**학습계획서**

|  |  |  |  |
| --- | --- | --- | --- |
| 팀 | 준비데이터 | 구성원 | 김호영, 안상현 |

|  |  |  |  |
| --- | --- | --- | --- |
| 일정 | 발제자 | 주제 | 주요내용 |
| 1일차  ( 5 / 27 ) | 안상현 | -빅데이터 이해  -관련 개념 | - 빅데이터 개념 정의  - 데이터베이스, NoSQL, SQL |
| 2일차  ( 5 / 28 ) | 안상현 | -빅데이터 분석 기획 | - 분석 기획 배경 및 필요성  - 다양한 분석 기획 방법 |
| 3일차  ( 5 / 29 ) | 김호영 | -빅데이터 전처리 | - 데이터 전처리 방법  - R의 데이터 타입과 구조 |
| 4일차  ( 5 / 30 ) | 김호영 | -빅데이터 처리 방법(Hadoop) | - HDFS  - MapReduce  - Hadoop Eco System |
| 5일차  ( 5 / 31 ) | 김호영 | -빅데이터 분석  (통계 분석1) | - 통계 기초 개념, 대푯값  - 표본과 모집단, 정규분포 |
| 6일차  ( 6 / 3 ) | 김호영 | -빅데이터 분석  (통계 분석2) | - t검증, 분산분석  - 상관분석, 회귀분석 |
| 7일차  ( 6 / 4 ) | 김호영 | -빅데이터 분석  (데이터마이닝1) | - 데이터마이닝의 개념과 분류  - 텍스트마이닝 개념 및 절차 |
| 8일차  ( 6 / 5 ) | 안상현 | -빅데이터 분석  (데이터마이닝2) | - 군집분석, SVM, 주성분분석  - 의사결정나무, 인공신경망 |
| 9일차  ( 6 / 7 ) | 안상현 | -빅데이터 분석  (신기술 동향) | - 딥러닝 개념 및 절차  - 딥러닝 소프트웨어(tensorflow, keras 등) |
| 10일차  ( 6 / 10 ) | 안상현 | -빅데이터 시각화 | - 시각화 이론 및 프로세스  - R을 활용한 시각화 방법(ggplot2) |

**학습 정리**

|  |  |  |  |
| --- | --- | --- | --- |
| 팀 | 준비데이터 | 구성원 | 김호영, 안상현 |

|  |  |  |  |
| --- | --- | --- | --- |
| 일정 | 발제자 | 주제 | |
| 5/27 | 안상현 | 1. 빅데이터 정의 및 관련 개념  2. 인터넷 강의 요약 | |
| 주요 내용 요약 | | | |
| **Ⅰ오늘의 ADsP 요약**   1. 빅데이터의 정의   기존 DBMS(데이터베이스 관리시스템)에 보관하기 힘든 양의 데이터 집합  이 집합에는 정형 데이터, 비정형 데이터 등 다양한 타입이 포함될 수 있다.  최근 SNS 사용 활성화, 데이터 처리 비용 감소, 클라우드 발달 등으로 빅데이터가 유망 산업에 포함되고 있다.   1. 빅데이터의 특징(3V + α)   ---- 전통 3V  Volume(양) : 데이터의 양이 매우 커졌다.(수십 TB ~ 수 PB)  Variety(다양성) : 생성/수집되는 데이터의 종류가 매우 다양하다(텍스트, 이미지, 동영상, 숫자 등)  Velocity(속도) : 데이터의 생성 속도가 매우 빠르다  ---- 새로운 V  Value(가치) : 방대한 양의 데이터에서 가치있는 정보를 이끌어낼 수 있다  Veracity(정확성) : 가치를 이끌어내기 전에 수집한 데이터가 정확한지 살펴야 한다  Variability(가변성) : 데이터는 맥락에 따라 의미가 달라질 수 있다  Visualization(시각화) : 빅데이터를 분석한 내용을 이해하기 쉽게 시각화해야 한다   1. 빅데이터 분석 방법   - 통계 분석 : 기술통계(빈도분석 등), 추리통계(카이제곱 검정, T검정, 분산분석, 상관분석, 회귀분석, 주성분 분석)  - 데이터 마이닝 : 로지스틱 회귀분석, 인공신경망, 의사결정나무, 앙상블 모형, 군집분석, 연관분석, 텍스트마이닝   1. 빅데이터 관련 개념   - RDBMS(관계형 데이터베이스)  엑셀과 비슷한 방식으로 데이터를 저장하는 소프트웨어이다. 엑셀에서 행과 열은 RDBMS에서 각각 row와 column으로 부른다.  컬럼은 데이터를 식별하는 역할을 맡는데, 구체적으로 기본키, 후보키 등으로 나눌 수 있다.  규칙적으로 데이터를 저장하므로 SQL을 통해 데이터를 삽입, 조회, 삭제, 수정이 가능하다.  자주 사용되는 소프트웨어로는 Oracle, MySQL, SQLite이 있다.  - NoSQL(Not Only SQL)  기존 RDBMS와 다른 방식으로 데이터를 저장하는 방식이다. 소프트웨어마다 각기 다른 방식으로 데이터를 저장하므로 통일성이 없다.  일반적으로 자주 사용하는 저장 방식은 JSON과 같이 key와 value로 저장한다.  따라서 SQL 대신 Key에 대한 Put(저장)과 Get(읽기)만 지원한다.  자주 사용되는 소프트웨어로는 몽고DB, 카산드라, Hbase가 있다.  - 데이터, 정보, 지식, 지혜  데이터는 가공 전의 순수한 수치나 기호를 나타낸다. raw data라고 불리며, 그 자체로는 어떠한 의미를 가지지 않는다.  정보는 데이터를 가공하고 연결시킴으로써 패턴을 찾아 의미를 부여한 경우에 사용한다.  지식은 정보를 연결함으로써 만든 결과물이다. 지식을 통해 어떻게 행동할지 예측할 수 있다.  지혜는 지식을 내면화시켜 개인별로 만든 창의적 아이디어이다.  **Ⅱ 오늘의 강의 요약**   1. **러닝패킷 : 통계의 기초 1강 - 공통**   - 통계가 필요한 이유  통계는 사회 및 다양한 현상을 이해해 규칙성을 발견할 수 있는 기술이므로 통계를 바탕으로 예측함으로써 위험을 대비할 수 있다.  - 자료의 2가지 형태  ① 비연속변수/질적변수/범주형변수 : 몇 개의 유형으로 분석대상을 구별하기 위한 척도, 사칙연산을 적용하기 어렵다.(명목변수 + 순서변수)  ② 연속변수/양적변수 : 분석대상의 성질 크기가 얼마인지 파악하는 척도, 사칙연산이 가능하다.(등간변수 + 수치변수)  ③ 명목변수 : 각 항목에 임의로 숫자를 붙여 구별하는 척도, 항목 간에 연관성이 없고 순서가 우열을 나타내지 않는다. ex) 성별, 종교, 혈액형  ④ 순서변수 : 각 항목을 순서적으로 나열하지만 항목 간 차이가 일정하지 않은 경우  ex) 계층(상/중/하), 만족(1/2/3/4/5)  ⑤ 등간변수 : 수치 간 차이는 일정하지만 절대 0점이 없는 경우 ex) 온도, IQ점수  ⑥ 수치변수 : 절대 0점이 있어 상대적인 크기비교가 가능한 척도 ex) 키, 몸무게  - 범주형 자료의 요약  범주형 자료는 사칙연산이 불가능하므로 우열을 나타낼 수 없다. 따라서, 단순 빈도로 비교하거나 빈도를 잘 보여주는 그래프를 이용해야 한다.  ex) 도수분포표, 막대그래프, 원도표  - 양적 자료의 요약  양적 자료는 사칙연산이 가능하므로 비교가 가능하다. 다양한 대표값과 그래프를 통해서 표현할 수 있다. 다만, 범위가 넓으므로 이용자가 한눈에 이해할 수 있게 그래프를 표시해야 한다.  ex) 산점도, 히스토그램, 막대상자, 꺾은선 그래프   1. **빅데이터 분석과 R프로그래밍 1강 - 김호영**  * R이란?   1995년 Ross Ihaka and Robert Gentleman 이 S언어를 개선하여 개발한 오픈소스 통계분석 프로그램으로 빅데이터 분석 도구이다.  다양한 통계기법과 시각화 도구 제공하며, 통계 기법에는 통계기법, 데이터마이닝, 기계학습, 텍스트마이닝, 인공지능 분석툴이 있고,  시각화 도구에는 Ggplot을 이용한 시각적분석, 공간분석 맵이 있다.  C++, JAVA, Python 등 다른 프로그램과 쉽게 연동 가능하다는 특징이 있다.   * 전세계 학계나 연구기관에서 범용적 사용하며 기업체에서도 R을 탑재엔진으로 기업형분석 플랫폼개발 중   + 확장자는 \*.r * R 패키지 종류   데이터마이닝 툴   * + svm, cluster, tree, …   텍스트마이닝 툴   * + NLP, wordcloud, tm   그래픽툴   * + ggplot, lattice * 추가패키지 설치(install.packages)   항상 인터넷이 연결된 상태로 사용해야한다.  설치는 한번만 사용  해당 패키지를 사용할때 library는 사용할때마다 필요하다.  패키지 이름을 모를 경우에는 웹검색을 통해 찾을 수 있으며, 설치 후 help를 이용하여 상세설명을 알 수 있다.   * R프로그램을 수행하는 단축키   + Ctrl + Enter   + 수행하고자 하는 프로그램 부분을 선택하고 수행   + 주석은 #으로 처리가능  1. **머신러닝기법과 R프로그래밍 1강 - 안상현**   - 데이터마이닝 기법 분류  예측 : 종속변수의 값이 연속적이다 ex) 다중회귀분석, 주성분분석, 신경망  분류 : 종속변수의 값이 범주적이다 ex) 로지스틱 회귀분석, 의사결정나무, SVM  군집/연관분석 : 종속변수의 값이 없고, 독립변수(특성)만 존재  - 다중회귀모형  종속변수를 설명하는데 k개의 독립변수가 있는 경우 사용  각 회귀계수의 해석 : 다른 독립변수들이 일정할 때, 해당 독립변수 1단위를 변화시켰을 때의 평균변화량  변수선택방법 : 전진선택법, 후진제거법, 단계별방법  회귀모형의 평가 : 잔차분석을 통해 확인(선형성, 독립성, 정규성, 등분산성, 이상치)  다중공선성 : 독립변수들 사이에 상관관계가 높은 경우 발생하기 쉬움  → 회귀계수의 부호가 바뀌거나 회귀모형의 안정성이 떨어짐  다중공선성이 발생한 경우 두 독립변수 중 하나만 선택하거나 주성분회귀법 사용  - 과적합  학습표본을 바탕으로 오분류율을 0이 되도록 인위적으로 만드는 경우  과적합을 시키면, 실제 데이터를 적용했을 때 더 높은 오분류율이 발생한다  과적합 발생 방지를 위해 교차검증(cross-validation) 실시한다  - 교차검증  ① 수집한 전체 데이터를 학습 데이터와 검증 데이터로 랜덤하게 나눈다  - 일반적으로 학습 데이터 70%, 검증 데이터 30%로 나눔  ② 학습 데이터로 분류규칙/모형을 만든다  ③ 만든 분류규칙에 검증 데이터를 넣는다  ④ 2개의 오분류율이 비슷하면 해당 규칙/모형을 현장에서 활용할 수 있다  **참고 문헌**  - 데이터 관련 개념  <https://ko.wikipedia.org/wiki/%EB%B9%85_%EB%8D%B0%EC%9D%B4%ED%84%B0>  <http://egloos.zum.com/yjhyjh/v/39721>  이종익, 박민석 편저, 「사회조사분석사 2급 필기」, 시대고시기획, 2014.  - 러닝패킷 강의  <http://www.kocw.net/home/search/kemView.do?kemId=694004>  - 빅데이터 분석과 R프로그래밍  <https://pabi.smartlearn.kr/courses/course-v1:POSTECH+DSB101+P902/course/>  - 머신러닝 기법과 R프로그래밍  <https://smartlearn.kr/courses/course-v1:POSTECH+DSB103+P902/course/> | | | |
|
|
|
|
|
|
|

**학습 정리**

|  |  |  |  |
| --- | --- | --- | --- |
| 팀 | 준비데이터 | 구성원 | 김호영, 안상현 |

|  |  |  |  |
| --- | --- | --- | --- |
| 일정 | 발제자 | 주제 | |
| 5/28 | 안상현 | 1. 빅데이터 분석 기획의 필요성  2. 빅데이터 분석 기획의 다양한 방법 | |
| 주요 내용 요약 | | | |
| **Ⅰ오늘의 ADsP 요약**   1. 빅데이터 분석 기획   기획 : 무엇을, 왜 할 것인지 결정하는 절차( ↔ 계획 : 어떻게 할 것인지 결정)  빅데이터 분석 기획 : 비즈니스 목표 달성 최적화를 위해 의사결정과 실행에서 필요한 정보와 인사이트를 과학적인 분석을 통해 제공하는 분석체계 수립  ( = 업무에 필요한 질문이 무엇인지 찾기 위해 분석질문을 먼저 정의하고 분석하기 위해 필요한 데이터가 무엇인지 정의하는 방식)  빅데이터 분석 기획의 필요성 : 특정 기업이 경쟁사보다 차별화된 상품/서비스를 제공하기 위해, 비용 절감 및 수익률 향상을 위해 실시   1. 빅데이터 분석 기획 방법론   ① KDD 분석 방법론 : 1996년 Fayyad가 체계적으로 정리한 데이터마이닝 프로세스  - 데이터셋 선택 → 데이터 전처리 → 데이터 변환 →  데이터마이닝 → 데이터마이닝 결과 평가  ② CRISP-DM 분석 방법론 : 1999년 SPSS, NCR 등의 기업들이 발표한 버전  - 업무 이해 → 데이터 이해 → 데이터 준비 → 모델링 → 평가 → 전개  각 단계별로 단절된 것이 아니라 피드백을 통해 단계별 완성도를 높임  ③ 빅데이터 분석 방법론  - 분석 기획 → 데이터 준비 → 데이터 분석 → 시스템 구현 → 평가 및 전개  - 분석기획 : 비즈니스 이해 및 범위 설정(관련자료 조사 및 방향 설정), 프로젝트 정의 수립, 프로젝트 위험계획 수립  - 데이터 준비 : 필요 데이터 정의, 데이터 획득방안 수립, 데이터 스토어 설계, 데이터 수집  - 데이터 분석 : 비즈니스 룰 확인, 텍스트 데이터 및 탐색적 분석, 모델링, 모델 평가  - 시스템 구현 : 시스템 설계 및 시스템 구현, 시스템 테스트  - 평가 및 전개 : 모델 발전 계획, 프로젝트 평가 및 보고  **Ⅱ 오늘의 강의 요약**   1. **러닝패킷 : 통계의 기초 2강 - 공통**   - 대푯값 : 데이터를 요약해 중심성, 패턴, 의미 등을 파악하는 기술통계 방법 중 하나  ① 산술평균 : 데이터의 총합을 데이터의 개수로 나눈 값  → 이상치의 영향을 받기 쉬움  ② 중앙값 : 데이터를 크기 순으로 나열했을 때 중앙에 오는 값  관측치 개수가 홀수면 (n+1)/2 번째 관측치, 짝수면 중앙에 있는 2개 관측치의 평균  → 이상치의 영향을 받지 않음  ③ 최빈값 : 도수가 가장 많은 데이터의 값  → 이상치의 영향을 받기 쉬움  ④ 최빈값, 중앙값, 평균의 관계  가) 정규분포 : 최빈값 = 중앙값 = 평균  나) 우꼬리분포 : 최빈값 < 중앙값 < 평균  다) 좌꼬리분포 : 최빈값 > 중앙값 > 평균  ⑤ 범위 : 최댓값 - 최솟값  → 이상치의 영향을 받기 쉬움  ⑥ 사분위수 : 데이터를 크기순으로 배열했을 때 4등분한 3개의 수치  → 제1사분위수(하위 25%), 제2사분위수(중앙값), 제3사분위수(하위 75%)  ⑦ 다섯 숫자 요약 : 최솟값, Q1(제1사분위수), 중앙값, Q3(제3사분위수), 최댓값  ⑧ 상자그림 : 위 다섯 숫자 요약을 바탕으로 만든 그림  - 제1사분위수 ~ 제3사분위수 : 상자로 나타냄  - 제2사분위수 : 상자 안에 굵은 선으로 표현함  - IQR : Q3 - Q1  - Q1 - (IQR \* 1.5) / Q3 + (IQR \* 1.5) : 상자의 바깥에 선으로 표시한다  - 그 외 : 최솟값이나 최댓값, 특이값이 위 선에 포함되지 않는 경우 o나 \*로 표시한다  ⑨ 분산 : 편차(평균으로부터의 차)를 제곱한 것의 평균 → 값이 커쳐 비교가 어려움  ⑩ 표준편차 : 분산의 제곱근 → 분산이나 표준편차의 값이 작을수록 데이터가 평균에 집중되었으며, 값이 커질수록 데이터가 평균에서 떨어져 있다   1. **빅데이터 분석과 R프로그래밍 2강 - 김호영**   ①벡터 생성  → x<-c(1,3,5,7,9)  - subset  → x[-1]  - sequence of 20 values  - y을 0부터 10까지, 20개의 값을 생성  → y<-seq(0,10, length=20)  - sequence of (1 to 10) by 0.5  - y2를 0부터 10까지 0.5씩 간격을 두고 값을 생성  → y2<-seq(0,10, by=0.5)  - replication 함수  - 1부터 4까지 두번을 반복하여 생성하라는 의미  → z1<-rep(1:4, 2)  ② 벡터의 결합  - cbind : column bine (열기준으로 결합)  → c2<-cbind(x, c1)  - rbind : column bine (열기준으로 결합)  → c2<rbind (x, c1)  ③ 행렬 생성  -matrix 함수 이용  - 행의 수, 열의 수  - 1열부터 채우는것이 default  → m<-matrix(1:6, nrow=2, byrow=T)  -array 함수 이용(고차원 행렬)  → a1<-array(c(1:18), dim=c(3,3,2))  ④ 객체이름 정의  -벡터이름 주기  - (0,1) 값을 갖는 벡터에 gender에 0=female, 1=male이라는 값을 부여  → gender<-c(0,1)  → names(gender)<-c("female", "male")  -범주형변수 생성(factor 사용)  -고유한 값이나 범주 수가 제한된 변수(ex: gender, 종교 등)  -factor는 범주형 변수로 정의하는 함수  - 순서가 기본적으로 없지만 ordered명령을 통해 순서 정할 수 있다.  → size<-c("S", "M","L","XL")  → size\_factor3 <- factor(size, ordered = TRUE, levels =c("S", “M","L","XL"))  - 행렬에 이름주기(dimnames, colnames, rownames)  -dimnames(x)[[2]]에서 [2]은 열이고, 만약 [1]이면 행을 의미한다.  → dimnames(x)[[2]]<-paste("x",1:3,sep="")  와  → colnames(x) <- c("x1", "x2", "x3")  은 같은 명령어이다.  ⑤ 행렬과 데이터 프레임 (as.data.frame, is.data.frame)  - 데이터 프레임은 객체값들을 행렬로 저장할 뿐 아니라 변수명, 관측치번호 등 여러가지 정보를 가지는 객체. 따라서 행렬을 데이터 프레임으로 인식하기 위해서는 as.data.frame로 정의 필요.  - 데이터 프레임의 정의  → 각 열은 변수열, 각 행은 관측치이다.  → x라는 데이터에서 x1 변수값들을 보여줌  → summary 함수를 통해 데이터 요약 통계량을 알 수 있다.  ⑥ 고유치와 고유벡터  - 고유 벡터는 그 선형 변환이 일어난 후에도 방향이 변하지 않는, 영벡터가 아닌 벡터이며, 이때의 상수배 값을 고유치라 한다.  -통계분석에서는 공분산 행렬의 특성을 파악하기 위해서 사용되거나, 데이터의 변환(ex.차원축소)에 사용됨.   1. **머신러닝기법과 R프로그래밍 2강 - 안상현**   ① k-인접기법  - k개의 가장 가까운 이웃들을 사용해 분류하는 방법  - 장점 : 단순하고 효율적이다, 데이터 분산을 추정할 필요가 없다  - 단점 : 모델을 생성하지 않으므로 설명하기 어려움, 분류 단계가 느림, k가 클 경우 많은 메모리가 필요하다  - 주의점 : k에 따라 모델의 정확성이나 효율성이 달라지므로 정확한 방법을 사용해야 한다  <1> 최적 k 탐색 : 일반적으로 1~20까지 탐색한다  <2> 가중치 : 단순 k만이 아니라 거리도 고려해서 분류방법을 사용한다  ② 판별분석  - 객체를 몇 개의 범주로 분류하되, 범주들을 가장 잘 구분하는 변수와 차이를 가장 잘 표현하는 함수를 도출한다  <1> 선형판별분석(LDA) : 정규분포의 분산-공분산 행렬이 범주에 관계없이 동일한 경우 사용  적은 파라미터를 사용하므로 빠르지만, 유연성이 낮다  <2> 이차판별분석(QDA) : 정규분포의 분산-공분산 행렬이 범주별로 다른 경우 사용  많은 파라미터를 사용하므로 느리지만, 유연성이 높다  이차판별분석을 사용하기 위해서는 모집단 등분산 검정이 유의해야 한다(biotools)  **참고 문헌**  - 데이터 기획  <https://bigdata-world.tistory.com/20>  <https://ymelissa.tistory.com/3>  <http://www.moonrepeat.org/wiki/doku.php?id=%EC%82%B0%EC%88%A0%ED%8F%89%EA%B7%A0_%EC%A4%91%EC%95%99%EA%B0%92_%EC%B5%9C%EB%B9%88%EA%B0%92%EC%9D%98_%EA%B4%80%EA%B3%84>  - 러닝패킷 강의  <http://www.kocw.net/home/search/kemView.do?kemId=694004>  - 빅데이터 분석과 R프로그래밍  <https://pabi.smartlearn.kr/courses/course-v1:POSTECH+DSB101+P902/course/>  - 머신러닝 기법과 R프로그래밍  <https://smartlearn.kr/courses/course-v1:POSTECH+DSB103+P902/course/> | | | |
|
|
|
|
|
|
|

**학습 정리**

|  |  |  |  |
| --- | --- | --- | --- |
| 팀 | 준비데이터 | 구성원 | 김호영, 안상현 |

|  |  |  |  |
| --- | --- | --- | --- |
| 일정 | 발제자 | 주제 | |
| 5/29 | 김호영 | 1. 데이터 전처리 방법  2. R의 데이터 타입과 구조 | |
| 주요 내용 요약 | | | |
| **Ⅰ오늘의 ADsP 요약**   1. 데이터 전처리   - 정의 : 일반적으로 데이터 분석 이전에 실시하는 데이터 확인 및 수정 단계  - 필요성 : 데이터 분석 중 업무 시간의 80% 정도를 데이터 수집/전처리에 사용함,  이 단계를 생략할 경우, 데이터 분석 결과가 잘못 나올 수 있어 세심한 주의가 필  - 종류  ① 데이터 셋 확인 : 독립변수/종속변수 정의, 각 변수 유형, 데이터 타입 확인  ② 변수 분석 : 히스토그램, 상자그림, 산점도를 통해 변수별 데이터 분포 확인  ③ 결측값 처리 : 삭제, 대체(평균, 최빈값, 중간값), 예측값 삽입(회귀분석)  ④ 이상값 처리 : 삭제, 대체, 변수화(이상값의 속성을 별개의 변수로 처리), 리샘플링, 케이스 분리  ⑤ Feature Engineering : Scaling(단위 변환), Binning/Dummy(연속형을 범주형으로 묶음), Transform(여러 변수를 묶어서 새로운 변수를 만듦)   1. R의 데이터 타입   ① 정수(integer) - 소수점이 없는 숫자(numeric) ex) 12, -29  ② 실수(double) - 소수점이 포함된 숫자(numeric) ex) 37.3, -27.4  ③ 문자형(character) - 하나, 여러 개의 문자로 이루어진 문자열 ex) “bigdata”, “A”  ④ 논리형(logical) - 참(TRUE, T), 거짓(FALSE, F)  ⑤ 기타 - 결측치(NA), 비숫자(NaN), 무한대(Inf, -Inf)   1. R의 데이터 구조   ① 스칼라(scala) : 1개의 데이터만 넣을 수 있는 객체 ex) a <- 5  ② 벡터(vector) : 1개의 데이터 유형(1차원 객체) ex) b <- c( )  ③ 행렬(matrix) : 1개의 데이터 유형(2차원 객체) ex) d <- matrix( )  matrix(데이터, ncol=숫자, nrow=숫자, byrow=T)  ④ 배열(array) : 1개의 데이터 유형(3차원 객체 이상) ex) e <- array( )  array(c(데이터), dim=c(숫자))  ⑤ 데이터프레임(data frame) : 여러 개의 데이터 유형(2차원 객체)  cf) 각 컬럼별로는 같은 데이터 타입이어야 한다  ex) f <- data.frame( )  ⑥ 리스트(list) : 여러 개의 데이터 유형, 여러 개의 객체 유형, 다차원 객체 구조  ex) g <- list( )  ⑦ 팩터(factor) : 1개의 데이터 유형 + 1차원 객체 구조  명목변수와 같은 형태로, 팩터 간에는 비교나 순서 없이 항목 구분 용도로만 사용됨 ex) h <- factor( )  **Ⅱ 오늘의 강의 요약**   1. **러닝패킷 : 통계의 기초 3강 - 공통**   ① 확률 : 어떤 사건이 일어날 가능성을 숫자로 표현한 것(0~1 사이)  ② 조건부 확률 : 특정 사건이 일어난 경우에서 다른 사건이 일어날 확률    ③ 확률변수 : 특정 값을 얻을 때의 확률이 정해져 있는 변수(X), 랜덤한 현상의 결과로 결정되는 수치적인 양  ④ 이산확률변수 : 확률변수가 구간 내 특정한 값만 취하는 경우  ⑤ 연속확률변수 : 확률변수가 구간 내 취할 수 있는 값의 범위가 무한한 경우  ⑥ 확률분포 : 확률변수 X의 값과 확률 p의 대응관계(값과 확률을 동시에 표현한 것)  - 확률분포의 기대값 :  - 확률분포의 분산 :  - 확률분포의 표준편차 :   1. **빅데이터 분석과 R프로그래밍 3강 - 김호영** 2. 데이터 불러들이기    1. csv파일 불러들이기 (read.csv)    2. .xls 파일       1. .xls 파일의 경우 데이터를 csv로 저장후, read.csv 함수를 이용하여 불러온다.    3. txt 파일 (read.table) 3. 현재 프로그램 작업폴더 (setwd)   → setwd("폴더주소")   1. attach    1. 현재 세션에서 나오는 변수들은 그 '데이터'의 변수로 인식한다는 의미    2. 데이터이름을 따로 지정하지 않아도 됨 2. detach    1. attach를 풀어줌 3. 데이터 추출(subset) 4. 요약통계치(aggregate) 5. csv로 내보내기 (write,table, write.csv) 6. 변수 특성변경    1. as.numeric    2. as.integer    3. factor 7. str(데이터 이름)    1. 데이터의 전체 구조 파악하기 8. summary(데이터이름)    1. 데이터 요약하기 9. table(데이터이름) 10. apply(변수리스트, (1=row,2=col), FUNC)     1. 몇개의 변수들의 요약통계치를 한번에 구하기 11. barplot(count)     1. 막대그래프 12. hist(변수이름, main="제목")     1. 히스토그램 13. scatterplot3d(변수이름, ..., main="제목")     1. 3D 산점도     2. packages 설치를 해야한다. (scatterplot3d) 14. lapply(변수리스트, FUNC)     1. 벡터화 요약치 15. readxl패키지     1. Excel 파일에 여러 worksheet가 있을때 사용     2. read\_excel을 통해 읽을수 있다.        1. 이때 전체 path를 적어야한다. 16. foreign 패키지     1. SPASS, SAS, ODBC 데이터 불러올 수 있음 17. sas7dat 패키지     1. SAS데이터 불러오기 가능 18. RODBC 패키지     1. SQL 데이터베이스와 연동하는 ODBC(open database connectivity) 기반의 패키지     2. SQL(Structured Query Language)서버에서 R로 데이터 불러오기 가능 19. **머신러닝기법과 R프로그래밍 3강 - 안상현**   - 내일 적을 예정(svm)  **참고 문헌**  - 데이터 전처리  <http://www.dodomira.com/2016/10/20/how_to_eda/>  <https://wikidocs.net/16582>  - 러닝패킷 강의  <http://www.kocw.net/home/search/kemView.do?kemId=694004>  나가노 히로유키, 「통계가 빨라지는 수학력」, 비전코리아, 2016.  이종익, 박민석 편저, 「사회조사분석사 2급 필기」, 시대고시기획, 2014.  - 빅데이터 분석과 R프로그래밍  <https://pabi.smartlearn.kr/courses/course-v1:POSTECH+DSB101+P902/course/>  - 머신러닝 기법과 R프로그래밍  <https://smartlearn.kr/courses/course-v1:POSTECH+DSB103+P902/course/> | | | |
|
|
|
|
|
|
|