29p 머신러닝
: 과거의 데이터를 통해 현재 혹은 미래를 예측하는 것
 - 과거의 데이터를 학습 데이터로 해서 예측 모양을 만드는 것
 - 과거의 패턴을 가지고 미래의 패턴이 어떻게 나올지를 예측한다.
ex. 과일 중에 바나나 구별하는 모델 - classification

일기 예보 - 연속된 값, - numerical( 뉴메리칼 value ) - regression

30p 이런 식으로 하는 것을 binary classification ( 0, 1로 분류하는 것 )

40p Decision tree

https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=laonple&logNo=220850892431

 - Decision Tree는 나무 모양의 그래프를 사용해 최적의 결정을 할 수 있게 하는 알고리즘
 - 어떤 항목에 대한 관측값(observation)에 대하여 branch 끝에 위치하는 기댓값(target)과 연결시켜 주는 predictive model이다.
 - 직관적이고, 새로운 데이터에 대해 쉽게 판단할 수 있다.
 - 속성이 여러개가 있는 경우, 어떤 속성을 root node에 두는지가 중요하다.
 - compact하게 만드는 방법 - 엔트로피
 - 엔트로피 : 데이터의 분포의 순도를 나타내는 척도
 - 데이터의 순도가 높다 : 엔트로피의 값이 낮다
 - 데이터의 순도가 낮다 : 엔트로피의 값이 크다

////아래는 뭔가 이상
엔트로피가 높다 - 정보를 많이 가지고 있다 -
엔트로피가 낮다 - 정보를 적게 가지고 있다 -

* 시계열 예측 분석
- 거시 쪽에서느 몇십년..된 분야
 - 페이스북에서 prophet이라는 모델을 만들어 뒀다.

71p
시계열 데이터에서, 학습 모델에서 작년 데이터는 버려야 한다.
작년 데이터를 버려야 하는 이유는? 예상하지 못한 코로나 상황이 발생했기 때문.
이 예측하는 과정에 EDA 이런 것들이 들어간다.

* Exploratory Data Analysis : 탐색적 데이터 분석

https://untitled-memo-2019.tistory.com/1

- 데이터를 수집했을 때, 다양한 각도에서 관찰하고 이해하는 과정
- 데이터 분석 전에 그래프 및 통계적 방법으로 자료를 직관적으로 보는 과정
- 데이터 분포 및 값을 검토하여 수집한 데이터가 어떤 것인지를 나타내는지 더 잘 이해할 수 있다.
- 자세한 방법은 따로 공부

데이터셋을 만들어야 한다. - 13p 정리

card 데이터에서
default-par 맨 끝 데이터를 종속변수, target,
cost, loss : 실젯값과 목표값의 차이
앞의 데이터들은 독립 변수, input parameter
독립변수 - 종속변수 관계는 인과관계
독립변수들끼리의 관계 -

PCA :공분산을 통해 아이겐벨류, 아이겐벡터,를 통해서 차원 축소를 하는 것

노랑 책 77p

t테스트?
아노바? - 아웃풋 밸류가 카테고리가 3개 이상이다?

covariance
- 차원이 많아지면서 공간에서의 데이터의 분포를 확인할 수 있는 것이 평균, 분산, covariance

이런 아노바 등의 상관계수를 어디에 배우려고 쓰는지?
차원의 저주에 이용하려고! 차원을 줄이려고.
차원을 줄이려는 이유는, 같은 데이터셋인데 차원이 많으면 밀도가 낮아져서, 좋은 모델이 나오기 힘들다.
어떻게?
복잡한 20차원 되는 것을 깔끔하게 줄인다. ( correlation 그림 )
그래서 PCA!

* PCA.pdf 확인
covariance metrix
Linear transformations
PCA.pdf linear transformations 부분에서 보면, 초록 피쳐는 없애고, 빨강 피쳐만 남는다는 것

* 표준화
standardization : 스케일은 평균 0, 표준편차 1로 다 맞추는 것
딥러닝 레이어에 들어가려면, input 데이터가 다 표준화가 되어야 한다.
skilearn을 통해서 표준화를 한번에 싹 다 시킬 수가 있다.
 - 딥러닝은 반.드.시 standardization을 해야 한다.

딥러닝은 feature exception까지 다 해버린다/
전통적인 머신러닝은 feature exception을 우리가 해줘야 한다.

? 트리 알고리즘은 데이터가 섞이질 않는다..

feature? 이거는 pca, rfe, 같은 것들로 한다.

classfication을 했으면 regression도 해야 한다.

xgboost? - 앙상블 기법?
https://lsjsj92.tistory.com/547

< 정리! >
전처리 과정에 대한 기본적인 이해!
regression, classification의 차이!
PCA, rfe,
( 카일스퀘어 검증? 이거는 내일! )

* 머신러닝의 표준화, 정규화
nomalization( 정규화 ) - 값을 모두 [0,1] 범위로 스케일링 하는 것

- standardization - 표준화도 스케일링 하는 건데.. 방법이 다르다.

https://bskyvision.com/849

nomalization : (0,1) 구간으로 다 맞춘다.

정규화는 이상값을 억제하기 위해 사용
표준화 : 전체에서 몇 퍼센트가 이런가? 이런 지표를 보기 위해서는 표준화를 더 사용.

Eigenvalue

PCA
- weight
- 전진, 후진은 전진 :

데이터 전처리

https://rk1993.tistory.com/entry/%EB%8D%B0%EC%9D%B4%ED%84%B0-%EC%A0%84%EC%B2%98%EB%A6%AC

* Feature Selection

https://junklee.tistory.com/8

<-> 차원 축소 (Dimesionality Reduction)와는 다르다
why? 차원 축소(PCA, SVD)는 새로운 특성 조합을 생성한다.
- 특성 선택은 값을, 특성을 변경하지 않고 포함 및 제외하는 방법이다.

4. 반복적 특성 선택
- 모든 특성 조합을 다 시도하고, 가장 좋은 set을 찾겠다는 의미 ( 가장 단순한 방법 )
- 1개의 feature에서 출발해 특성 개수에 도달할 때 까지 특성을 추가하는 방법
- 모든 특성에서 출발해 지정한 특성 개수에 도달할 때까지 특성을 제거하는 방법 ( RFE )

* RFE
https://ratsgo.github.io/machine%20learning/2017/04/02/logistic/
logistic regression

linear regression
- 변수 x의 값은 독립적으로 변할 수 있는 변수, 독립변수
- 변수 x에 따라 종속적으로 변하는 변수 = 종속변수
- 독립변수가 1개인 경우 단순 선형 회귀
- 여러 개인 경우 multiple linear regression Analysis

simple linear regression인 경우
- y = Wx + b, W= weight, b = bias(편향)
- 이렇게 해서 x와 y의 관계를 적절히 표현할 수 있다.

multiple linear regression의 경우
- y = W1x1 + W2x2 +... + Wn Xn+ b 이렇게 해서
- 여러개의 독립변수인 x1~xn과 bias의 b를 통해 식으로 나타낼 수 있다.

weight와 b를 찾는 방법은? - cost function
: 모든 오차의 제곱의 평균을 구하는 방법
https://wikidocs.net/22647
여기서부터 계속

https://wikidocs.net/22647
머신러닝 쪽 기본 내용에 대한 내용이 잘 나와 있다.