Documentatie

Detectarea anomaliilor cerebrale

Dilirici Mihai, grupa 243

**Naive Bayes**

Descriere model

Bayes naiv este o tehnică simplă pentru construirea clasificatorilor: modele care atribuie etichete de clasă pentru instanțe noi, reprezentate ca vectori de valori pentru diverse caracteristici, unde etichetele de clasă fac parte trase dintr-o mulțime finită.

Clasificatorii bayesieni naivi presupun că valoarea unei anumite caracteristică este independentă de valoarea oricărei altă caracteristici, dat fiind variabila de clasă. De exemplu, un fruct poate fi considerat a fi un măr dacă este roșu, rotund, și de aproximativ 10 cm în diametru. Un clasificator naiv bayesian consideră că fiecare dintre aceste caracteristici contribuie în mod independent la probabilitatea ca acest fruct să fie un măr, indiferent de eventualele corelatii între culoare, rotunjime și diametru, ceea ce poate duce la subestimarea interacțiunilor dintre caracteristici.

Modelul Naive Bayes este adecvat pentru probleme de clasificare a datelor cu multiple caracteristici discrete sau continue și cu clase echilibrate. În cazul în care datele de antrenament sunt foarte mari, se poate întâmpla să apară probleme de sub-estimare, din cauza lipsei legaturii dintre caracteristici.

Implementare

Pentru a implementa acest model, se încarcă imaginile într-o listă de numpy arrays, apoi se normalizează valorile pixelilor împărțind la 255. Imaginile sunt apoi salvate în fișiere numpy. În același mod se încarcă și etichetele pentru imagini.

Modelul Naive Bayes este antrenat cu funcția fit() pe setul de antrenament și se prezice clasa imaginilor din setul de validare folosind funcția predict(). Este evaluată performanța modelului pe setul de validare cu ajutorul măsurii F1-score.

De asemenea, se generează un fișier CSV cu etichetele prezise pentru imaginile din setul de testare și se calculează matricea de confuzie și raportul de clasificare pentru setul de validare.

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| Tumor | 0.93 | 0.66 | 0.77 |
| Non-tumor | 0.25 | 0.68 | 0.36 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| Macro avg | 0.59 | 0.67 | 0.57 |
| Weighted avg | 0.83 | 0.67 | 0.72 |

|  |  |  |
| --- | --- | --- |
|  | C1 | C2 |
| L0 | 1143 | 581 |
| L1 | 87 | 189 |

**Random forest**

Descriere

Modelul de Random Forest este o metoda de invatare automata supervizata de clasificare si regresie. Scopul acestui model este de a crea un set de arbori de decizie, unde fiecare arbore este antrenat pe un subset diferit de date aleatorii din setul de antrenare.

Pentru sarcinile de clasificare, rezultatul pădurii aleatoare este clasa selectată de cei mai mulți arbori. Pentru sarcinile de regresie, este returnată media sau predicția medie a arborilor individuali. Pădurile decizionale aleatoare corectează obiceiul arborilor de decizie de a supra-ajusta setul lor de antrenare. În general, pădurile aleatoare performează mai bine decât arborii de decizie.

Modelul de Random Forest este adesea preferat in cazul datelor complexe, datorita capacitatii sale de a reduce varianata si de a evita overfitting-ul. De asemenea, acest model este usor de interpretat si poate fi folosit cu succes pentru clasificarea multor tipuri de date.

Implementare

Algoritmul a fost folosit pentru a clasifica imagini in tumore maligne sau benigne pe baza unui set de caracteristici extrase din imaginile medicale. Imaginile au fost prelucrate si transmise ca vectori de caracteristici la algoritmul de Random Forest, care a fost antrenat pe aceste date.

Dupa antrenare, modelul a fost folosit pentru a face predictii pe setul de validare si setul de testare, iar performanta modelului a fost evaluata prin calcularea scorurilor F1. De asemenea, a fost generat un raport de clasificare si o matrice de confuzie pentru a evalua performanta modelului in detaliu.

Model basic

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| Tumor | 0.91 | 0.96 | 0.93 |
| Non-tumor | 0.58 | 0.38 | 0.46 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| Macro avg | 0.74 | 0.67 | 0.69 |
| Weighted avg | 0.86 | 0.88 | 0.87 |

|  |  |  |
| --- | --- | --- |
|  | C1 | C2 |
| L0 | 1659 | 75 |
| L1 | 172 | 104 |

Model imbunatatit

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| Tumor | 0.94 | 0.90 | 0.92 |
| Non-tumor | 0.51 | 0.66 | 0.57 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| Macro avg | 0.72 | 0.78 | 0.75 |
| Weighted avg | 0.88 | 0.86 | 0.87 |

|  |  |  |
| --- | --- | --- |
|  | C1 | C2 |
| L0 | 1547 | 177 |
| L1 | 94 | 182 |

In primele doua seturi de tabele nu sunt diferenta mari, insa in al 3lea tabel, cu matricea de confuzie, modelul imbunatatit a nimerit de aproape 2 ori mai multe imagini cu tumori fata de cel normal, pe cand diferenta la imaginile fara tumori nu este asa de semnificativa.