Zadanie 7

Naiwny Bayes

Kacper Kania

28.04.2024

Treść zadania

Naiwny Bayes zakłada brak zależności między zmiennymi objaśniającymi. Mimo że to założenie ma mało wspólnego z rzeczywistością (zazwyczaj zmienne objaśniające tworzą skomplikowane hierarchie), okazuje się, że jest to „wystarczająco dobre'' mierzone skuteczność klasyfikatora.

Zazwyczaj Naiwny Bayes jest używany dla klasyfikacji w przypadku atrybutów kategorycznych. Aby metoda mogła działać na atrybutach ciągłych, zakłada się, że wartości atrybutów pochodzą z rozkładu normalnego (Gaussowski Naiwny Bayes). Efektywnie, taki klasyfikator ma stałą liczbę parametrów, niezależną od liczby różnych wartości, jaką atrybut może przyjąć (Naiwny Bayes modeluje prawdopodobieństwo wystąpienia każdej wartości atrybutu w danych treningowych względem każdej z klas).

Parametry są uczone na podstawie danych treningowych, a klasyfikacja odbywa się na podstawie maksymalizacji prawdopodobieństwa a posteriori. Rozkład a priori to rozkład normalny. Uczenie się odbywa metodą największę wiarygodności. Taki estymator dla rozkładu normalnego ma następujące postacie:

$$\hat{\mu}_{i=a|c=k} = \frac{1}{N[c=k]} \sum_{j=1}^{|\mathcal{D}|} x_{j|i=a}$$

$$\hat{\sigma}_{i=a|c=k} = \sqrt{\frac{1}{N[c=k]-1} \sum_{j=1}^{|\mathcal{D}|} \left(x_{j|i=a} - \hat{\mu}_{i=a|c=k}\right)^2},$$

gdzie $N[c=k]$ to liczba przykładów z klasy k, $|D|$ to liczba przykładów w zbiorze treningowym, a $x_{j|i=a}$ to wartość atrybutu a dla j-tego przykładu.

Prawdopodobieństwo, że nowa próbka $\boldsymbol{x}$ należy o klasy y określa się wzorem:

$$P(y \mid x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y).$$

$P(y)$ to prawdopodobieństwo wystąpienia danej klasy, liczone jako:

$$P(y = k) = \frac{N[c = k]}{|D|},$$

czyli ile razy dana klasa występuje w zbiorze treningowym. Natomiast do wyliczenia prawdopodobieństwo wartości atrybutu pod warunkiem określonej klasy określa się wzorem na rozkład normalny:

$$P\left(x_{i=a \mid c=k}\right) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{i=a|c=k}}} \exp\left(-\frac{\left(x_i - \hat{\mu}_{i=a|c=k}\right)^2}{2\hat{\sigma}^2_{i=a|c=k}}\right).$$

Waszym zadaniem jest:

- zaimplementować klasyfikator Gaussowskiego Naiwnego Bayesa,
- wnioskowanie dla tego klasyfikatora (predykcja do jakiej klasy należy nowa próbka) oraz uczenie parametrów.

Sam algorytm nie posiada żadnych dodatkowych hiperparametrów, które wymagają strojenia i przepadania.

W sprawozdaniu należy:

- umieścić tabelę z wynikami miar jakości klasyfikacji (accuracy, precision, recall, F1) dla zbioru danych iris[1] (dostępnego w pakiecie sklearn.datasets),
- porównać wyniki z klasyfikatorem drzewa decyzyjnego oraz SVM z zadania 4. (proszę wybrać parametry, dla których osiągaliście najlepszego wyniki).

Do badania jakości klasyfikacji proszę użyć walidacji krzyżowej z 5 podziałami z biblioteki sklearn. Należy podać średnią wartość miar jakości klasyfikacji oraz odchylenie standardowe (oznaczane jako średnia $\pm$ odchylenie). Do podziału można wykorzystać funkcję cross_val_score z biblioteki sklearn.model_selection (będzie to wymagało od Państwa odziedziczenie klasy Estimator z sklearn[2], żeby zadziałało—jest to

[1] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html

[2] https://scikit-learn.org/stable/developers/develop.html

swoją drogą bardzo użyteczne później w praktyce). Proszę zadbać o to, żeby był wykorzystany podział stratyfikowany (StratifiedKFold) (spójrzcie na dokumentację).