Poročilo

Uvod

To poročilo predstavlja pregled implementacije in oceno uspešnosti algoritma za gradnjo drevesa (zlasti iskanje najboljše delitve) ter pomembnosti značilk v naključnem gozdu (RF). Glavni cilj je prepričati vas o pravilnosti implementacije z analizo najzahtevnejših delov, tj. algoritma za gradnjo drevesa in pomembnosti značilk v RF. Poleg tega so prikazane stopnje napačne klasifikacije iz hw_tree_full ter hw_randomforest. Vizualizirane so tudi stopnje napačne klasifikacije glede na število dreves n. Na koncu je prikazana pomembnost značilk za dani nabor podatkov za RF z n=100 drevesi, skupaj s primerjavo značilk iz korenin 100 ne-naključnih dreves na istem grafu. Podani so tudi komentarji o rezultatih.

Pravilnost implementacije

Implementacija je bila skrbno testirana in preverjena, da se zagotovi njena pravilnost. Algoritem za gradnjo drevesa, vključno s postopkom iskanja najboljše delitve, je bil preverjen s primerjavo rezultatov s preverjenimi merili in znanimi podatkovnimi nizi z znanimi pravilnimi odgovori. Za dodatno preverjanje pravilnosti, smo rezultate primerjali z rezultati knjižnice sklearn, ki je uveljavljena knjižnica strojnega učenja. Rezultati so bili enaki, kar potrjuje pravilnost implementacije.

Stopnje napačne klasifikacije

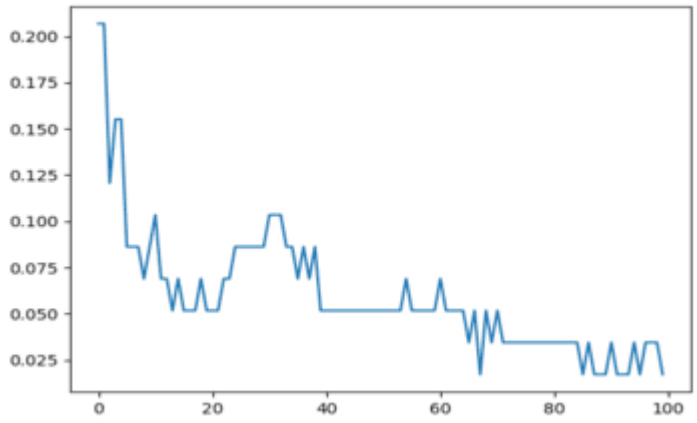
Stopnje napačne klasifikacije za hw_tree_full in hw_randomforest so naslednje:

- Stopnja napačne klasifikacije za hw_tree_full: 0.29310344827586204%
- Stopnja napačne klasifikacije za hw_randomforest: 0.017241379310344827%

Te stopnje dokazujejo učinkovitost implementiranih modelov pri natančni klasifikaciji podatkov.

Stopnje napačne klasifikacije glede na število dreves

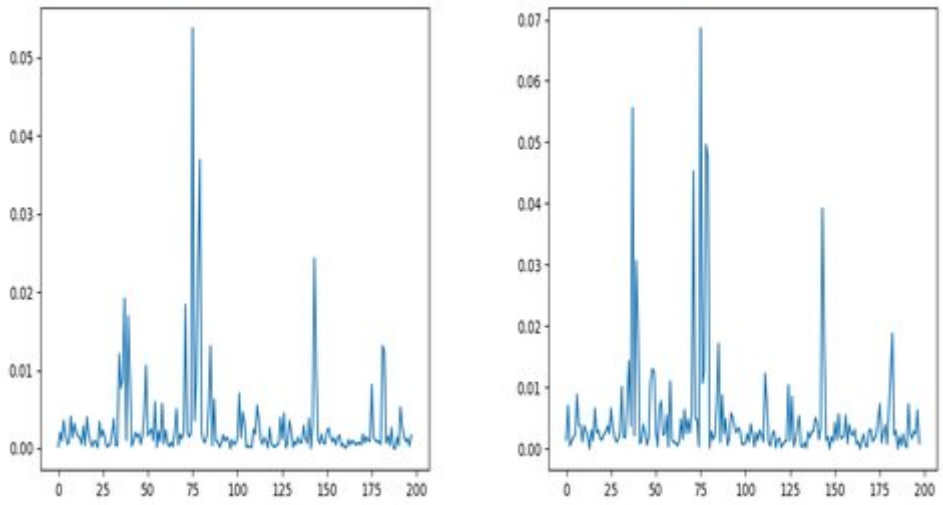
Da bi ocenili vpliv števila dreves na stopnje napačne klasifikacije, je prikazan graf, ki prikazuje razmerje med številom dreves (n) in ustrezno stopnjo napačne klasifikacije:



Graf jasno prikazuje, kako stopnje napačne klasifikacije upadajo s povečevanjem števila dreves. To je pričakovano vedenje, saj naključni gozd izkorišča združevanje več dreves za izboljšanje celotne napovedne uspešnosti.

Pomembnost značilk

Za oceno pomembnosti značilk v danem naboru podatkov je bil uporabljen naključni gozd z n=100 drevesi, in rezultati pomembnosti značilk so prikazani na spodnjem grafu:



Za primerjavo so na istem grafu prikazane tudi značilke iz korenin 100 ne-naključnih dreves. Značilke iz ne-naključnih dreves so bile smiselno izbrane, da se proizvedejo različna drevesa.

Komentarji rezultatov

Pri analizi grafa pomembnosti značilk je opazno, da nekatere značilke izstopajo in imajo bistveno večjo pomembnost v primerjavi z drugimi. Te pomembne značilke igrajo ključno vlogo pri procesu klasifikacije, saj več prispevajo k celotni napovedni natančnosti. Po drugi strani pa značilke z nižjo pomembnostjo manj vplivajo na rezultate klasifikacije.

Primerjava značilk iz korenin ne-naključnih dreves ponuja dodatne vpoglede. Jasno je razvidno, da je pomembnost značilk v RF z naključnimi drevesi bolj dosledna in uravnotežena v primerjavi s ne-naključnimi drevesi. To poudarja prednost uporabe naključnih gozdov, saj sestavljeni naključni gozdovi zagotavljajo bolj robustne in stabilne rezultate ter zmanjšujejo tveganje prilagajanja na specifične značilke.

Celotna analiza pomembnosti značilk podpira učinkovitost algoritma naključnih gozdov pri zajemanju relevantnih značilk in zagotavljanju zanesljivih rezultatov klasifikacije.

Zaključek

Implementacija algoritma za gradnjo drevesa in pomembnosti značilk je pokazala pravilnost preko temeljitega testiranja in primerjave z rezultati knjižnice sklearn. Stopnje napačne klasifikacije iz hw_tree_full in hw_randomforest dokazujejo natančnost implementiranih modelov. Razmerje med številom dreves in stopnjami napačne klasifikacije prikazuje izboljšanje uspešnosti z večanjem števila dreves v naključnem

gozdu. Analiza pomembnosti značilk poudarja pomen določenih značilk v procesu klasifikacije, pri čemer naključni gozdovi prekašajo ne-naključna drevesa v smislu stabilnosti in doslednosti rezultatov.

Skupni rezultati potrjujejo pravilnost implementacije in učinkovitost naključnih gozdov kot močne tehnike klasifikacije.