Dokumentácia USU Sem 2

Bagging and regression trees

Popis programu:

Program obsahuje 2 triedy, a to main.py a Bag.py. main.py je hlavným súborom programu. Načítajú sa v ňom dáta, kategorické dáta sa premenia na numerické, rozdelia na trénovaciu a testovaciu množinu. Následne sa použije bootstrap aggregating (bagging) a operácie spojené s experimentami.

Trieda Bag.py je implementáciu bootstrap aggregatingu (baggingu) pričom na predikciu využíva implementáciu regresného stromu z knižnice scikit-learn od pythonu. Do konštruktora vchádzjú parametere num\_of\_bags a length\_of\_subtrain, ktoré udávajú počet vytvorených náhodných podmnožín trénovacích dát a percentuálnu veľkosť vytvorenej trénovacej podmnožiny ku pôvodnej množine. V metóde fit sa vytvárajú trénovacie podmnožiny, pričom sa prvky z pôvodnej množiny môžu opakovať. Ďalej sa tu nad vytvorenou podmnožinou natrénuje inštancia triedy DecissionTreeReggresor a pridá sa do poľa modelov. Metóda predict slúži na predikciu všetkých modelov, ktoré sú uložené v poli a potom z nich urobí priemer.

Prezentácia výsledkov:

Pri experimentoch sa skúšali rôzne kombinácie množstva vytvorených podmnožín a percentuálnej veľkosti podmnožiny voči pôvodnej množine, a ako budú ovplivňovať výsledné MSE (mean squared error). Pre každú kombináciu sa vykonalo 1000 opakovaní a následne sa spravil priemer.

Vplyv počtu modelov :

Pozorovalo sa vplyv počtu podmnožín na výsledné MSE pričom sa začínalo na 5 a konečný počet bol 100. Pri zvyšovaní počtu podmnožín dochádzalo ku znižovaní MSE čo naznačovalo, že väčší počet zlepšuje schopnosť lepšie generalizovať a tým pádom aj lepšie predikovať. Pri postupnom zvyšovaní z 5 až na 50 sa výsledky rádovo zlepšili. Pri 5 bolo MSE 157.30 a pri 50 bolo MSE 138.40. Avšak pri následnom zvýšení na 100 sa MSE znížilo na 137.43 čo je skoro nebadateľné. Preto sa už ďalej so zvyšovaním nepokračovalo. Výsledný počet podmnožín, ktoré by mohli byť použité by mal byť niekde medzi 50-100.

Vplyv percentuálnej veľkosti podmnožiny:

Pozorovalo sa vplyv percentuálnej veľkosti podmnožiny na výsledné MSE pričom sa brali do úvahy tieto 40, 50, 60 a 70. Bol zaznamenaný trend, kedy sa MSE znižovalo pri použití väčšieho množstva dát. Rozdiely sú relatívne mále, ale naznačujú že väčšia podmnožina má pozitívny vplyv na zníženie MSE.

Optimálny výber parametrov:

Pri výbere optimálnych parametrov by som odporučil väčší počet podmnožín 50-100 a percentuálnu veľkosť podmnožiny by som volili od 50-70.

Podrobné výsledky experimentu:

A screen shot of a computer

Description automatically generated

Experiment s parametrami DecisionTreeRegressor

Experimenty s parametrami stromu boli vykonané s optimálnymi parametrami pre Bag, čo je lengh\_of\_subtrain = 60 a num\_of\_bags = 50.

Max Depth:  
 Do experimentu vstupobali hodnoty  [3, 5, 10, None], pričom hodnota MSE pri 3 bola vyššia čo naznačuje, že model bol príliš jednoduchý. Pri 5 sa MSE znížilo, čo naznačuje kompromis medzi komplexnosťou a jednoduchosťou modelu. Pri vyšších hĺbkach sa mSE znovu zhoršovalo.  
  
Min Samples Split:

Pri hodnote nižšej ako napr. 2 je hodnota MSE nižšia ako pri vyšších, to naznačuje, že rozdelenie aj pri nižšom počte vzoriek pomáha modelu zachytiť detaily a tým znížiť MSE.

Optimálny výber parametrov:

Pri výbere optimálnych parametrov sa najlepšie MSE dosiahlo kombináciou max\_depth = 5 a min\_samples\_split = 2.

Podrobné výsledky experimentu:

A screen shot of a computer screen

Description automatically generated