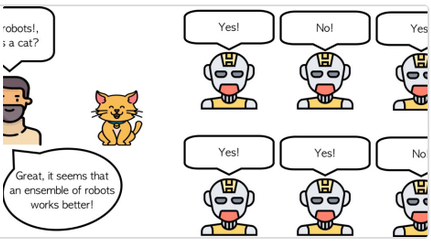
7 앙상블 학습

Ensemble Learning: Stacking, Blending & Voting

Great, now that you're familiar with the Blending architecture, let's see how we do this in code: Let's analyze the key parts of this model. In line 4 we are defining the 5 base classifiers that we will

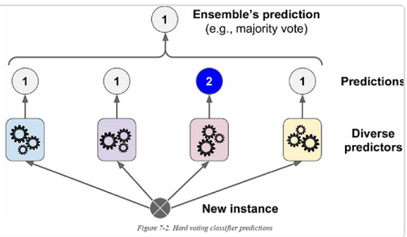
tds https://towardsdatascience.com/ensemble-learning-stacking-blending-voting-b37737c4f483



[Python] Voting Classifiers(다수결 분류)의 정의와 구현

Voting Classifiers Voting Classifiers는 "다수결 분류"를 뜻하는 것으로, 두 가지 방법으로 분류할 수 있습니다. 1. Hard Voting Classifier 여러 모델을 생성하고 그 성과(결과)를 비교합니다. 이 때 classifie..

https://nonmeyet.tistory.com/entry/Python-Voting-Classifiers%EB%8B%A4%EC%88%98%EA%B2%B0-%EB%B6%84%EB%A5%98%EC%9D%98-%EC%A0%95%EC%9D%98%EC%99%80-%EA%B5%AC%ED%98%84



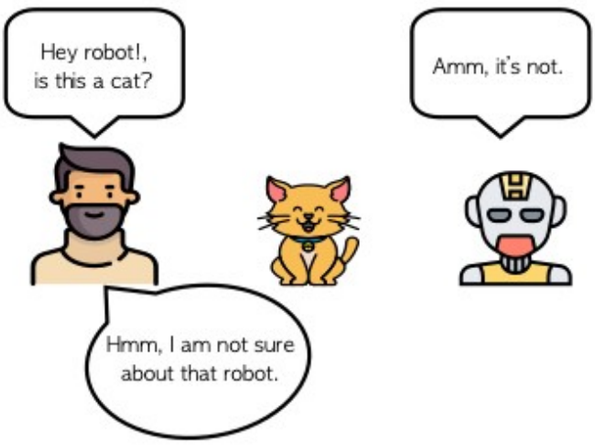
Ensemble methods: bagging, boosting and stacking

This post was co-written with Baptiste Rocca . "Unity is strength". This old saying expresses pretty well the underlying idea that rules the very powerful "ensemble methods" in machine learning.

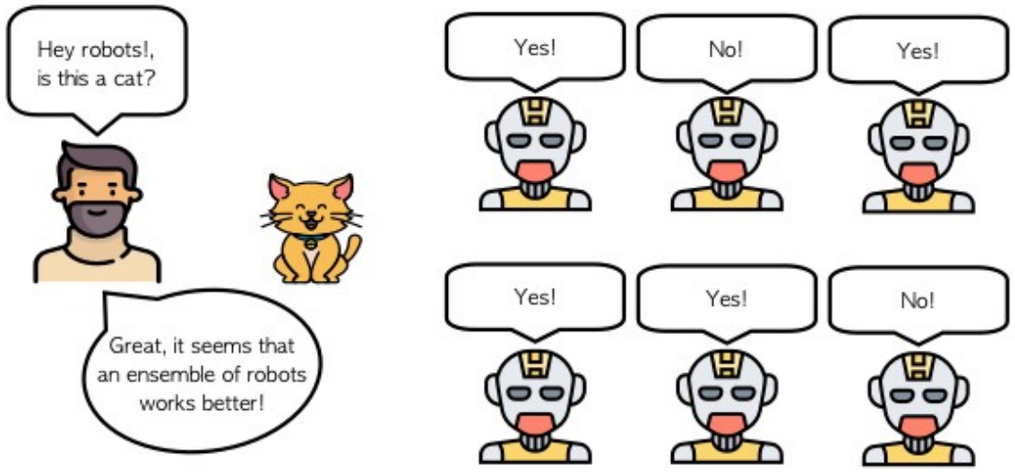
tds https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Ensemble

- 서로다른 객체 학습기들로 구성
- 일련의 예측기(즉, 분류나 회귀 모델)로부터 예측을 수집하면 가장 좋은 모델 하나보다 더 좋은 예측을 얻을 수 있다.



a) Traditional Learning



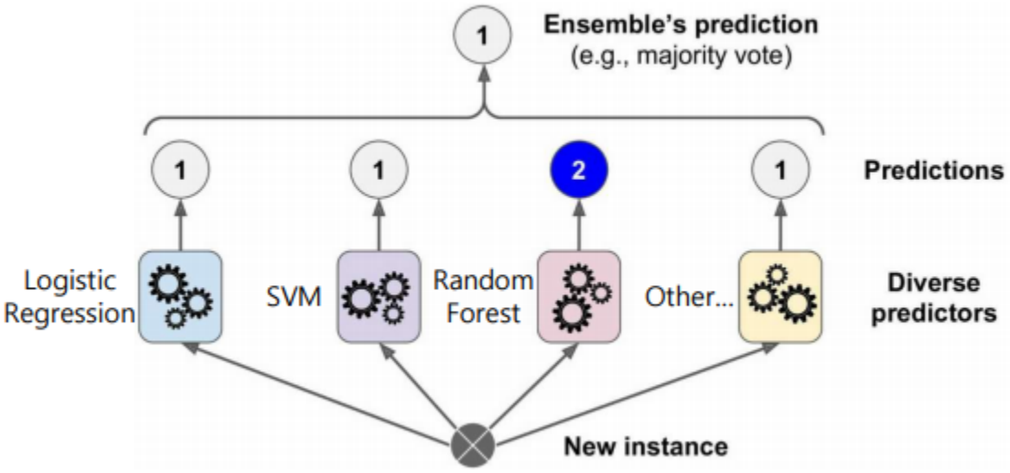
b) Ensemble Learning

동질적 객체 학습기

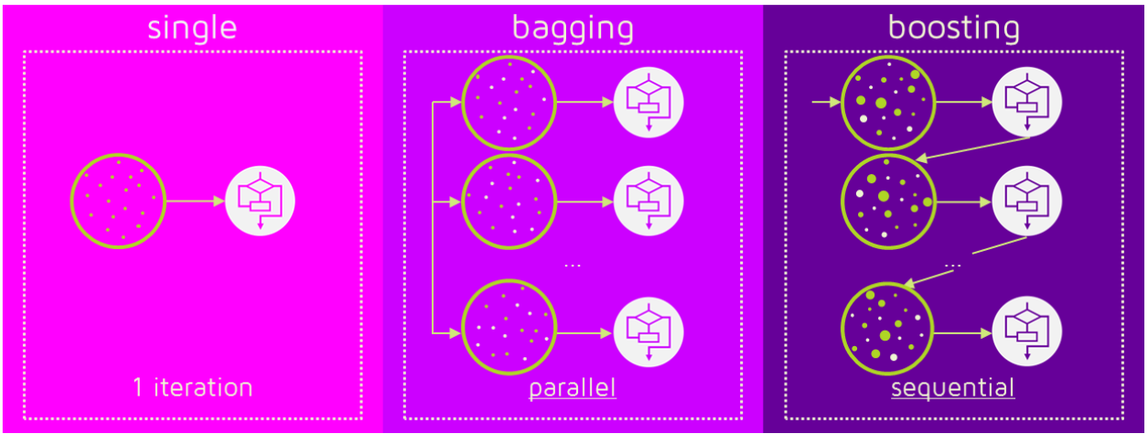
- 같은 객체 학습기를 사용하는 방법
- Random Forest

이질적 객체 학습기

- 서로 다른 객체 학습기를 사용는 방법
- Forst + NurealNetwork



객체 학습기 사이의 관계



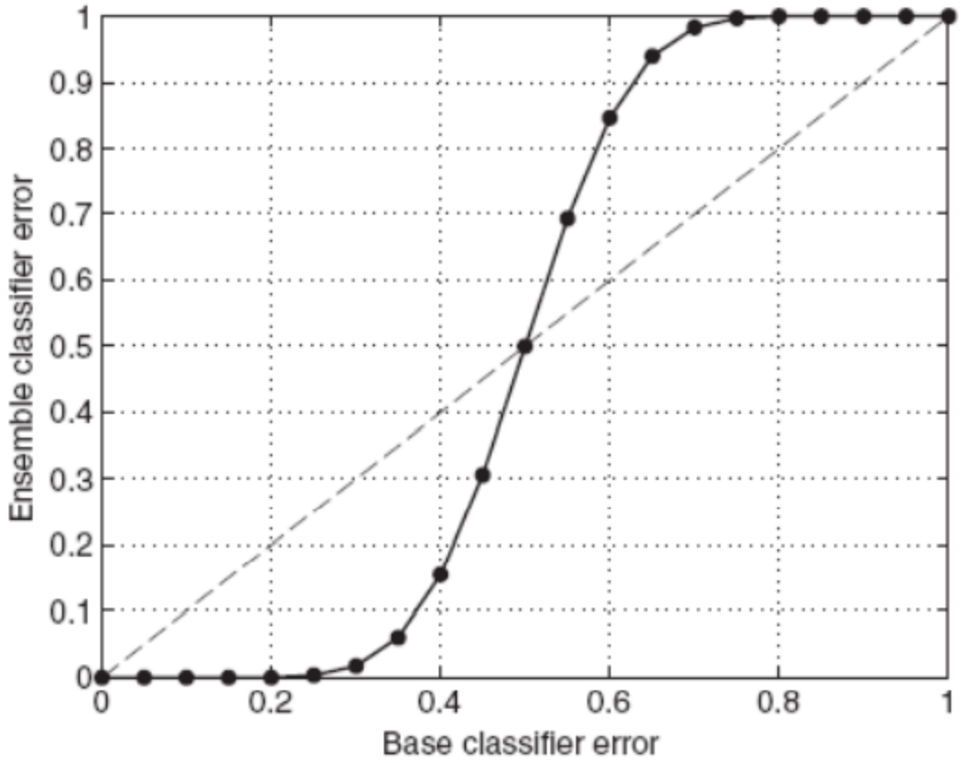
- 각 학습기들간의 의존관계에 따라서 Model 이 분류된다
- 강한 의존관계
  - Serial로 생성되는 연속적인 방법으로 분류가 진행됨
  - 해결하기 어려운 문제를 해결
    - 틀린 문제들에 가중치를 부여해 잘 해결할수 있도록 한다.
  - Boosting

- 약한 의존관계
  - 동시에 생성가능한 Parallel 한 방법으로 학습이 진행됨
  - 일반적으로 사용할수 있는 Model 을 만들어내는게 목적

- Bagging, RandomForest

객체 학습기의 훈련 상태와 성능의 관계

- 각 객체 학습기의 성능이 Ensemble 모델 성능에 영향을 미친다.



	테스트 샘플$_1$	테스트 샘플$_2$	테스트 샘플$_3$
h_1	√	√	×
h_2	×	√	√
h_3	√	×	√
앙상블	√	√	√

(a) 앙상블 성능 향상

	테스트 샘플$_1$	테스트 샘플$_2$	테스트 샘플$_3$
h_1	√	√	×
h_2	√	√	×
h_3	√	√	×
앙상블	√	√	×

(b) 앙상블 효과 없음

	테스트 샘플$_1$	테스트 샘플$_2$	테스트 샘플$_3$
h_1	√	×	×
h_2	×	√	×
h_3	×	×	√
앙상블	×	×	×

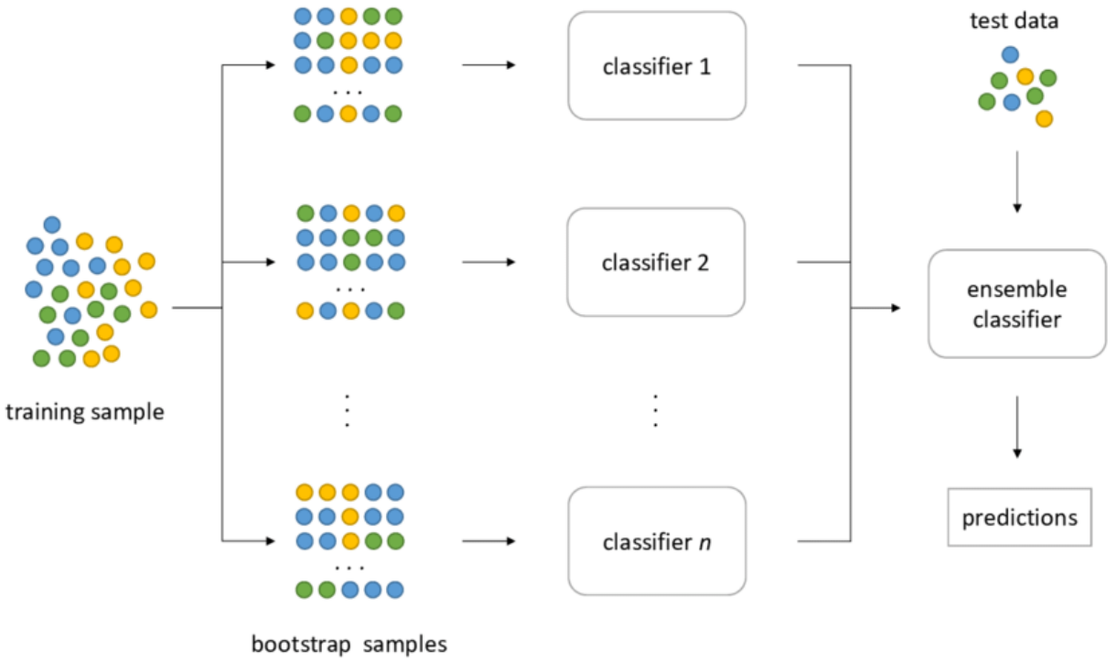
(c) 앙상블 부작용

그림 8.2 ＼ 앙상블 모델의 객체 학습기는 좋으면서 다양해야 한다
(h_i는 i번째 분류기를 뜻함)

Sampling

- 중복을 허용한 리샘플링(resampling)을 **부트스트래핑**(bootstraping)
- 중복을 허용하지 않는 샘플링 방식을 **페이스팅**(pasting) 이라고 한다.

Bootstrapping



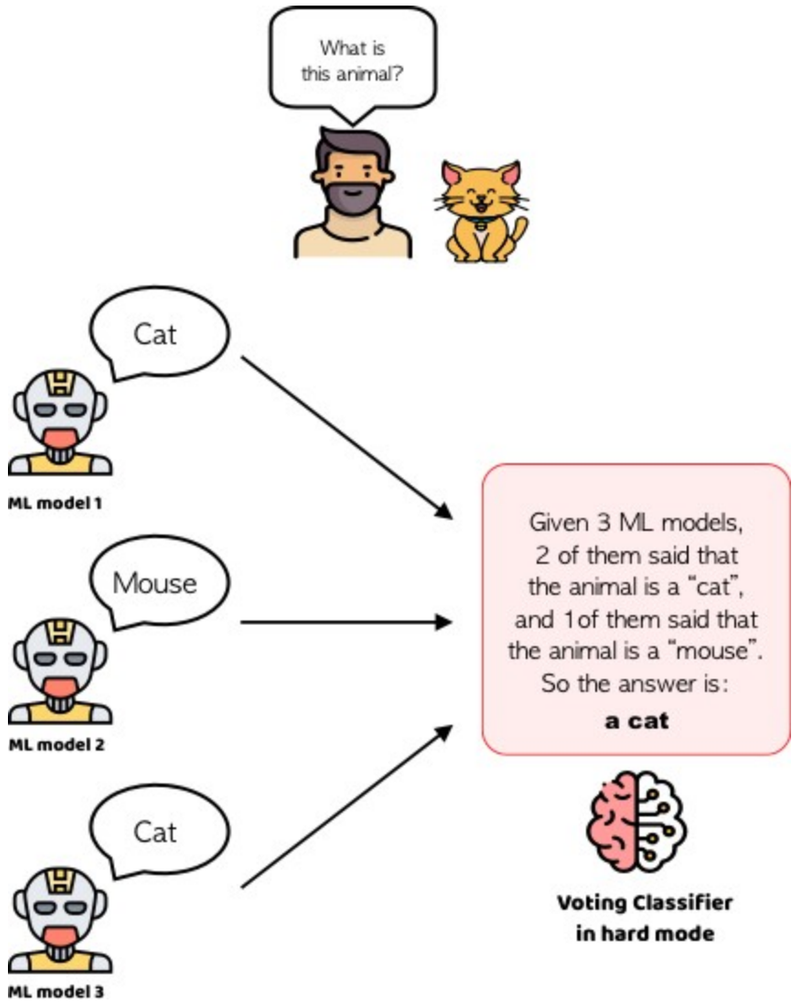
- 각 객체학습기는 서로 다른 Train Data Set 을 사용
- 복원추출 (Sampling with replacement) 을 통해 원래 데이터의 크기만큼을 갖도록 Sampling 한다.
- 이때 복원추출된 각각의 개별 DataSet을 Bootstrap Set이라고 부른다.
- 복원 추출이지만 한개의 Dataset이 Bootstrap에 선택되지 않을 확률은 0.368 이다.
- $\lim_{n\to\infty}(1-\frac{1}{n})^n = e^{-1}$

Aggregation

- 결과를 취합하는 방식

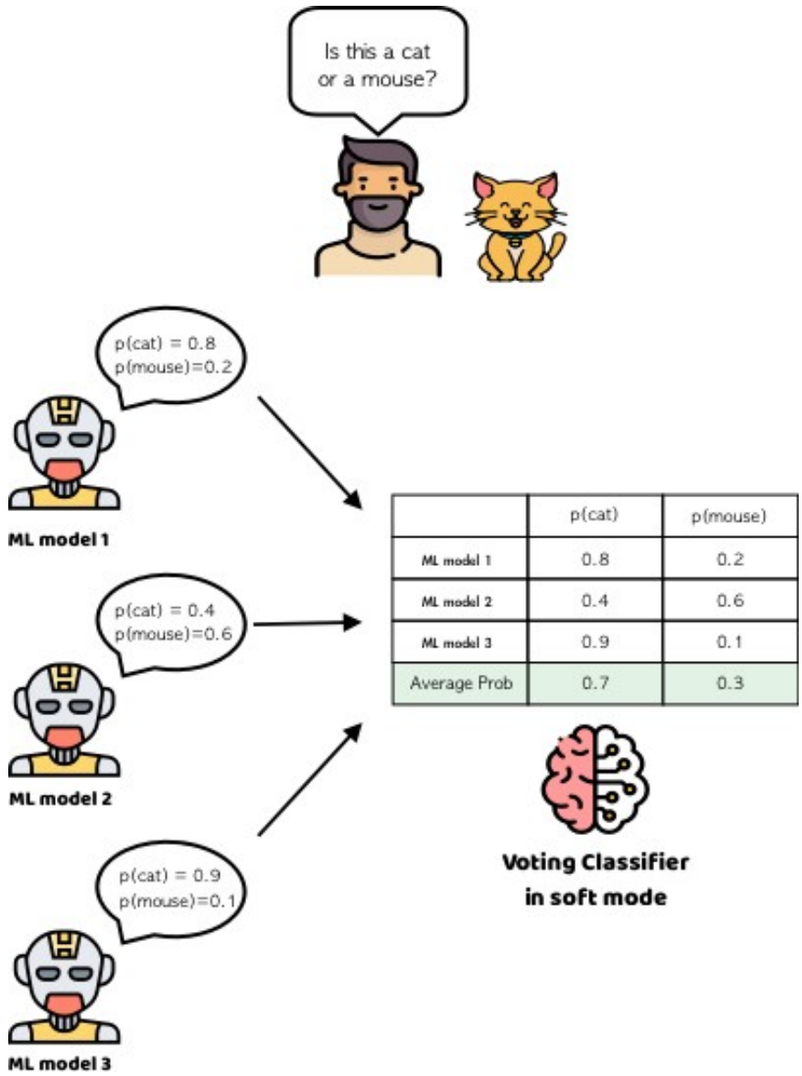
Majority Voting(Hard Voting)

- 가장 많이 투표된 값으로 선택



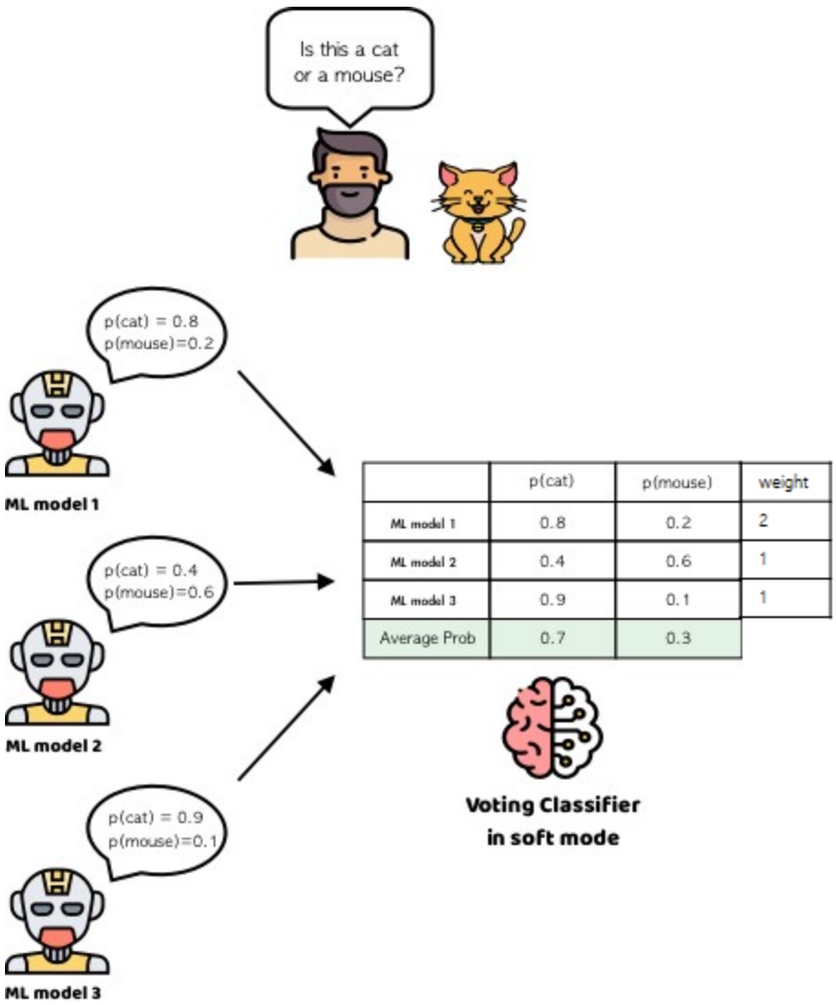
Probability Voting(Soft Voting)

- 각 Class 가 내놓은 예측 확률값의 평균



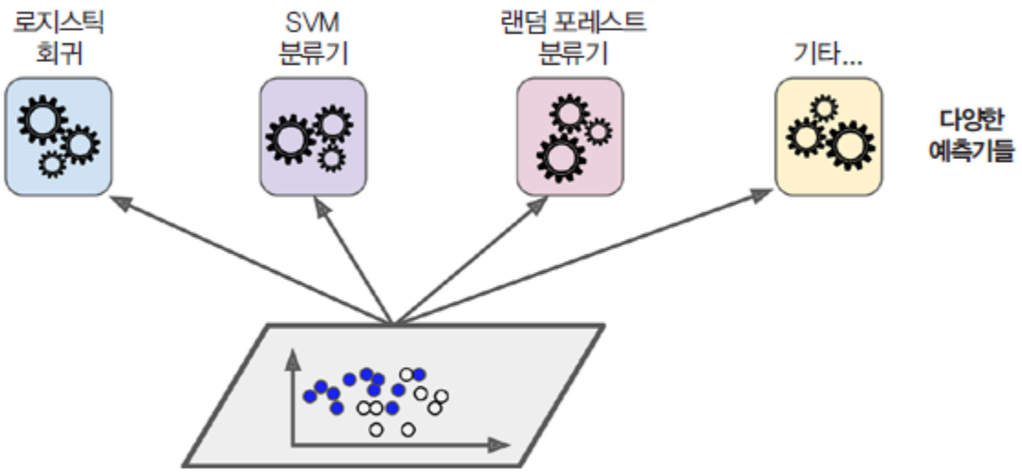
Weighted Voting

- 각 Class별 지정된 weight 로 사용해 판단한다.
- Weight 는 training accuracy 등을 사용할수 있다.



Voting Classifier

- Basic Ensnmbel Classifier
- 정확도가 80%인 분류기 여러 개를 훈련시켰다고 가정: 로지스틱 회귀 분류기, SVM 분류기, 랜덤 포레스트 분류기, K-최근접 이웃 분류기 등



▲ 그림 7-1 여러 분류기 훈련시키기

Hard voting

- `VotingClassifier` `voting='hard'`
- 직접 투표(hard voting) 분류기: 다수결 투표로 정해지는 분류기. 더 좋은 분류기를 만드는 매우 간단한 방법은 각 분류기의 예측을 모아서 가장 많이 선택된 클래스를 예측하는 것임
- 각 분류기의 예측값(레이블)을 가지고 **다수결 투표**를 통해 최종 앙상블 예측이 이루어진다.
- 강한 학습기
- 약한 학습기
- 큰 수의 법칙

Soft Voting

- `VotingClassifier` `voting='soft'`
- 각 분류기의 예측값(레이블)의 **확률을 가지고 평균을 구한 뒤, 평균이 가장 높은 클래스**로 최종 앙상블 예측이 이루어진다. 이러한 방법을 **간접 투표** (soft voting) 이라고 한다.

Ensemble 의 종류

- Bagging
  - Random Forest
- pasting
- Boosting
  - AdaBoost(Adaptive Boost)
  - Gradient Boosting

    - XGBoost
    - LightGBM
    - GLMBoost
- Staking