계분석 : lm(종속~독립, 데이터) 2. 다중공선성 검증 : vif(mregmodel) - 독립변수가 서로 비슷한지 확인. 10 이상이면 문제 있음. 3. 변수제거 : step(회귀모델, direction="backward" or "forward", trace=T) 4. 결과 확인 : AIC 값이 증가하면 모델이 안 좋아지는 것.		http cee2
통계 검정	회귀분석 잔차 검정	1. 정규성 검정 : shapiro.test(lmmodel$residuals) 2. 독립성 검정 : lmtest 패키지의 dwtest(더빈왓슨) 3. 등분산성 : plot(lmmodel)		http cee2
데이터 전처리	정규화(0~1)	1. 데이터 불러오기 2. 변수별(열)로 max와 min 구하기(apply, 1 or 2, 적용함수) 2) 3. 정규화(scale 함수) scale(data, center=min, scale=max-min)		
데이터 전처리	요인분석	1. 데이터 정규화(apply로 max,min 구한 뒤 scale) 2. 요인분석 factanal(정규화 결과, 요인 수)		
데이터 전처리	결측값	1. 결측값 확인 : colSums(is.na(train.data)) # colSums 칼럼별 합계 / summary 2. 결측값 제거 : # na.omit() 결측값이 있는 행 전체를 데이터셋에서 제거 3. 결측값 대체 : train.data$Age[is.na(train.data$Age)] ← median(train.data$Age, na.rm=T) table(is.na(titan.full$Age))		
데이터 전처리	결측값 대체	회귀모델 사용 (lm) * 회귀모델 만들 때 이상치도 제거된 깨끗한 데이터를 기반으로 돌린다. - boxplot.stats(full.data$Fare) [1]보다 적은 값, [5]보다 큰값은 아웃라이어 - outlier.filter ← full.data$Fare < boxplot.stats(full.data$Fare)[5] * 아웃라이어가 아닌 데이터 행만 취하겠다! ⇒ full.data[outlier.filter,] : 아웃라이어 없는 깨끗한 데이터 * lm 돌림 - lm(공식, data=full.data[outlier.filter,], **na.action=na.omit)** - predict 함수를 쓸 때 lm 공식에서 선택된 변수들은 돌릴 데이터의 열과 일치해야 한다. **- newdata에 들어가야 할 값은 full.data[fare가 결측된 행 index, lm 공식에 들어갔던 열 정보]** - fare.row ← full.data[is.na(full.data$Fare), c("독립변수1", "독립변수2".....) - predict(아까 lm 모델, data=fare.row) * 결측값 집어넣기 ifelse 함수 사용하면 됨		http fe1a
데이터 전처리	파생변수	**1) transform 함수 사용** 데이터프레임명 ← transform(데이터프레임명, 새변수명=함수(벡터명, ...)) * transform만 덜렁 쓰지 말고 해당 데이터프레임에 적용을 해야지 반영된다. * car.df ← transform(car.df, result=scale(qsec, center=min.qsec, scale=max.qsec-min.qsec)) **2) dplyr 패키지의 mutate 함수 사용** * 데이터프레임에 적용을 해야지 반영이 된다!!!!!!!!!!!!!!!! car.df ← car.df %>% mutate(result=scale(qsec, center=min.qsec, scale=max.qsec-min.qsec))	dplyr	
데이터 전처리	연속형 변수를 범주형 변수로 구간 나누기	**transform 과 cut 함수 사용** * transform(데이터프레임명, 새변수명 =cut(열값, breaks=c(0,10,20...), labels=c("0이상 10 미만" ....), right=F)		http fe1a
데이터 전처리	더미변수 만들기	1) library(caret) 2) dummy ← **dummyvars**("~범주형 변수1+범주형 변수 2 ...", data=데이터프레임명) 3) predict(dummy, data=아까랑 같은 데이터프레임명) 4) 3번을 as.data.frame 더미바 함수 쓸 때 "~범주형변수 +..." 쌍따옴표 쓴다,,,	caret	

속성	분야	순서	패키지	지
데이터 전처리	소속	A벡터 %in% B벡터 : A에 있는 값들 중에 B에 있는 애들이 있니? ⇒ 각 행마다 있으면 트루, 없으면 펄스 fo1 ← fifa.data$Position_Class %in% c("LS","ST", "RS", "LW","LF","CF","RF", "RW") fifa.data$Position_Class ← ifelse(fo1, "Forward", fifa.data$Position_Class)		http fe1a
데이터 전처리	Date로 바꾸기	library(caret) as.Date 함수	caret	
최종결과 제출	데이터 합치기 (data.frame), write 함수	# PassengerId ← test.titan$PassengerId #테스트데이터의 패신져아이디 # output.df ← as.data.frame(PassengerId) #output.df라는 데이타프레임 만듦 # output.df$Survived ← Survived # 내가 구한 서바이브값 넣음 # (주의) 씨바인드 하면 안됨. **씨바인드는 데이터프레임간의 결합. value라면 데이터프레임부터 만들어야.** # write.csv(output.df, file="./output.csv", row.names=F) write.csv 함수에 들어가는 x는 데이터프레임이어야 한다!!!!!!!!!!!!!!!!!!!!		
최종결과 제출	특정 조건의 열의 개수 세기	* 열의 개수를 세고 싶거든 **length와 which**를 사용하는 방법이 있다. length(which(car.df$result > 0.5)) * (주의) nrow 함수는 데이터의 특정 열에 적용 할 수 없고, 벡터에도 적용할 수 없습니다. nrow를 굳이 써야한다면 데이터 프레임에 해당 벡터를 집어 넣어서 행을 구해야..... **nrow(car.df[car.df$result > 0.5,])**		http cee2
데이터 전처리	엑셀 데이터 불러오기	1. library(readxl) 2. read_excel("data/ddd.xlsx", sheet="시트명", col_names=T)		