Pytanie **1**

Odpowiedź zapisana

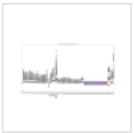
Punkty maks.: 0,80

- Przypomnij sobie z wykładu, w jaki sposób możemy łączyć klasyfikatory ze sobą (kilka architektur) oraz co jest niezbędne (jakie warunki muszą być spełnione) do tego, żeby takie połączenia działały skuteczniej od ich elementów składowych.
- Nawiązując do informacji z wykładu przeczytaj dokumentację pakietu scikit-learn na temat metod zespołowych w klasyfikacji (pomiń regresję; skup się tylko na BaggingClassifier, RandomForestClassifier, AdaBoostClassifier, VotingClassifier i StackingClassifier).
- Pobierz zbiór danych o nazwie odpowiadającej Twojemu numerowi albumu i przeprowadź jego wstępną eksplorację: proporcja klas, liczba i rodzaje atrybutów, ich zakresy i rozkłady wartości. Pokaż rozkłady wartości wszystkich atrybutów obok siebie na jednym szerokim wykresie pudełkowym lub skrzypcowym; na osi poziomej umieść nazwy atrybutów. Opisując wnioski (wystarczy kilka zdań) możesz pogrupować (o ile to możliwe) atrybuty pisząc np. "73 atrybuty są takie a takie, 22 atrybuty charakteryzują się tym a tym, wyjątkowy jest atrybut taki a taki", itp.

W zbiorze znajduje się 125 atrybutów warunkowych przyjmujących wartości ciągłych. Wartości przyjmowane przez większość tych atrybutów znajdują się w zakresie <0; 1>. Wyjątkowymi atrybutami są 'diffminus', który przyjmuje tylko wartości ujemne z przedziału <-0,4; 0) oraz 'stat76', który jako jedyny posiada wartość średniej oraz poszczególnych kwartyli większy od 1. Poza tymi atrybutami można wyróżnić pewne grupy atrybutów o podobnych rozkładach, takie jak: 'stat32' - 'stat35' (4 atrybuty) o rozkładzie wartości <0,02; 0,36>, średniej 0,23 i odchyleniu standardowym 0,05, 'stat36' - 'stat67' (32 atrybuty) o rozkładzie wartości <-0,14; 0,14>, średniej 0 i odchyleniu standardowym 0,1, 'stat68' - 'stat70' (3 atrybuty) o rozkładzie wartości <-0,16; 1>, średniej 0,28 i odchyleniu standardowym 0,18, 'stat72' - 'stat74' o rozkładzie wartości <-0,06; 0,79>, średniej 0,065 i odchyleniu standardowym 0,11. Atrybut decyzyjny przyjmuje wartości '0' lub '1', z czego aż 95,57% to '0' - dane są mocno niezbalansowane.

Maksymalny rozmiar dla nowych plików: 1GB

Pliki



wykres_pudel...

Pytanie **2**

Odpowiedź zapisana

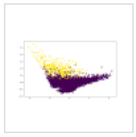
Punkty maks.: 0,80

Stosując wiedzę nabytą na wcześniejszym laboratorium z transformacji przestrzeni atrybutów, zwizualizuj ten zbiór w 2D i 3D podając procent wariancji zachowany przy rzutowaniu oryginalnej przestrzeni do 2D i 3D. Na wykresach pokaż przypadki obu klas jako kropki o dwóch różnych kolorach. Rozwiąż ewentualny problem zasłaniania kropek, na czym może cierpieć mniej liczna klasa. Czy na podstawie tej wizualizacji możesz dostrzec pole do współpracy różnych klasyfikatorów – czy pewne fragmenty przestrzeni wydają się trudne dla jednych metod, a proste dla innych? Czy ten rodzaj wizualizacji uprawnia do wyciągania tego typu wniosków?

Procent wariancji zachowany przy rzutowaniu oryginalnej przestrzeni do 2D: 42,11%
Procent wariancji zachowany przy rzutowaniu oryginalnej przestrzeni do 3D: 48,64%
Można zauważyć na wizualizacji pewne obszary, które będą łatwiejsze dla pewnych specyficznych metod. Z uwagi na niski procent zachowanej wariancji i utratę informacji ten rodzaj wizualizacji nie uprawnia do wyciągania tego typu wniosków.

Maksymalny rozmiar dla nowych plików: 1GB

Pliki



wykres_pca_2...

wykres_pca_3...

Pytanie **3**

Odpowiedź zapisana

Punkty maks.: 0,40

Przejrzyj dokumentację RandomForestClassifier, AdaBoostClassifier, VotingClassifier, StackingClassifier. Które z nich mają parametr `n_estimators`? Czym on się różni od parametru `estimators` oraz `base_estimator`? Pamiętaj o różnorodnych, poznanych do tej pory klasyfikatorach – DecisionTreeClassifier, SVC, MLPClassifier, GaussianNB i QuadraticDiscriminantAnalysis.

Parametr n_estimators posiadają klasyfikatory RandomForestClassifier, AdaBoostClassifier. Parametr n_estimators określa liczbę klasyfikatorów base_estimator wykorzystanych do zbudowania zespołu klasyfikatorów - estimators.

Pytanie **4**

Odpowiedź zapisana

Punkty maks.: 1,00

Postaraj się uzyskać na swoim zbiorze danych jak najwyższą trafność klasyfikacji za pomocą czterech wymienionych w poprzednim punkcie metod zespołowych. Używaj oryginalnych (niezmienionych) atrybutów. W przypadku AdaBoost, Voting i Stacking poeksperymentuj z różnymi zestawami bazowych klasyfikatorów (przynajmniej dwie próby na jedną architekturę) – to może być iteracyjna praca, polegająca na odkrywaniu, które klasyfikatory nawzajem sobie pomagają podnosząc jakość klasyfikacji. Jakość klasyfikacji oceniaj za pomocą G-mean techniką 10-fold stratified CV. Opisz przeprowadzone próby i wyciągnij wnioski.

Eksperymenty zostały przeprowadzone dla następujących architektur:
- AdaBoost z bazowymi klasyfikatorami: GaussianNB i DecisionTreeClasifier
- VotingClassifier z zestawami klasyfikatorów: [DecisionTree, SVC], [SVC, GaussianNB], [DecisionTree, GaussianNB, SVC], [DecisionTree, SVC, GaussianNB]
- StackingClassifier z zestawami klasyfikatorów: [DecisionTree, SVC], [SVC, GaussianNB], [DecisionTree, GaussianNB, SVC], [DecisionTree, SVC, GaussianNB] i domyślnym końcowym klasyfikatorem (LogisticRegression)
- RandomForestClassifier

Zdecydowana większość architektur osiągnęła średnią wartość G-mean powyżej 90%. Wariant VotingClassifier z zestawem bazowych klasyfikatorów DecisionTree i SVC osiągnął tę wartość nieco niższą. Metodą zespołową, która wyróżnia się wyraźnie i osiągnęła najsłabsze wyniki była AdaBoost z klasyfikatorem GaussianNB - 59,5% średnia wartość G-mean przy odchyleniu standardowym wynoszącym aż

Maksymalny rozmiar dla nowych plików: 1GB

Pliki

wykres_gmea...

Pytanie **5**

Odpowiedź zapisana

Punkty maks.: 0,80

Powtórz cały poprzedni eksperyment (poszukiwanie najwyższej jakości) jeszcze raz (tym razem bez eksperymentowania z różnymi zestawami bazowych klasyfikatorów), aby porównać wyniki (robiąc np. wykres różnic) wykorzystania oryginalnych atrybutów oraz atrybutów znormalizowanych. Wybierz jedną metodę normalizacji i uzasadnij, dlaczego taką wybrałe/aś. Zwróć uwagę, jak należy podejść do skalowania, kiedy mamy zbiór uczący i testujący, i nie wolno nam "dotykać" zbioru testowego podczas uczenia.

Jako metodę normalizacji wybrałem StandardScaler ze względu na założenie/wymaganie dla klasyfikatora GaussianNB, jak i większości problemów ML, że rozkład wartości atrybutów przypomina rozkład normalny o średniej 0 i jednostkową wariancją. Standaryzacja zgodnie z oczekiwaniami poprawiła uzyskiwane rezultaty, zwłaszcza dla wariantu AdaBoost z klasyfikatorem bazowym GaussianNB.

Maksymalny rozmiar dla nowych plików: 1GB

Pliki

wykres_poro...

Pytanie **6**

Nie udzielono odpowiedzi

Punkty maks.: 0,90

Wybierz najbardziej obiecującą złożoną architekturę i spróbuj dostroić jej parametry – zarówno samej architektury, jak i klasyfikatorów bazowych. Możesz się częściowo wspomóc znanym już GridSearchCV. Ponieważ klasy nie są zbalansowane, sprawdź, czy użycie `class_weight='balanced'` przynosi poprawę. Ile wynosi zysk z dostrojenia parametrów w porównaniu do rozwiązania początkowego?

Najlepszym rozpatrywanym klasyfikatorem złożonym okazał się klasyfikator VotingClassifier z klasyfikatorami bazowymi DecisionTree, GaussianNB oraz SVC. Użycie class_weight='balanced' przynosi niewielką poprawę. Zysk z odpowiedniego dostrojenia parametrów wyniósł 1,63%, gdzie początkowa wartość G-mean wynosiła 94,18%.

Maksymalny rozmiar dla nowych plików: 1GB

Pliki

wykres_poro...

Pytanie **7**

Nie udzielono odpowiedzi

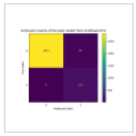
Punkty maks.: 0,70

Obejrzyj macierze pomyłek dla najlepszych uzyskanych zespołów klasyfikatorów i podsumuj wnioski: Jaki był wpływ normalizacji? Jakie architektury i ich parametry dały najlepsze G-mean i ile ono wyniosło? Czy widzisz zysk ze współpracy klasyfikatorów (w porównaniu do pojedynczego klasyfikatora) i ich uzupełniające się kompetencje? Jak duże są odchylenia standardowe wartości zwracanych przez 10-fold stratified CV i czy różnice w jakościach porównywanych klasyfikatorów są istotne?

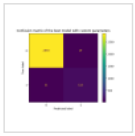
Dzięki normalizacji większość klasyfikatorów była w stanie uzyskać lepsze wyniki. Zwłaszcza jest to widoczne dla wariantu klasyfikatora AdaBoost z klasyfikatorem bazowym GaussianNB, który zakłada rozkład normalny poszczególnych atrybutów. Najlepsze G-mean uzyskały architektury wykorzystujące zespoły klasyfikatorów bazowych do uzyskania końcowej predykcji - VotingClassifier oraz StackingClassifier. Uzyskana średnia wartość G-mean przekroczyła w obu przypadkach 92% wraz z dowolną kombinacją następujących klasyfikatorów bazowych: SVC, DecisionTree, GaussianNb. Zysk wynikający ze współpracy klasyfikatorów jest zauważalny. Poszczególne klasyfikatory wzajemnie się uzupełniają, redukując swoje słabe strony oraz ryzyko overfittingu. Odchylenia standardowe dla prawie wszystkich przykładów (z wyjątkiem AdaBoost w połączeniu z GaussianNB, gdzie odchylenie wynosi 23%) wynoszą ok. 3%. Różnice w jakościach porównywanych klasyfikatorów nie są istotne ze względu na ich zbliżone wartości, z drobnym wyjątkiem w przypadku wariantu AdaBoost wraz z klasyfikatorem bazowym GaussianNB, gdzie średnia wartość G-mean wyniosła zaledwie 65% więc jest to klasyfikator dla rozważanego przypadku gorszy od pozostałych.

Maksymalny rozmiar dla nowych plików: 1GB

Pliki



macierz_pom...



macierz_pom...