

---

# ENSEMBLE LEARNING

# Ensemble learning

- Motivations:
  - Ensemble model improves accuracy and robustness over single model methods.
  - A complex problem can be decomposed into multiple sub-problems that are easier to understand and solve (divide-and-conquer approach).

یادگیری گروهی  
انگیزه ها:  
مدل گروهی دقت و استحکام را نسبت به روش های تک مدل بهبود می بخشد.  
یک مسئله پیچیده را می توان به چندین مساله فرعی که درک و حل آن آسان تر است تجزیه کرد (رویکرد تفرقه بیانداز و حکومت کن).

# Ensemble learning

## Popular methods:

- Bagging
- Boosting
- Stacking

روش های محبوب:

- کیسه زدن
- افزایش
- پشته سازی

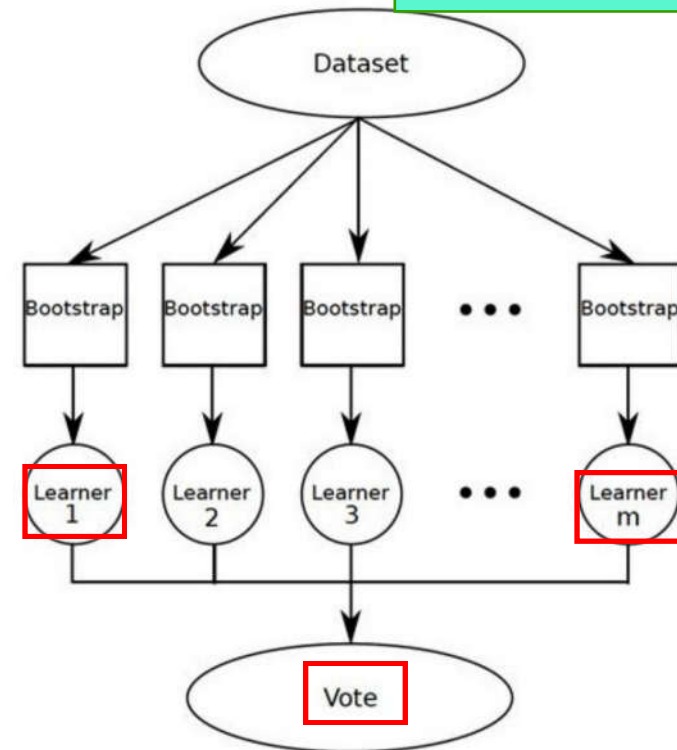
تکنیک هایی که نظر چندتا کلسیفایر را میپرسند.  
ایده: چون داریم از چندتا مدل استفاده میکنیم  
انتظار داریم نتایج بهتر باشه احتمال اورفیت شدن  
یکی کمتر میشه  
یه مساله دیگه هم اینه که بعضی از مسائل اونقدر  
سخت هستند که نباید انتظار داشته باشیم یه  
کلسیفایر بتواند حلشون کنه به تنهایی

۳تا رویکرد برا شکستن مساله به بخش های  
کوچک تر

# Bagging: Bootstrap Aggregation

کل داده ها را به چندین بخش تقسیم میکنیم و هر بخش را به یک مدل میدهم  
 وقتی داده ی جدیدی اومد بین این  $m$  تا مدل رای گیری میکنیم هر جوابی که بیشترین رای را آورد را به عنوان نتیجه نهایی در نظر میگیریم.

- **Training**
  - Given a set  $D$  of  $d$  tuples, at each iteration  $i$ , a training set  $D_i$  of  $d$  tuples is sampled with replacement from  $D$  (i.e., bootstrap)
  - A classifier model  $M_i$  is learned for each training set  $D_i$
- **Classification**: to classify an unknown sample  $X$   
 Each classifier  $M_i$  returns its class prediction •  
 The bagged classifier  $M^*$  counts the votes and assigns the class with the most votes to  $X$
- **Regression**: take the average value instead of voting  
 رگرسیون: به جای رای دادن، مقدار متوسط را بگیرد
- **Bagging** produces a combined model that often performs significantly better than the single model built from the original training data, and is never substantially worse.



عروف ترین روش  
bagging

- **Example**
  - **Random forest**

با یک تعداد زیادی درخت، داده هامون را مدل میکنیم و هر درختی مسئول یه بخشی از داده هاست.

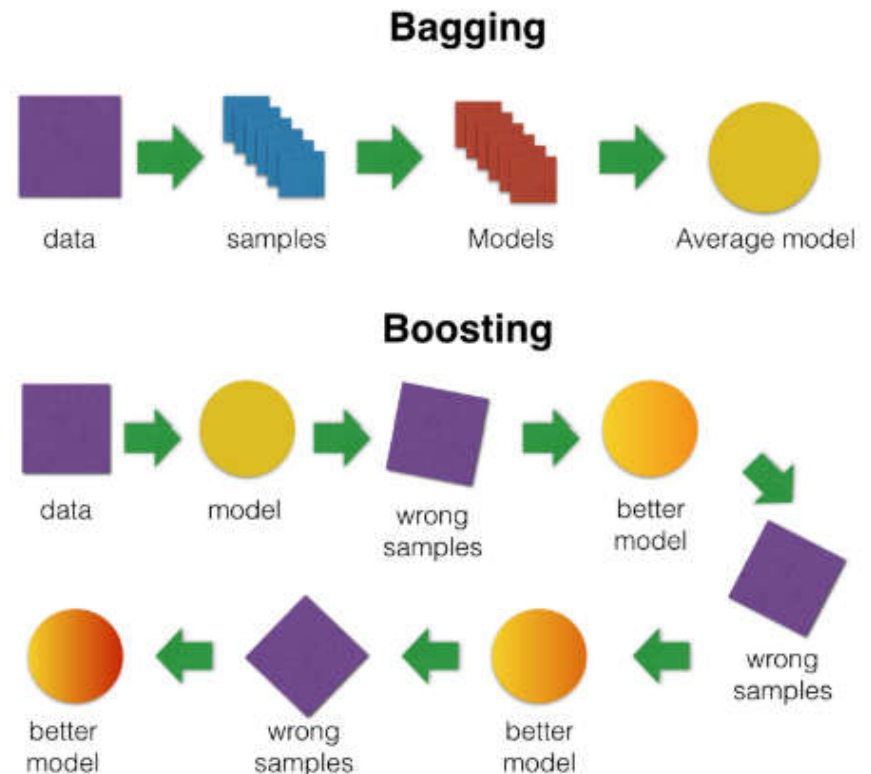
در مسائل رگرشن یک میانگین از خروجی این مدل ها میگیریم به عنوان نتیجه نهایی

Bagging یک مدل ترکیبی تولید می کند که اغلب به طور قابل توجهی بهتر از مدل منفرد ساخته شده از داده های آموزشی اصلی عمل می کند و هرگز به طور قابل توجهی بدتر نیست. مثال- جنگل تصادفی

طبقه بندی: طبقه بندی یک نمونه ناشناخته  $X$  هر طبقه بندی کننده  $M_i$  پیشبینی کلاس خود را برمیگرداند طبقه بندی کننده کیسه ای  $M^*$  آرا را می شمارد و کلاسی را که بیشترین آرا را دارد به  $X$  اختصاص می دهد.

## 2. Boosting

- Comparing with bagging:
- The same base classifiers are used in both
- Boosting uses weighted voting/averaging
- Bagging is parallel while boosting is sequential
- Boosting tends to have greater accuracy, but it also risks overfitting the model to misclassified data.
- Example
  - adaboost
  - XGBoost



2. تقویت  
مقایسه با کیسه بندی:  
طبقه بندی کننده های پایه یکسان در هر دو استفاده می شود  
تقویت از رای گیری وزنی/میانگین استفاده می کند  
بسته بندی موازی است در حالی که تقویت متوالی است  
تقویت تمایل به دقت بیشتری دارد، اما همچنین خطر تطبیق بیش از حد مدل را با داده های طبقه بندی نادرست دارد.

# 3. Stacking

- It introduces the concept of a **metalearner**, which replaces the voting procedure.
- Normally is used to **combine** models of **different types**.
- Because most of the work is already done by the level-0 learners, it makes sense to choose a rather **simple algorithm** for the level-1 classifier.
- Use **out-of-fold predictions (OOF)** as the **training data** for the level 1 classifier.

