Istraživanje podataka

Vežbe 8

4. Mai 2018

Outline

Outline

Algoritam

- Density-based spatial clustering of applications with noise
- Algoritam DBSCAN može pronaći klastere proizvoljnog oblika

Algoritam

Parametri

- *Eps* - prag za rastojanje suseda. Dve instance su susedne ako im je rastojanje manje ili jednako *Eps*
- *MinPts* - prag za broj suseda instanci

Algoritam

Podela instanci

- Instance u jezgru klastera - instanca je u jezgru klastera ako je broj suseda na rastojanju *Eps* bar *MinPts* .
- Instance na granici klastera - instanca nije u jezgru, ali je na rastojanju do *Eps* nekoj instanci koja je u jezgru klastera.
- Šum - instanca koja nije ni u jezgru ni na granici klastera.

Algoritam

Koraci

1. Za svaku instancu odrediti tip: u jezgru, na granici ili šum.
2. Eliminisati instance koje su šum.
3. Povezati sve instance u jezgru koje su na međusobnom rastojanju do *Eps*.
4. Napraviti poseban klaster za svaku grupu instanci u jezgru koje su povezane.
5. Svaku instancu na granici dodeliti klasteru kojem pripada instanca u jezgru u čijem je susedstvu ta instanca na granici.

Outline

Klasifikacija

Ulazni podatak u klasifikaciju je skup slogova. Svaki slog je oblika (x, y) gde je x skup atributa, a y ciljni atribut koji sadrži oznaku klase. Potrebno je naći klasifikacioni model (funkciju) koji preslikava svaki skup atributa x u jednu od predefinisanih oznaka klasa y.

Klasifikacija

Ulazni podaci se dele u **dva** dela:

- podatke za trening pomoću kojih se pravi model
- podatke za testiranje koji se koriste za proveru ispravnosti modela

Provera modela

Preciznost

$$\textit{Preciznost} = \frac{\text{Broj slogova čija klasa je dobro predviđena modelom}}{\text{Ukupan broj slogova}}$$

Stopa greške

$$\text{Stopa greške} = \frac{\text{Broj slogova čija klasa nije dobro predviđena modelom}}{\text{Ukupan broj slogova}}$$

Drveta odlučivanja

- Model klasifikacije se predstavlja kao drvo odlučivanja koje ima
  - unutrašnje čvorove. Svaki unutrašnji čvor sadrži uslov nad test atributom koji služi za podelu slogova koji imaju različite karakteristike. Grane koje izlaze iz unutrašnjeg čvora odgovaraju mogućim vrednostima test atributa.
  - listova. Svakom listu je dodeljena jedna klasa.

Drveta odlučivanja - klasifikacija sloga

Klasifikacija sloga: počevši od korena drveta odlučivanja, primenjuje se test uslov nad slogom i prati se grana koja odgovara dobijenom rezultatu. Ukoliko se pri spuštanju niz drvo odlučivanja naiđe na unutrašni čvor, postupak se ponavlja (test uslov se primenjuje na slog i prati se grana koja odgovara rezultatu testa). Ako se naiđe na list, slogu se dodeljuje klasa koja je pridružena tom listu.

Drveta odlučivanja - pravljenje drveta odlučivanja

Opšti algoritam

1. Neka je D_t skup slogova za trening koji se nalaze u čvoru t, a $y = y_1, ..., y_c$ su oznake klasa
2. Ako D_t sadrži samo slogove koji pripadaju jednoj klasi y_t, tada je t list označen sa y_t
3. Ako D_t sadrži slogove koji se nalaze u više od jedne klase, tada se koristi test atribut radi podele podataka u manje podskupove. Na dobijene podskupove se zatim rekurzivno primenjuje kompletna procedura.

Mere nečistoće

$p(j|t)$ je relativna frekvencija klase j u čvoru t

Ginijev indeks

$$Gini(t) = 1 - \sum_j [p(j|t)]^2$$

Entropija

$$Entropy(t) = -\sum_j p(j|t) * \log p(j|t)$$

Greška klasifikacije

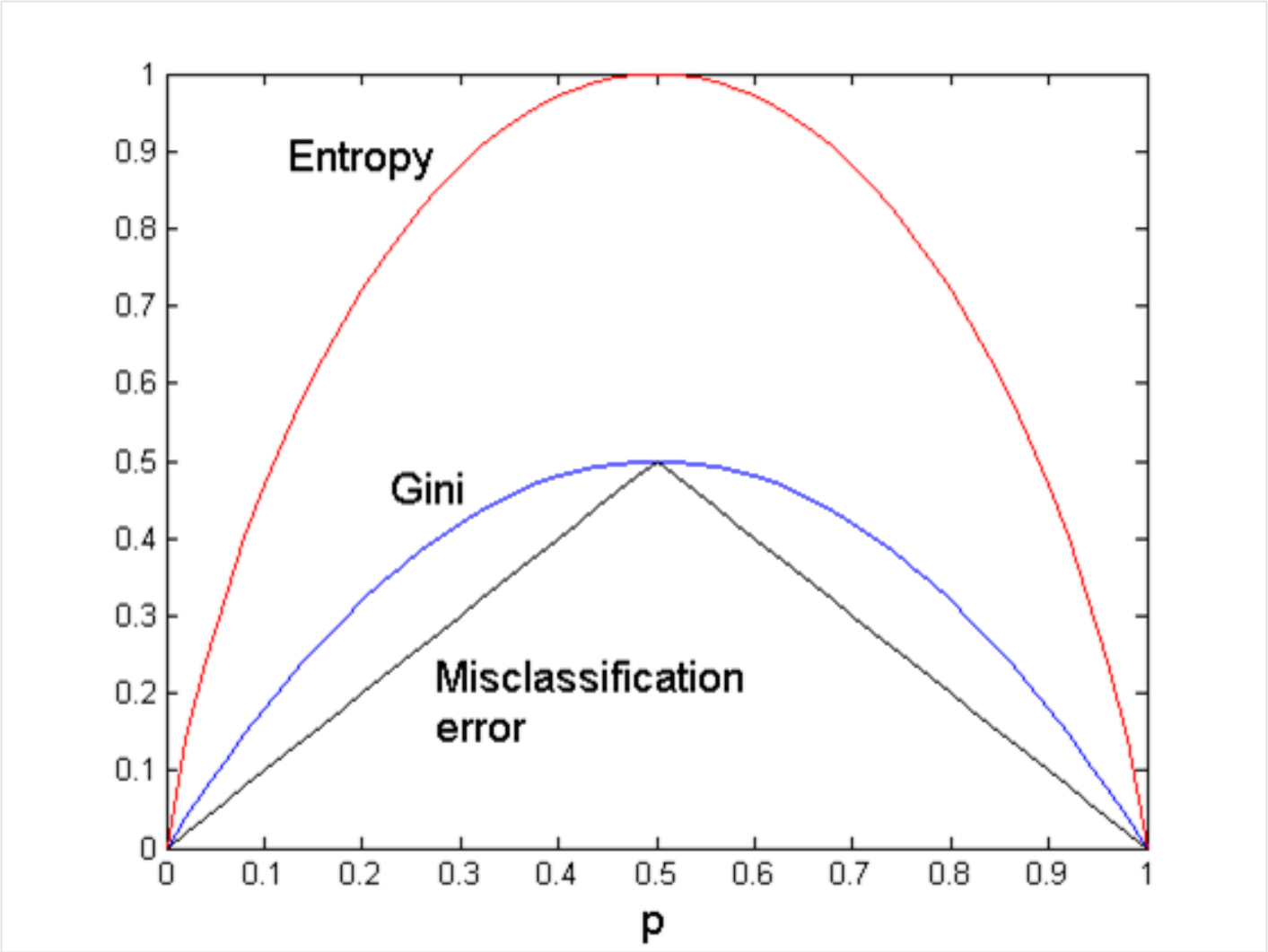
$$Error(t) = 1 - \max_j p(j|t)$$

Mere nečistoće

Dobit

$$\Delta = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} * I(v_j)$$

Mere nečistoće



Zadatak 1

Dati su trening primeri za problem binarne klasifikacije.

- Kolika je entropija skupa trening podataka?
- Kolika je informaciona dobit za a_1, a kolika za a_2 na ovim trening podacima?
- Za a_3, koji je neprekidan atribut, izračunati informacionu dobit za svaku moguću podelu.
- Koja je najbolja podela (između a_1, a_2 i a_3) prema informacionoj dobiti?
- Koja je najbolja podela (između a_1 i a_2) prema grešci klasifikacije?

Zadatak 1

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

Zadatak 2

Na osnovu datih podataka o životinjama iz trening skupa proceniti da li je životinja osobinama (*Velika*, *Biljke*, *Da*) opasna ili ne korišćenjem stabla odlučivanja dubine 2 uz korišćenje Ginijevog indeksa.

Veličina	Ishrana	Otrovnost	Opasna
Velika	Meso	Ne	Da
Mala	Meso	Ne	Ne
Mala	Biljke	Ne	Ne
Velika	Meso	Da	Da
Mala	Meso	Da	Da
Mala	Biljke	Ne	Ne
Mala	Biljke	Da	Da
Velika	Biljke	Ne	Da