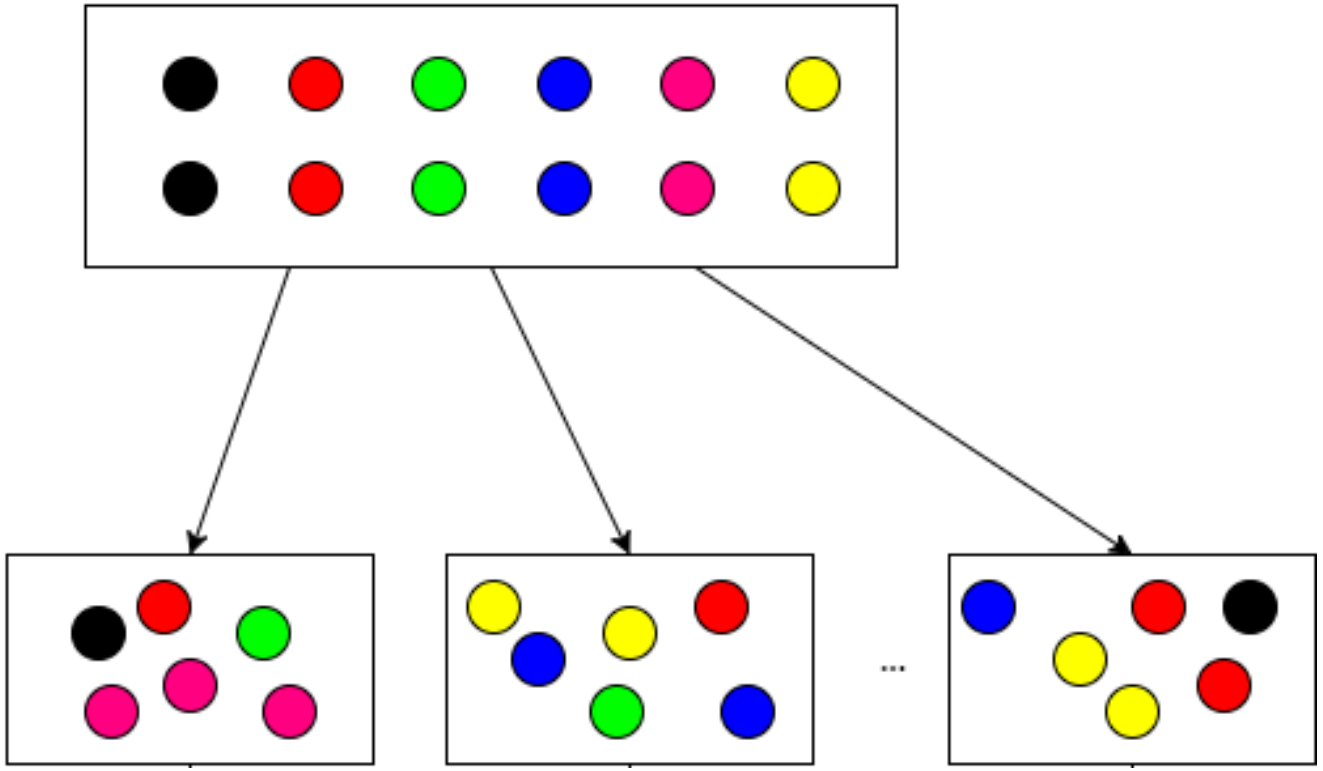
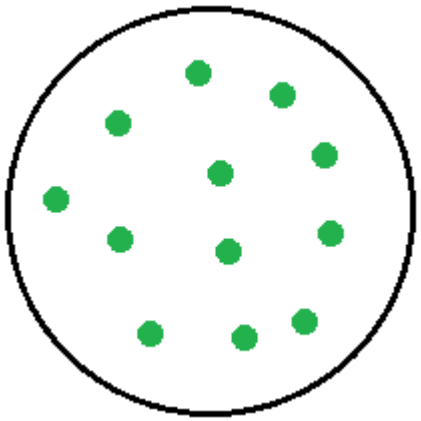


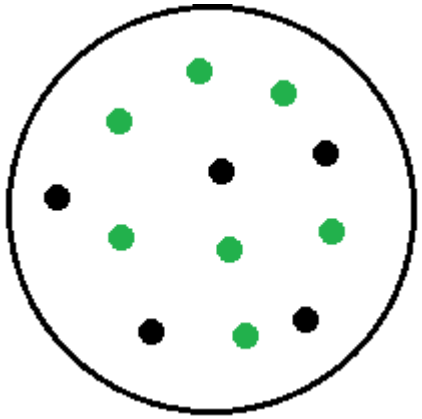
Bagging

Manipulacija trening skupom

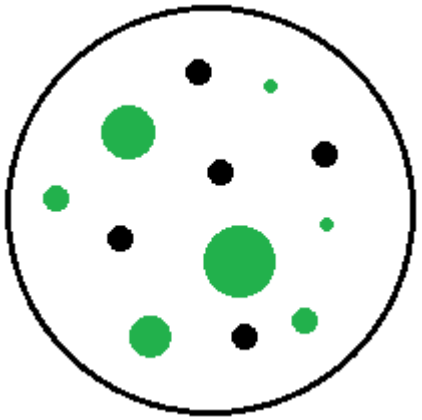
- Na *različitim* podacima više puta trenirati *isti* model
- Kako da dobijemo različite podatke (u praksi, raspolažemo sa samo jednim skupom podataka)?



Jedan model
koristiti ceo skup podataka



Bagging
Slučajno uzorkovanje sa zamenom



Boosting
Slučajno uzorkovanje sa zamenom, pri čemu neki primeri imaju veću šansu da budu selektovani

* Zelene tačke su selektovane za treniranje modela

Bagging (*Bootstrap Aggregation*)

Ulaz	• $T = \{(x^{(i)}, y^{(i)}), i = 1, \dots, N\}$ – trening skup • M – broj slabih prediktora • Osnovni model L (korišćen za obuku pojedinačnih članova ansambla)

Postupak	• for $m = 1, \dots, M$
	  • Napraviti podskup skupa podataka T_m uzorkovanjem *N* primera sa zamenom iz *T*
	  • Obučiti model $h_m = L(T_m)$

Izlaz	$\hat{y}(x) = \frac{1}{M} \sum_{m=1}^{M} h_m(x)$

Različitost trening skupova

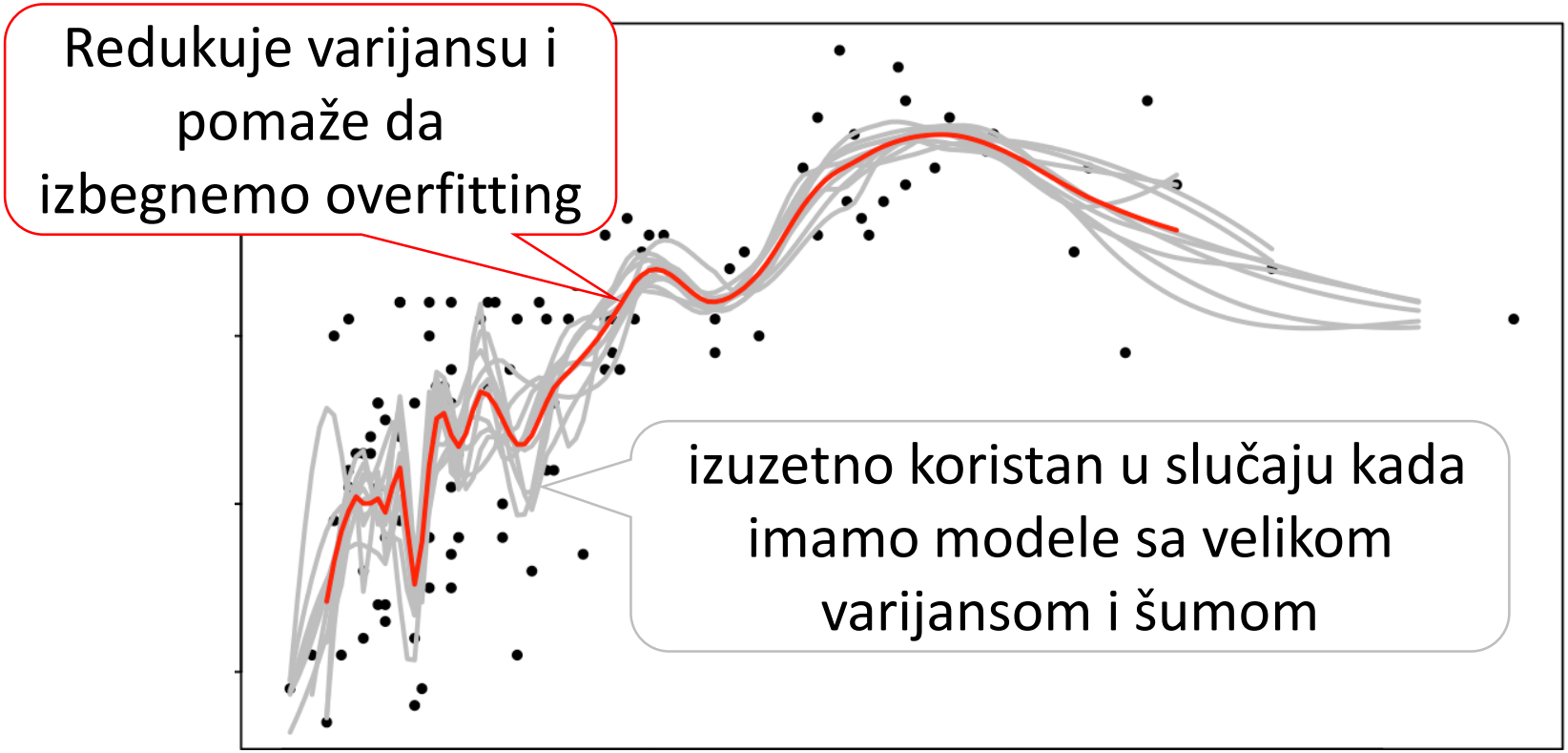
- Svaki podskup T_m ima isti broj primera *N* kao i originalni skup podataka *T*
- Neki primeri iz *T* se ponavljaju u T_m, a neki su izostavljeni
- Podskupovi se u razumnoj meri razlikuju
  - Instanca ima verovatnoću $\left(1 - \frac{1}{N}\right)^N \approx e^{-1} \approx 0.368$ da ne bude odabrana za trening skup
  - To znači da će u T_m biti oko 63.2% originalnih/jedinstvenih instanci

Random forest

- Najpopularniji algoritam je *Random forests*
- Kao osnovni model se koristi stablo odluke (*decision tree*)
- Umesto podskupova primera, koriste se podskupovi obeležja (svaki model je zasnovan na $\sqrt{\text{no. features}}$ obeležja)

Bagging prednosti

- Iznenađujuće kompetitivne performanse



- Manje osetljiv na nerelevantna obeležja