Techniki oceny klasyfikacji
SK.UMA.7
Dokumentacja wstępna

Projekt realizowany w ramach przedmiotu
Uczenia maszynowego
autorstwa Karoliny Romanowskiej i Marianny Gromadzkiej
kierunek: Informatyka
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Spis treści

1 Wstęp

W ramach projektu przedmiotu Uczenie maszynowe został nam przydzielony poniżyszy temat:

> Zaimplementuj techniki oceny klasyfikacji dla zestawów danych dotyczących raka piersi, które dostępne są w http://archive.ics.uci.edu/ml/datasets

2 Interpretacja problemu

Skupiamy się na różnych metodach oceny klasyfikatorów, dlatego zaimplementujemy jedenaście różnych algorytmów za pomocą biblioteki scikit-learn. Jakość predykcji każdego z nich zostanie poddana analizie na podstawie metryk dla klasyfikatorów binarnych.

3 Zbióry danych

3.1 Wybrany zbiór

Wybrałyśmy zbiór Breast Cancer Wisconsin (Dataset) Data Set składający się z 569 rekordów o 30 atrybutach w postaci liczb rzeczywistych. 357 przypadów to przypadki łagodne, 212 złośliwe.

3.2 Atrybuty

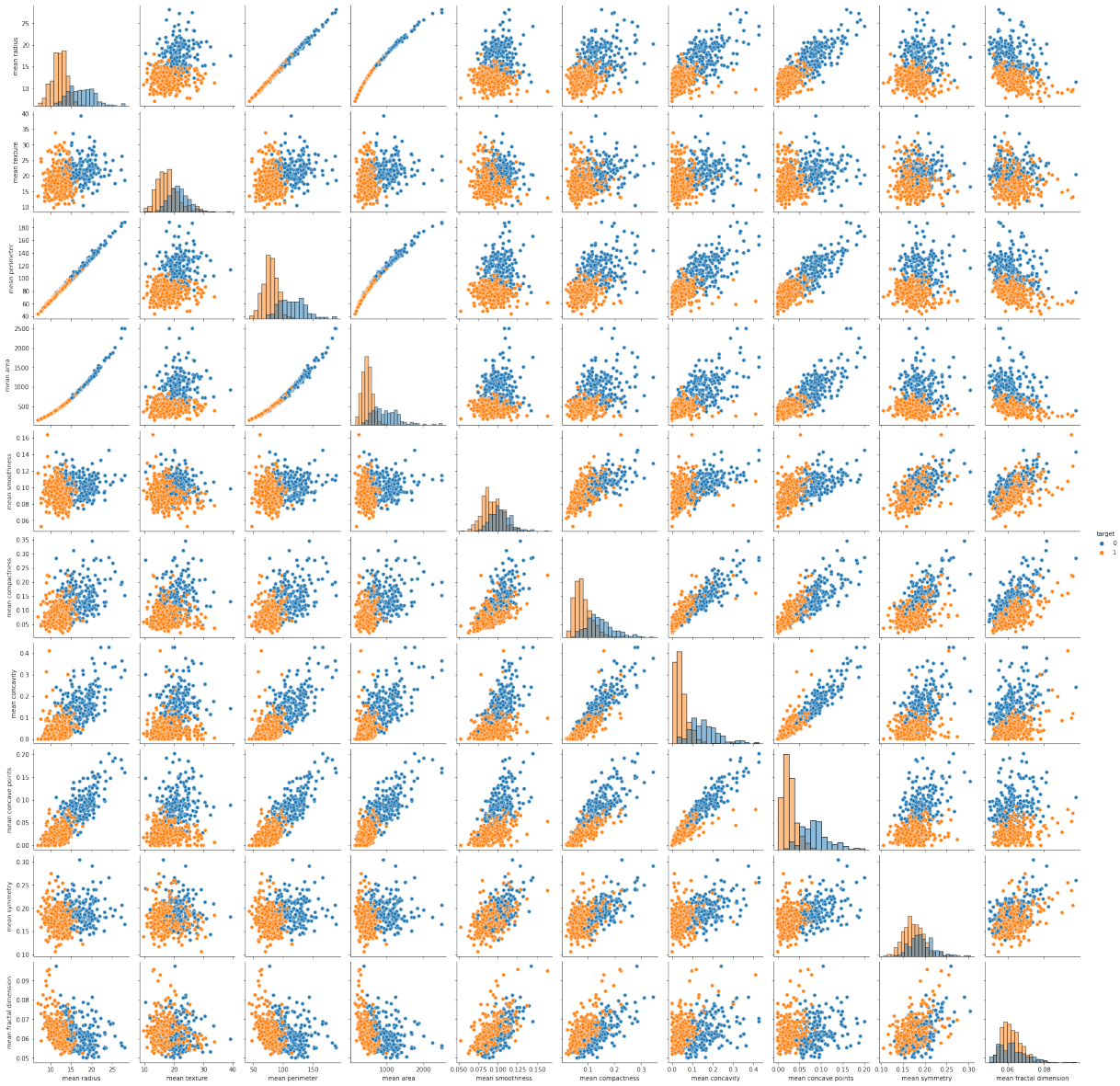
Atrybuty opisują właściwości jąder komórkowych widocznych na zdjęciach z biopsji:

- promień
- tekstura
- obwód
- powierzchnia
- gładkość (lokalna zmienność długości promieni)
- wklęsłość (dotkliwość wklęsłych fragmentów konturu)
- punkty wklęsłe (liczba wklęsłych części konturu)
- symetria
- wymiar podobieństwa
- ścisłość

Dla każdej właściwości podana jest średnia, błąd standardowy oraz "najgorsze przypadki" (średnia trzech największych wartości dla danej właściwości).

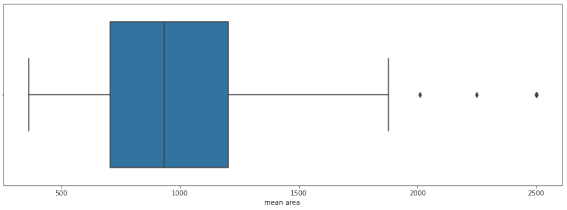
3.3 Analiza danych

Na wykresach przedstawiłyśmy histogramy (na przekątnej) średnich wartości właściwości komórek oraz pairploty. Można zauważyć, że dla tekstury, symetrii oraz wymiaru podobieństwa histogramy pokrywają się i klasy są trudne do rozdzielenia.

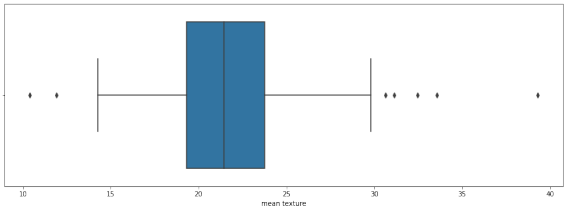


Rysunek 1: Legenda: pomarańczowy -przypadek łagodny, niebieski - przypadek złośliwy

W obu klasach w wielu atrybutach możemy zaobserwować górne wartości ekstremalne, np.:



Rysunek 2: Przypadek złośliwy



Rysunek 3: Przypadek łagodny

3.4 Wnioski

Z powodu za dużego podobieństwa w wartościach między klasami odrzucimy atrybuty: tekstura, symetria i wymiar podobieństwa. Odrzucone też zostaną przypadki w których wartości atrybutów są odstające.

3.5 Zbiór trenujący i testowy

Dataset zostanie podzielony na zbiór testowy (50%), walidacyjny (20%) i trenujący (30%).

4 Badane algorytmy

Algorytm	Opis
Maszyna wektorów nośnych	Ma na celu wyznaczenie hiperpłaszczyzny rozdzielającej przykłady z klas z maksymalnym marginesem
Perceptron wielowarstwowy	Sztuczna sieć neuronowa składająca się z warstw pojedynczych neuronów
K najbliższych sąsiadów	Wyznacza k badanych sąsiadów, do których badany element ma najbliżej dla wybranej metryki
Klasyfikacja procesem gaussowskim	Wykorzystuje leniwe uczenie się i miarę podobieństwa między punktami (funkcję jądra), aby przewidzieć wartość niewidocznego punktu na podstawie danych uczących
Drzewo decyzyjne	Etykietowane drzewo, w którym każdy węzeł wewnętrzny odpowiada przeprowadzeniu pewnego testu na wartościach atrybutów i wyznacza optymalny punkt podziału
ExtraTree	Etykietowane drzewo, w którym każdy węzeł wewnętrzny odpowiada przeprowadzeniu pewnego testu na wartościach atrybutów i wyznacza losowy punkt podziału
Las losowy	Polega na konstruowaniu wielu drzew decyzyjnych w czasie uczenia i generowaniu klasy, która jest dominantą klas (klasyfikacja) lub przewidywaną średnią (regresja) poszczególnych drzew
Klasyfikator AdaBoost	Silny klasyfikator, łącząc wiele klasyfikatorów o słabej wydajności. Podstawową koncepcją Adaboost jest ustawienie wag klasyfikatorów i próbek danych uczących w każdej iteracji, tak aby zapewnić dokładne przewidywania nietypowych obserwacji
Naiwny Klasyfikator Bayesowski	Prosty klasyfikator probabilistyczny oparty na założeniu o wzajemnej niezależności zmiennych objaśniających.
Kwadratowa analiza dyskryminacyjna	Jej zadaniem jest rozstrzyganie, które zmienne niezależne w najlepszy sposób dzielą dany zbiór przypadków na występujące w naturalny sposób grupy, opisane jakościową zmienną zależną.
Regresja logistyczna	Jest szczególnym przypadkiem uogólnionego modelu liniowego, znajdującym zastosowanie, gdy zmienna zależna jest dychotomiczna.

Tabela 1: Opisy rozważanych algorytmów

5 Wykorzystane metryki

5.1 Tablica pomyłek

Tablica stosowana podczas oceny klasyfikacji binarnej. Wykorzystywane są tutaj etykiety przypisane przez model oraz klasy rzeczywiste. W tablicy można znaleźć ilość próbek prawdziwie pozytywnych - tp, fałszywie pozytywnych - fp, fałszywie negatywnych - fn i prawdziwie negatywnych - tn. Oś Y przedstawia klasy rzeczywiste, oś X przedstawia klasy przewidywane.

	Klasyfikacja pozytywna	Klasyfikacja negatywna
Stan pozytywny	Prawdziwie dodatnia - TP	Fałszywie ujemna - FN
Stan negatywny	Fałszywie dodatnia - FP	Prawdziwie ujemna - TN

Tabela 2: Tablica pomyłek

5.2 Błąd pierwszego rodzaju — Miara fałszywie pozytywna

Błąd pierwszego rodzaju przedstawia ilość przypadków, w których odrzucona została hipoteza zerowa, pomimo jej prawdziwości. Często stosuje się tę metrykę jako pomocniczą do innej metryki, zazwyczaj do pary ze wskaźnikiem fałszywych odkryć. Można ją interpretować jako częstość fałszywych alarmów

$$FPR = \frac{fp}{fp + tn} \tag{1}$$

5.3 Błąd drugiego rodzaju — Miara fałszywie negatywna

Błąd drugiego rodzaju przedstawia ilość przypadków, w których hipoteza zerowa nie została odrzucona, pomimo jej fałszywości. Często stosuje się tę metrykę jako pomocniczą do innej metryki.

$$FNR = \frac{fn}{tp + fn} \tag{2}$$

5.4 Swoistość — Odsetek prawdziwie negatywnych

Swoistość przedstawia stosunek ilości wyników prawdziwie negatywnych do sumy prawdziwie negatywnych oraz fałszywie pozytywnych. Przedstawia w jakim procencie klasa będąca w rzeczywistości negatywną, została przewidziana jako negatywna. Wraz z czułością metryka wskazuje poziom dokładności predykcji. Używana dla potwierdzenia poprawności predykcji. Warto z niej skorzystać gdy w naszych przewidywaniach chemy uchwycić jak najwięcej przypadków negatywnych.

$$TNR = \frac{tn}{fp + tn} \tag{3}$$

5.5 Czułość — Odsetek prawdziwie pozytywnych

Czułość przedstawia stosunek ilości wyników prawdziwie pozytywnych do sumy prawdziwie pozytywnych oraz fałszywie negatywnych. Można ją interpretować jako prawdopodobieństwo poprawności klasyfikacji pod warunkiem rozpatrywania przypadku pozytywnego. Wraz ze swoistością metryka wskazuje poziom dokładności predykcji. Warto z niej skorzystać gdy w naszych przewidywaniach chcemy uchwycić jak najwięcej przypadków pozytywnych.

$$TPR = \frac{tp}{tp + fn} \tag{4}$$

5.6 Precyzja — Wartość predykcyjna dodatnia

Precyzja przedstawia stosunek ilości wyników prawdziwie pozytywnych do sumy prawdziwie pozytywnych oraz fałszywie pozytywnych. Metryka ta mierzy ile predykcji pozytywnych jest w rzeczywistości pozytywne. Stosowana gdy zależy nam na wysokiej pewności w stosunku do prawdziwości naszych predykcji

$$PPV = \frac{tp}{tp + fp} \tag{5}$$

5.7 Wartość predykcyjna ujemna

Wartość predykcyjna ujemna przedstawia stosunek ilości wyników prawdziwie negatywnych do sumy prawdziwie negatywnych i fałszywie negatywnych. Metryka ta mierzy ilość prawdziwie negatwnych przypadków która została zakwalifikowana poprawnie. Można ją interpretować jako precyzję wyznaczania klas negatywnych.

$$NPV = \frac{tn}{tn + fn} \tag{6}$$

5.8 Wskaźnik fałszywych odkryć

Wskaźnik fałszywych odkryć mierzy ile predykcji ze wszystkich pozytywnych było w rzeczywistości fałszywymi. Można ją interpretować jako częstość fałszywych odkryć. Często podaje się tę miarę wraz z Miarą fałszywie pozytywną.

$$FDR = \frac{fp}{fp + tp} \tag{7}$$

5.9 Dokładność

Dokładność mierzy ile predykcji, zarówno pozytywnych jak i negatywnych, była poprawnie zaklasyfikowana. Nie należy używać tej miary przy niezbalansowanych zbiorach danych dla problemu - wtedy łatwo uzyskać wysoką dokładność, poprzez klasyfikacje wszystkich obserwacji do klasy większościowej. Metryki warto używać gdy każda z klas jest dla nas równie ważna.

$$ACC = \frac{tp + tn}{tp + tn + fp + fn} \tag{8}$$

5.10 Współczynnik korelacji Matthews'a

Współczynnik korelacji Matthews'a to korelacja między przewidywaną klasą a obserwowaną. Przyjmuje wartości w zakresie $[-1, 1]$, gdzie wartość 1 - oznacza klasyfikację bez pomyłek, 0 - jakość klasyfikacji równą klasyfikatorowi losowemu i 1 - klasyfikator zawsze zwracający błędne wyniki. Współczynnik uważany jest za zrównoważoną miarę - oznacza to, że znajduje swoje zastosowanie, również gdy nasze dane są niezbalansowane pod względem przynależności do klas.

$$MCC = \frac{tp * tn - fp * fn}{(tp + fp)(tp + fn)(tn + fp)(tn + fn)} \tag{9}$$

5.11 Wskaźnik F1

Wskaźnik F1 zdefiniowana jakoś średnia harmoniczna z precyzji i czułości. Średnia harmoniczna, w porównaniu do średniej arytmetycznej, jest nie czuła na wysokie wartości pojedynczych parametrów. Z tego powodu wykorzystywana jest gdy chcemy sprawdzić czy istnieje balans między precyzją a czułością w naszym modelu.

$$F_1 = \frac{2 * PPV * TPR}{PPV + TPR} \tag{10}$$

5.12 Wskaźnik F-Beta

Wskaźnik F-Beta pozwala na dostosowanie metryki do naszych potrzeb. Im bardziej zależy nam na czułości, tym wyższe beta powinniśmy zastosować. Parametr ten przyjmuje zazwyczaj wartości od 0.1 do 10. Szczególnie często przyjmuje wartości 0.5 - gdy zależy nam bardziej na precyzji i 2 - gdy zależy nam na czułości.

$$F_{beta} = (1 + \beta^2)\frac{PPV * TPR}{\beta^2 * PPV + TPR} \tag{11}$$

5.13 Strata logarytmiczna — Strata entropii krzyżowej — Strata logistyczna

Metryka ta jest używana gdy wyjściem klasyfikatora jest prawdopodobieństwo przewidywania klasy. Przyjmuje wartości w zakresie $[0, \infty)$. Wartości bliskie 0 wskazują na dokładność predykcji. Dążymy do minimalizacji tego wskaźnika.

$$logloss_{(N=1)} = -\frac{1}{N}\sum_{i=1}^{N}[y_i * ln(p_i) + (1 - y_i) * ln(1 - p_i)] \tag{12}$$

5.14 Współczynnik Kappa Cochena

Współczynnik Kappa Cochena przedstawia o ile oceniany model jest lepszy od losowego klasyfikatora. Parametr p_e oznacza sumę prawdopodobieństwa przypadkowej zgodności prognozy z rzeczywistymi wartościami klasy pierwszej oraz prawdopodobieństwo przypadkowej zgodności z rzeczywistymi wartościami klasy drugiej. Współczynnik ten przyjmuje wartości w zakresie $[-1, 1]$. Współczynnik ten wyniesie niższe wartości dla danych niezrównoważonych.

$$\kappa = \frac{ACC - p_e}{1 - p_e} \tag{13}$$

5.15 Krzywa ROC

Jest to wykres który obrazuje zależność między metrykami TPR (czułością) i FPR (błędem pierwszego rodzaju). Dla każdego progu wyliczamy miary i wykreślamy je na wykresie. Krzywa ROC pozwala na szybkie porównanie ze sobą klasyfikatorów oraz wyboru dla nich progu odcięcia klas. Dążymy do uzyskania wysokiej wartości TPR i niskkiej FPR.

5.16 Wskaźnik ROC AUC

Wskaźnik ROC AUC przedstawia wartość pola powierzchni pod krzywą ROC. Jest to prawdopodobieństwo, że oceniany model klasyfikacji binarnej przydzieli wyższe prawdopodobieństwo losowo wybranemu przypadkowi pozytywnemu, niż losowo wybranemu przypadkowi negatywnemu. Chcemy maksymalizować tą metrykę tak aby nasza krzywa ROC sięgała jak najbliżej lewemu górnemu rogu wykresu.

5.17 Wynik Briera

Wynik Briera to metryka służąca do sprawdzenia poprawności przewidywanego wyniku. Jest bardzo podobna do błędu średniokwadratowego, ale można ją zastosować jedynie dla prawdopodobieństwa wyniku z modelu będącego w zakresie $[0, 1]$ Metryka używana do kalibracji prawdopodobieństwa modeli uczenia maszynowego. Stanowi dobre uzupełnienie metryki ROC AUC.

$$brierloss = (y_{pred} - y_{true})^2 \tag{14}$$

5.18 Krzywa PR

Krzywa PR obrazuje zależności między metrykami PPV i TPR. Pomaga dobrać próg odcięcia klas.

5.19 Wskaźnik PR AUC

Podobnie jak ROC AUC, metryka PR AUC przedstawia wartość pola powierzchni pod krzywą PR. Ta metryka może być również interpretowana jako średnia z precyzji obliczana dla każdej wartości czułości.

5.20 Wykres skumulowanego zysku

Wykres skumulowanego zysku pozwala ocenić, jak duży będzie zysk z wykorzystania modelu klasyfikacji w porównaniu z modelem losowym dla danej miary najwyżej ocenionych predykcji. Należy uporządkować predykcje od najwyższych do najniższych i dla każdego percentyla wyliczana jest miara prawdziwych pozytywów. Wykres pozwala dostrzec korzyści wynikające z zastosowania klasyfikatora.

5.21 Kolmogorov-Smirnov plot

Wykres Kolmogorova-Smirnova przedstawia dystans pomiędzy proporcją sukcesu:

$$\frac{liczba\ pozytywow\ w\ danym\ percentylu}{liczba\ wszystkich\ pozytywow} \tag{15}$$

a proporcją porażki:

$$\frac{liczba\ negatywow\ w\ danym\ percentylu}{liczba\ wszystkich\ pozytywow} \tag{16}$$

wyliczaną dla kolejnych percentyli uporządkowanych malejąco predykcji.

5.22 Kolmogorov Smirnov statistics

Jest to największy dystans pomiędzy wykresami Kolmorova-Smirnova.Przydatny gdy zależy nam równo na klasach pozytywnych co negatywnych.

6 Bibliografia

Klasyfikacja procesem gaussowskim
- Regresja logistyczna
- Maszyna wektorów nośnych
- Drzewa decyzyjne
- Las losowy i extra tree
- Klasyfikator AdaBoost
- Naiwny klasyfikator bayesowski
- Analiza dyskryminacyjna
- Regresja logistyczna