Ch2. End-to-End Machine Learning Project -Part1

김성봉

다루는 내용

가정 : 당신은 부동산 회사의 데이터 사이언티스트로 고용이 되었음

예제 Project를 처음부터 끝까지 아래와 같은 과정을 거쳐서 훑어볼 예정

1. Look at the big picture.
2. Get the data.
3. Discover and visualize the data to gain insights.
4. Prepare the data for Machine Learning algorithms.
5. Select a model and train it.
6. Fine-tune your model.
7. Present your solution.
8. Launch, monitor, and maintain your system.

Working with Real Data

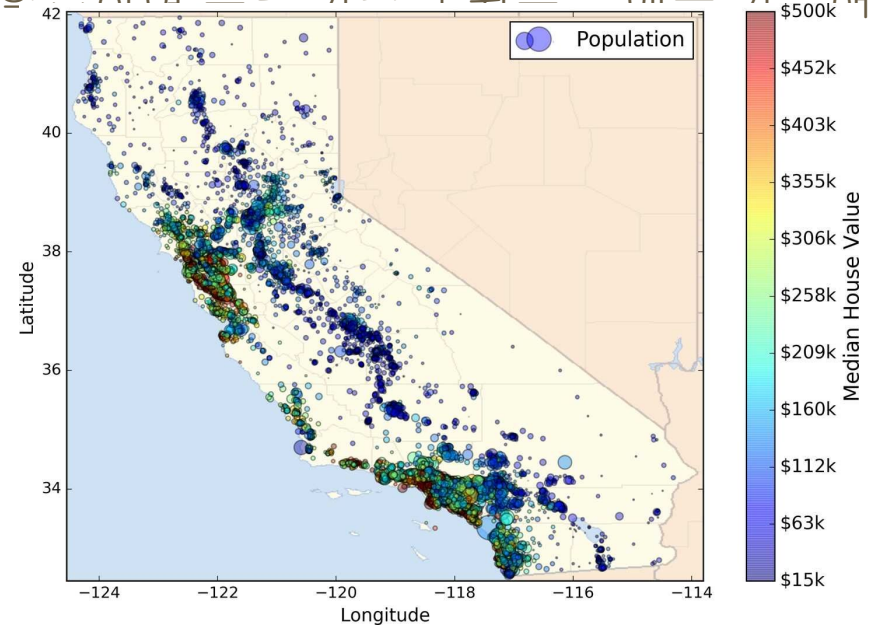
Machine Learning을 배울 때, 실제 데이터를 가지고 연습하는 것이 Best

다행히 수천개의 open dataset이 모든 분야에 걸쳐 존재함

- Popular open data repositories:
  - UC Irvine Machine Learning Repository
  - Kaggle datasets
  - Amazon's AWS datasets
- Meta Portals
  - http://dataportals.org/
  - http://opendatamonitor.eu/
  - http://quandl.com/
- Other pages
  - Wikipedia의 Machine Learning dataset 목록 (https://goo.gl/SJHN2k)
  - Quora.com의 질문 (http://goo.gl/zDR78y)
  - Reddit의 dataset 목록 (https://www.reddit.com/r/datasets)

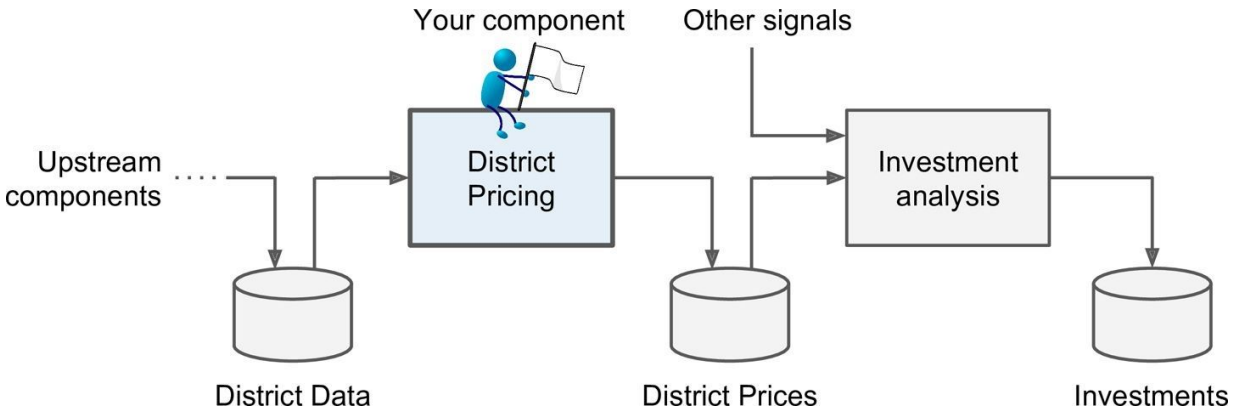
Look at the Big Picture

- 입사한 회사 : Machine Learning Housing Corporation
- 요청사항 : California 인구조사 데이터를 사용하여 주택 가격에 대한 모델 구축
- 데이터에 포함된 내용
  - 인구, 소득 중위수, 주택 가격 중위수 등이 block 단위로 제공됨
  - block : US Census Bureau에서 발간하는 데이터의 지정학적 최소 단위 (이하 “districts”로 표기)
- 목표 : 데이터를 학습하여 주택 가격 중위수를 예측 가능해야 함



Look at the Big Picture - Frame the Problem (1/2)

- 첫 번째로 질문해야 할 사항 : 비즈니스 목적은 무엇인가?
- 모델을 만드는 것보다 회사의 모델을 활용 방안이 문제 정의에서 중요
  - 알고리즘 선택
  - 모델 성능 평가 방법
  - 모델 수정에 들일 노력의 정도
- 개발한 모델(district의 평균 주택 가격 예측)은 다른 ML 시스템에 다른 signal과 함께 제공 예정임 - From Boss



Look at the Big Picture - Frame the Problem (2/2)

- 두 번째로 질문해야 할 사항 : 현재 솔루션은 무엇인가?
- 현재 솔루션 조사는 모델 성능에 참조 가능하며, 문제 해결 방법에 인사이트 제공
- 현재 솔루션
  - 전문가들에 의한 평균 가격 산정
  - 복잡한 룰을 사용해 산정 중
  - 비용과 시간 소모적인 방법
  - 산정 결과는 보통 15% 가량 오차가 존재함
- 정보를 조합하여 ML 시스템 디자인을 수행
  - 지도? 비지도? 강화? 학습 선택 ⇒ 지도학습 (with Labeled examples)
  - 분류? 회귀? 그 외? 수행 알고리즘 선택 ⇒ 회귀 (값을 예측) ※ 분류는 (범주를 예측)
  - 배치 학습? 실시간 학습? 선택 ⇒ 배치 학습 (지속적인 데이터 제공이 없으므로)

Look at the Big Picture - Select a Performance Measure (1/4)

- 모델 성능 평가를 위한 방법 선택이 필요

※ sklearn 제공 평가 방법 (http://scikit-learn.org/stable/modules/model_evaluation.html)

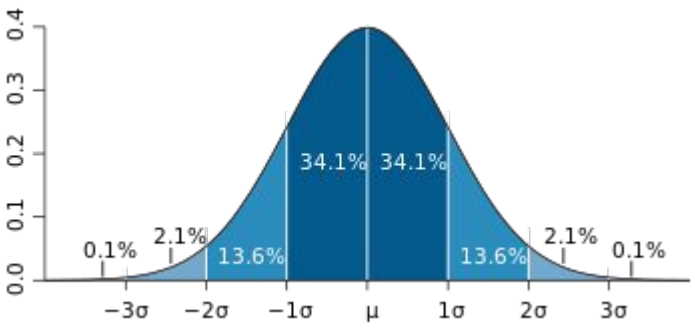
  - Mean Absolute Error (MAE)
  - Mean Squared Error / Root Mean Squared Error
  - Mean Squared Log Error / Root Mean Squared Log Error
  - Median Absolute Error (MAE?)

Look at the Big Picture - Select a Performance Measure (2/4)

- Root Mean Squared Error (RMSE) : 보통 회귀 문제의 평가 방법

$$\mathrm{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(\mathbf{x}^{(i)}) - y^{(i)})^2}$$

- Error (=예측값 - 실제값)의 표준편차를 측정
- 예시 : RMSE가 50,000일 경우,
  - 예측값 오차의 68%(±1σ)가 ± $50,000 안에 속함
  - 예측값 오차의 95%(±2σ)가 ± $100,000 안에 속함



Look at the Big Picture - Select a Performance Measure (3/4)

- Mean Absolute Error (MAE)

$$\text{MAE}(\mathbf{X}, h) = \frac{1}{m}\sum_{i=1}^{m} | h(\mathbf{x}^{(i)}) - y^{(i)} |$$

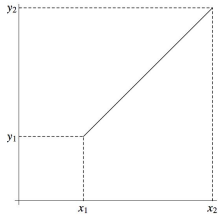
- Error (예측값 - 실제값)의 절대값을 측정
- Outlier가 많을 경우 고려할 수 있음

Look at the Big Picture - Select a Performance Measure (4/4)

- RMSE와 MAE 모두 두 벡터 사이의 거리를 측정하는 방법
  - RMSE : Euclidean norm, L2 norm

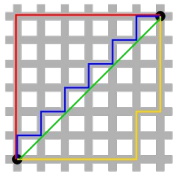
$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$



  - MAE : Manhattan norm, L1 norm

$$d_1(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_1 = \sum_{i=1}^{n}|p_i-q_i|$$

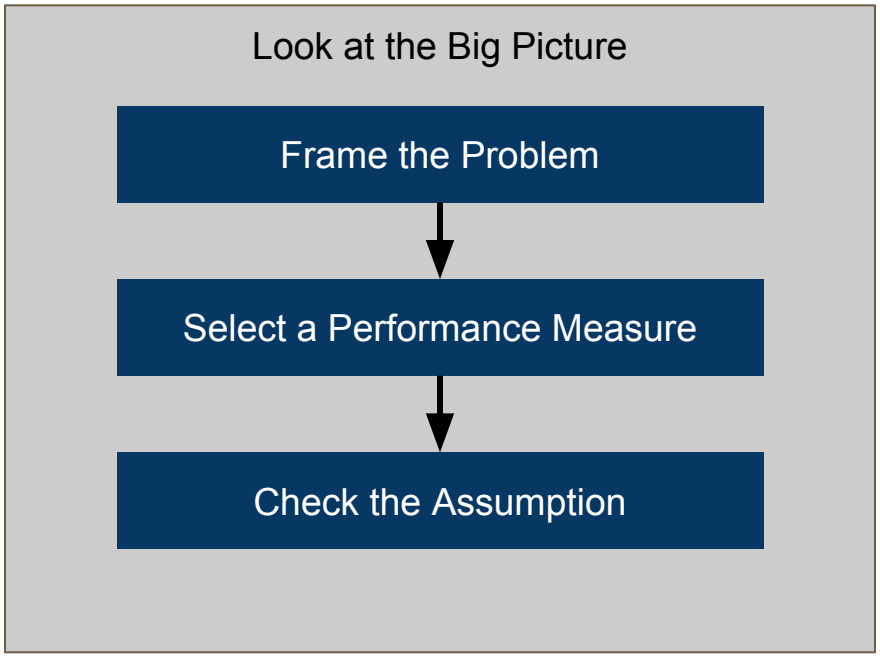


  - 일반화 하면 $\|\mathbf{v}\|_k = (|v_0|^k + |v_1|^k + \cdots + |v_n|^k)^{\frac{1}{k}}$, Lk norm
  - L0의 경우 벡터의 element 수를 반환
  - L∞의 경우 벡터중 최대 절대값을 반환
- 높은 차수의 norm의 경우 큰 값에 집중하고 작은 값은 무시하는 결과가 나옴
- 그래서 RMSE가 MAE보다 outlier에 민감하게 반응 → 산출 오차가 과하게 증가
- 종 모양의 분포일 경우는 RMSE가 더 선호됨

Look at the Big Picture - Check the Assumption

- 현재까지의 가정들의 목록을 작성하고 검토
  - 수행자 : 본인 또는 다른 사람 (누구든지)
  - 심각한 이슈를 미리 발견할 수 있음
- 예를 들어, 다음 프로세스에 전달할 값의 형태에 대한 가정 점검이 필요
- 현재 가정 : 평균 주택 가격 값을 전달
- 하지만 평균 주택 가격의 범주로 전달 받기를 원할 수도 있음 (저렴, 중간, 비쌈)
- 이 경우, 평균 주택 가격 값이 완벽히 맞아도 소용이 없으며, 분류 문제로 변경됨
- 다행히, 다음 프로세스에 확인 결과 현재의 가정이 맞음을 확인 완료

Look at the Big Picture - Summary



Get the Data

- 이후 내용은 jupyter notebook을 이용해 실습하며 진행
- 코드 : https://github.com/ageron/handson-ml
- 환경 설정 : 교재 및 jupyter notebook 파일 참조
- 데이터 다운로드 : 교재 및 jupyter notebook 파일 참조

Get the Data - Take a Quick Look at the Data Structure (1/2)

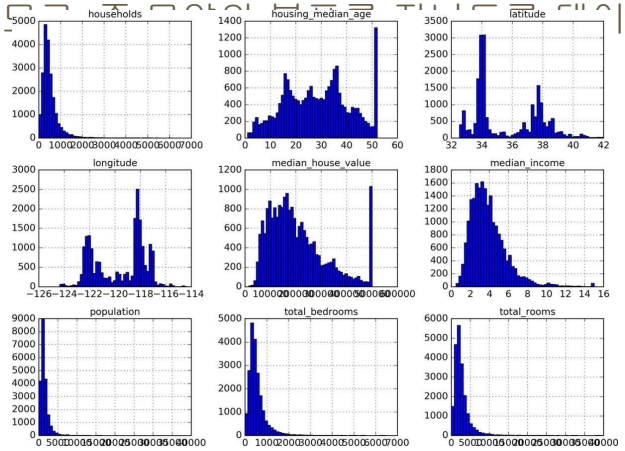
- housing 데이터 샘플 확인 : pandas DataFrame.head()
  - 10개 속성(attribute)이 존재
  - longitude : 경도 / latitude : 위도 / housing_median_age : 평균 주택 연식
  - total_rooms : 전체 방 수 / total_bedrooms : 전체 침실 수 / population : 인구
  - households : 가구 수 / median_income : 평균 수입 / median_house_value : 평균 주택 가격
  - ocean_proximity : 해안 접근성

- 간략한 데이터 설명 확인 : pandas DataFrame.info()
  - 데이터 수 : 20,640
  - total_bedroom에 207 district가 null, ocean_proximity는 범주형 변수임을 확인

- 범주형 속성이 보유한 개별 값들의 수 확인 : pandas DataFrame.value_counts()
- 숫자형 속성들의 기초 통계 요약 정보 확인 : pandas DataFrame.describe()

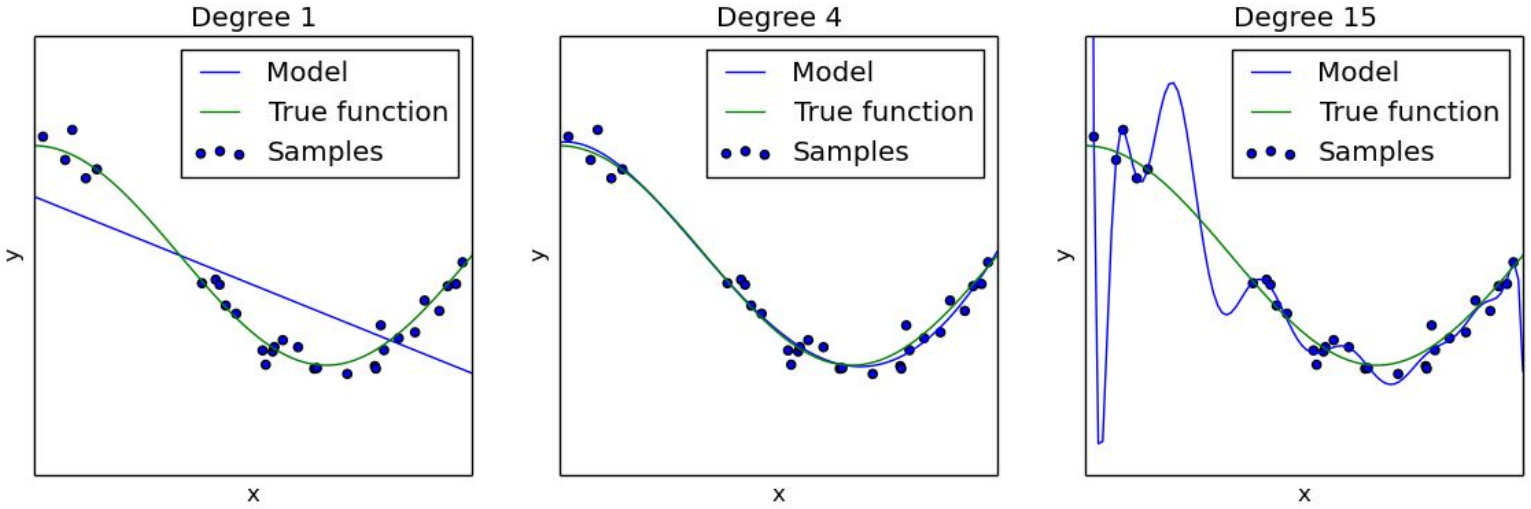
Get the Data - Take a Quick Look at the Data Structure (2/2)

- 히스토그램을 통한 데이터 분포 확인
  - 평균 수입 : US Dollars로 표현되지 않음, 0.5~15 사이의 값을 가지도록 scaling 되어 있음
  - 평균 주택 연식/가격 : scaling 되어 있음, 가격은 예측 값이므로 이후 프로세스 팀과 논의가 필요
  - 가격은 $50,000의 값을 가지도록 scaling 되어 있음
  - $50,000 이상의 값을 정확히 예측해야 할 경우 ① 조정된 district의 정상 데이터를 수집 또는 ② $50,000 이상 데이터를 제거
  - 모든 속성들은 다른 scale을 가짐
  - 많은 그래프들이 우측으로 꼬리가 긴 형태를 지니며 (Positive Skewness), 이는 ML 알고리즘이 패턴을 찾기 어렵게 하므로 [illegible] 데이터 변형을 시도 예정



Get the Data - Create a Test Set (1/5)

- 사람은 뇌는 훌륭한 패턴 감지 시스템이지만, overfitting하는 경향이 있음
- Test 데이터를 미리 본다면, 특정 패턴에 치우쳐 일반화된 모델을 만들 수 없음
- 따라서 Test용 데이터를 분리해 놓는 작업이 필요함



http://scikit-learn.org/0.15/auto_examples/plot_underfitting_overfitting.html

Get the Data - Create a Test Set (2/5)

- Test Set을 만드는 방법 : 임의의 instance(데이터)를 선택하여 분리 (보통 20%)
- 프로그램으로 구현 (Random Sampling)
  - 데이터 수만큼 정수 순열을 랜덤하게 생성하여 index 번호로 사용
  - [1, 15, 4, 3, ...] → 1번째, 15번째, 4번째, 3번째 데이터 순으로 재정렬하는 효과
  - 생성한 index 번호 개수의 20% 개수에 해당하는 데이터를 선택하여 분리
- 문제점 및 대안
  - 프로그램 실행 시마다 다른 랜덤 숫자 생성 ⇒ seed 지정, np.random.seed(42)/random_state 등
  - 데이터 추가/업데이트 시 여전히 문제
  - 특정 속성(컬럼)값의 hash 끝자리를 기준으로 20%를 선택 (끝자리는 16*16=256가지 값을 가짐?)
  - 또는 속성(컬럼)값의 조합으로 안정적인 ID컬럼을 생성 후 그 값을 기준으로 20%를 선택
- Scikit-learn 모듈(sklearn)에서는 train_test_split이라는 함수를 제공

hash 설명 참조 : http://bcho.tistory.com/1072

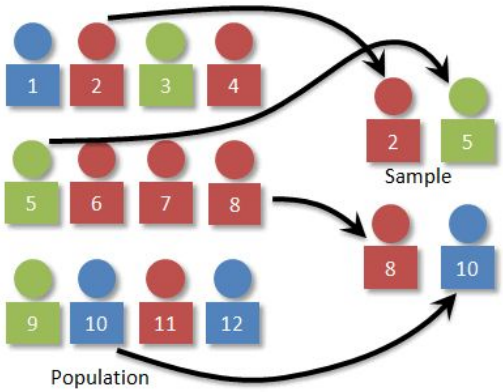
Get the Data - Create a Test Set (3/5)

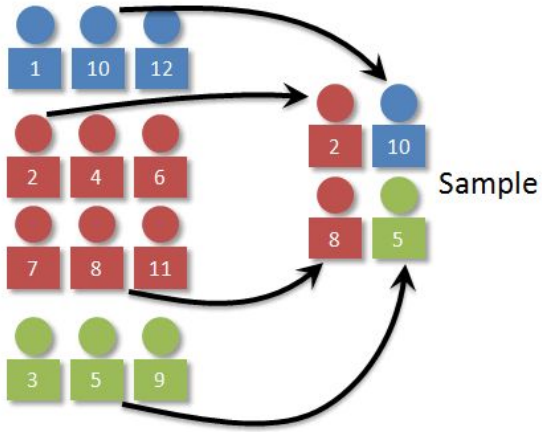
- dataset이 충분히 클 경우(특히 속성), 일반적으로 Random Sampling은 좋으나, 아닐 경우, 샘플링된 데이터 집단이 전체 모수를 대변하지 못할 수 있음
- 1000명 설문조사 기획 시, 인구 비중이 여자 51.3%, 남자 48.7%라면 여자 513명, 남자 487명 샘플링 하는 것이 잘 설계된 설문조사
- Stratified Sampling : 카테고리 별, 비중을 유지한 샘플링 방식
  - strata : 하위그룹, stratum : strata에 속한 데이터

- 이해가 잘 안되는 부분 : 일반 random sampling 시, 12%의 확률로 skewed test set을 샘플링할 가능성이 있다고 하는데 사유를 모르겠음
- (49% 미만의 여자 비중 또는 54% 초과의 여자 비중을 얻을 가능성)

Get the Data - Create a Test Set (4/5)

- 샘플링 방법
  - Simple Random Sampling : 완전 랜덤한 추출
  - Stratified Sampling : 모수를 계층화(군집?)하고, 계층별로 비중에 비례한 샘플 수를 랜덤하게 추출
  - Systematic Sampling : 모수에서 매 k번째 데이터를 추출, (k=3일때, 1/4/7/10/...번째 데이터 추출)





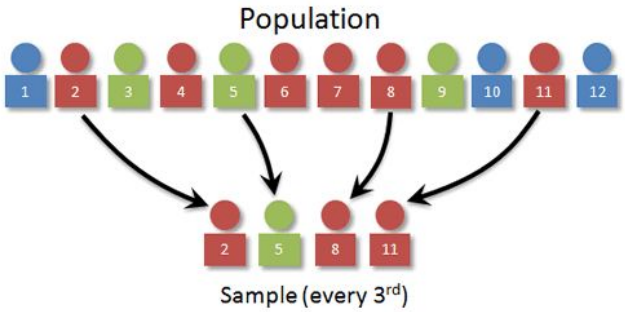
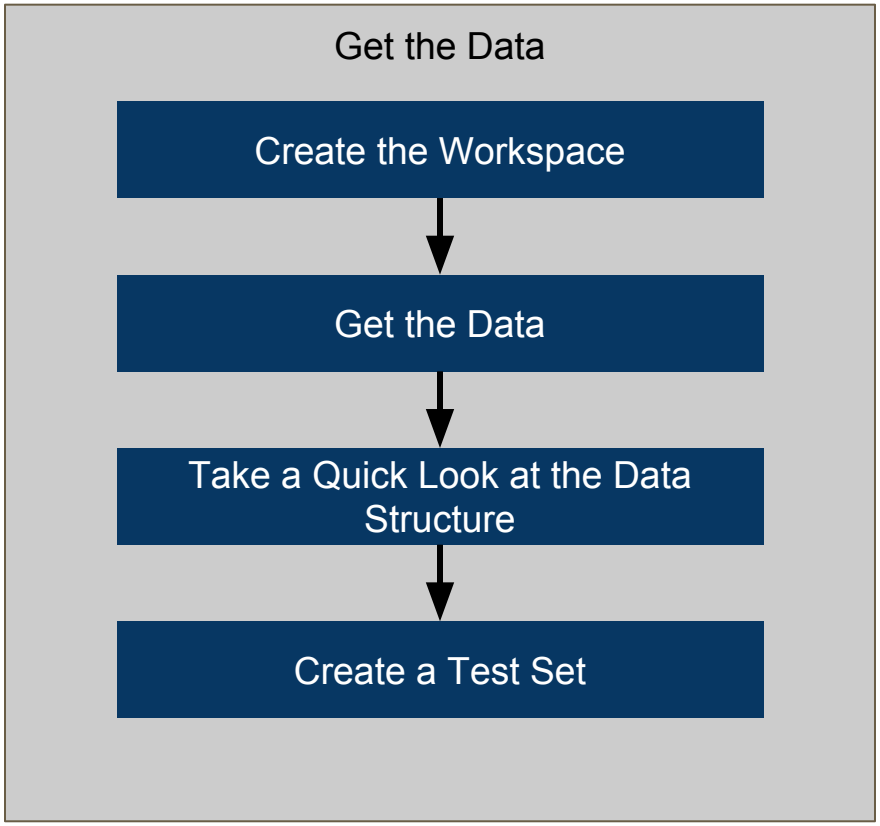


그림 참조 : https://faculty.elgin.edu/dkernler/statistics/ch01/1-4.html

Get the Data - Create a Test Set (5/5)

- 주택 가격 예측 문제에서 평균 소득은 예측에 중요한 인자임
- 전체 모수의 다양한 평균 소득 범주를 Test set에 반영하고자 함
  - 평균 소득은 실수형 데이터이므로, 범주형으로 변환 필요
  - 평균 소득 분포는 대부분 2~5만 달러에 분포하지만, 6만달러 이상도 존재
  - strata에 포함되는 데이터 수의 불균형이 있으면, 데이터 중요도가 편향 가능
  - 균형잡힌 stratum 구축을 위해 5만 달러 이상은 5만달러로 통일
  - 그 외 소득은 1.5로 나누고 올림 처리하여 범주화 수행 ⇒ 1, 2, 3, 4, 5의 소득 범주 생성
- 소득 범주별로 모수/Random Sampling/Stratified Sampling에서의 비중 비교
  - Random Sampling : 5개 범주 평균 절대 오차 약 2.87%
  - Stratified Sampling : 5개 범주 평균 절대 오차 약 0.07%
  - Stratified Sampling이 모수의 범주 비중을 잘 반영함을 알 수 있음

Get the Data - Summary

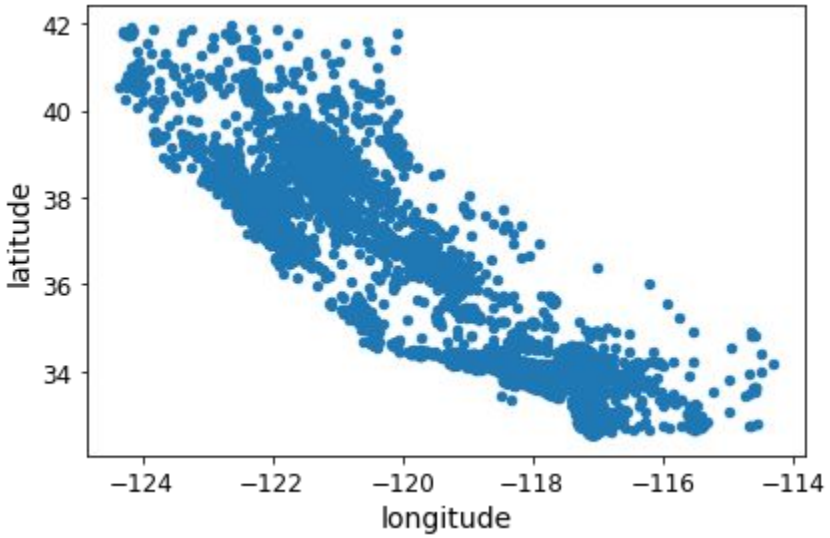


Discover and Visualize the Data to Gain Insights

- 이전 단계까지는 데이터에 대해 일반적인 내용을 알기 위해 살펴봄
- 이 단계에서는 조금 더 깊은 이해를 얻고자 함
- 이 단계에서부터 Training set만 가지고 탐색을 시작
- Training set이 매우 크다면 탐색을 위한 샘플 set을 만들어서 사용
- 이 단계의 주요 내용은 시각화 및 변수간 상관관계 분석을 통한 인사이트 도출

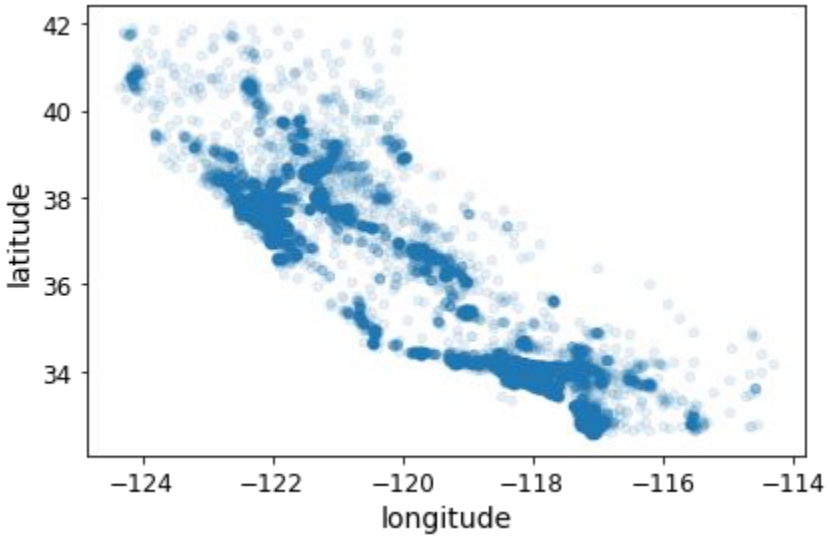
Discover and Visualize the Data to Gain Insights - Visualizing Geographical Data (1/3)

- x축 경도(longitude), y축 위도(latitude)로 산점도를 그려 탐색
- 점이 밀집되어 있어서 특별한 패턴이 보이지 않음



Discover and Visualize the Data to Gain Insights - Visualizing Geographical Data (2/3)

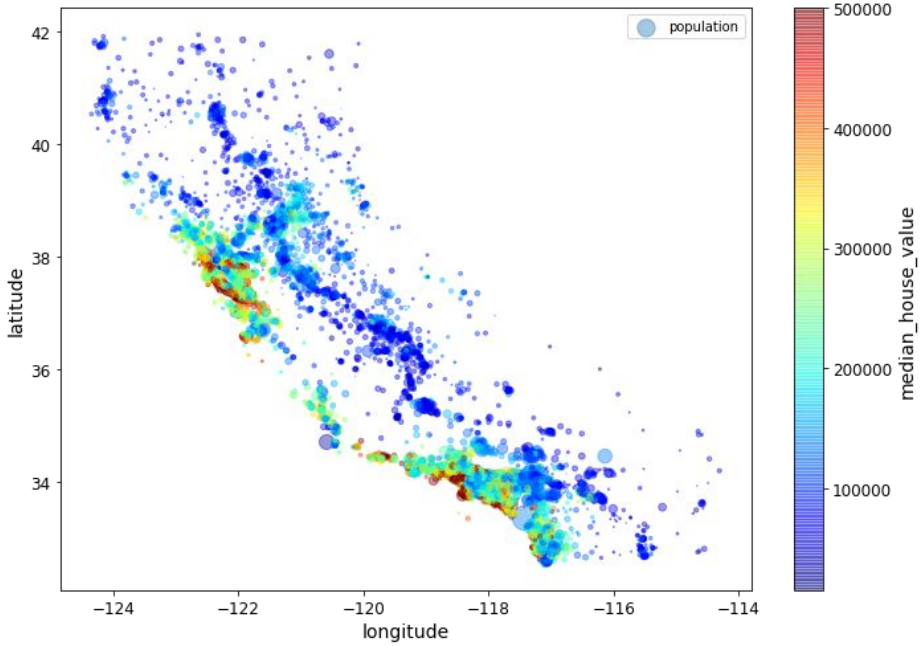
- 산점도에 투명도(alpha 옵션)을 추가하여 탐색
- 점이 밀집된 지점이 명확히 눈에 들어옴
- Bay Area, Los Angeles, San Diego, Central Valley, Sacramento, Fresno

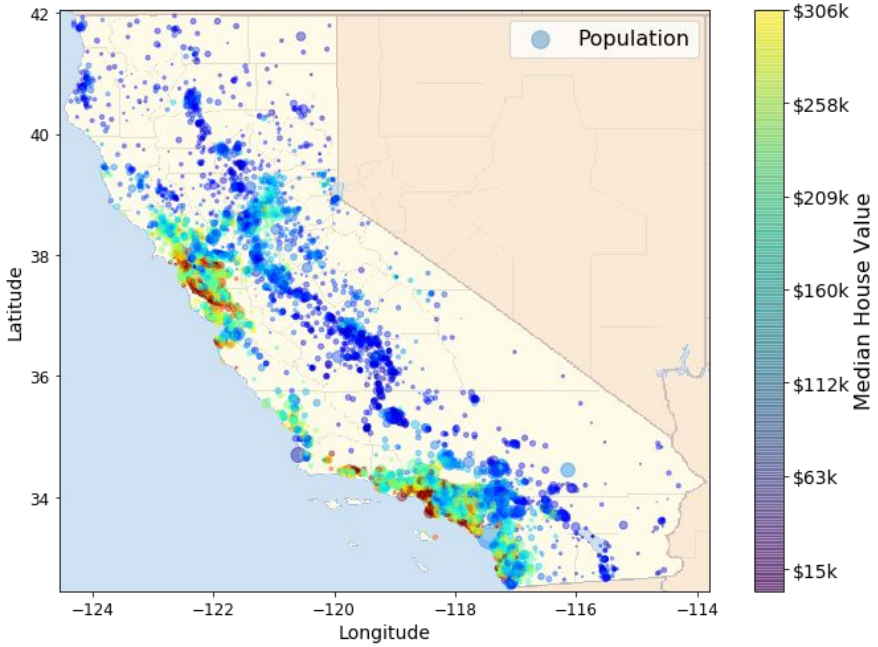


Discover and Visualize the Data to Gain Insights -

Visualizing Geographical Data (3/3)

- 산점도 + 점 크기(인구 수) + 점 색(평균 주택 가격)으로 표현
- 지도 이미지 결합하여 시각화
- 해안가에 인접할 수록, 인구가 많을 수록 주택 가격이 높은 경향

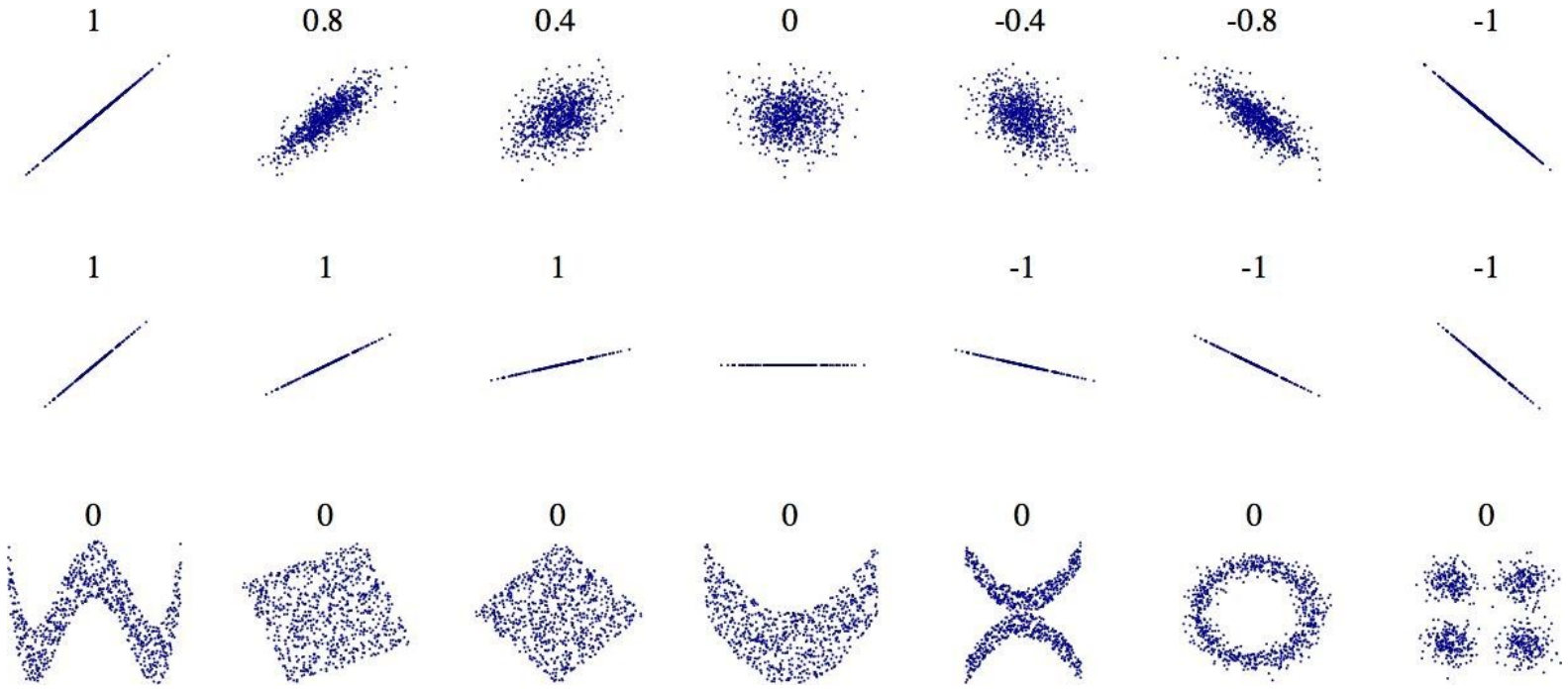




Discover and Visualize the Data to Gain Insights -

Looking for Correlations (1/2)

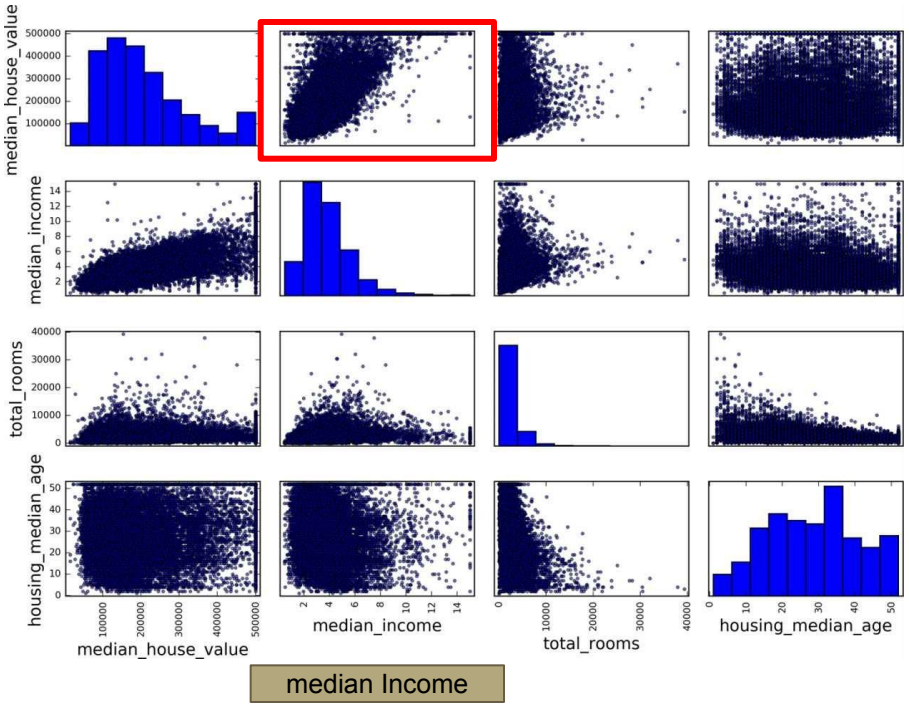
- 상관 분석 : 두 변수간에 어떤 선형적 관계를 갖고 있는 지를 분석하는 방법
- 상관 계수 : 두 변수간 관계의 강도를 나타내는 수치, -1 ~ 1 사이의 값을 가짐



Discover and Visualize the Data to Gain Insights - Looking for Correlations (2/2)

- housing 데이터에서 예측하려는 평균 주택 가격과의 상관 계수 탐색
- pandas DataFrame의 corr() 메소드를 사용
  - 평균 소득 0.68, 총 방 개수 0.13, 평균 주택 연식 0.11 순으로 형성
- scatter_matrix() 함수를 사용 (모든 변수의 시각화가 어려우므로 4개만 선택 수행)
  - 시각적으로 평

median house value



median Income

Discover and Visualize the Data to Gain Insights - Experimenting with Attribute Combinations

- 현재까지 데이터에 대한 탐색 및 인사이트를 도출함
  - 몇몇 데이터는 ML에 데이터 적용 전 Cleaning이 필요
  - 몇몇 속성들은 상관성을 갖고 있음
  - 몇몇 속성들은 꼬리가 길게 형성된 분포를 가지고 있어서 변형이 필요함을 발견
- ML을 위한 데이터 준비 과정의 마지막으로 다양한 속성들의 조합을 시도할 수 있음
  - 가구 당 방 개수 : total_rooms / households
  - 방 개수 중 침실의 비중 : total_bedrooms / total_rooms
  - 가구 당 인구 수 : population / households
  - 기타 등등
- 위의 추가된 속성을 포함한 상관관계 분석 결과
  - 가구 당 방 개수 : 0.19
  - 가구 당 인구 수 : -0.02
  - 방 개수 중 침실의 비중 : -0.26
- 가구 당 방 개수가 총 방 개수보다 상관계수가 높으며, 방 개수 중 침실의 비중은 음의 상관관계를 가짐을 알 수 있음

Discover and Visualize the Data to Gain Insights - Summary