WSI – Raport z zadania 4

Michał Pędziwiatr

Numer indeksu: 331 421

Zamysł algorytmu

Model regresji logistycznej jest powszechnie używaną techniką do przewidywania wyników mających dwie możliwe odpowiedzi, obliczając prawdopodobieństwo na wystąpienie danej z nich. Dzieje się to, dzięki algorytmowi „gradient descent”, opisanego w raporcie z zadania 1, dzięki któremu, operując na próbce uczącej zawierającej prawdziwe wartości, możemy zminimalizować różnicę pomiędzy wartością prawdziwą a naszymi przewidywaniami. Te natomiast, generowane są dzięki funkcji przypisującej każdej z wartości wejściowych parametr wagi, decydujący o wpływie jej wartości na ostateczne przewidywania.

W naszym przypadku wartości wejściowe są różnymi pomiarami guza piersi. Dobierając odpowiednie wagi tychże wartości, nasz model przewidywać będzie czy guz ten jest złośliwy, czy łagodny.

Wyniki algorytmu – brak ingerencji w dane

Iteracje	Krok	Dokładność	F1	AUROC	Normalizacja	Usunięte kolumny
100	0,000001	87,41%	0,88	0,97	nie	brak
300	0,000001	90,21%	0,89	0,95	nie	brak
500	0,000001	89,51%	0,88	0,95	nie	brak
1000	0,000001	90,21%	0,88	0,95	nie	brak
100	0,00001	78,32%	0,69	0,87	nie	brak
300	0,00001	89,51%	0,88	0,97	nie	brak
500	0,00001	90,21%	0,89	0,97	nie	brak
1000	0,00001	91,61%	0,91	0,98	nie	brak
100	0,0001	46,15%	0,63	0,94	nie	brak
300	0,0001	88,11%	0,88	0,98	nie	brak
500	0,0001	86,01%	0,87	0,98	nie	brak
1000	0,0001	88,81%	0,89	0,98	nie	brak
100	0,001	46,15%	0,63	0,51	nie	brak
300	0,001	81,12%	0,83	0,86	nie	brak
500	0,001	90,21%	0,9	0,92	nie	brak
1000	0,001	61,54%	0,71	0,66	nie	brak
100	0,01	46,15%	0,63	0,5	nie	brak
300	0,01	72,03%	0,77	0,75	nie	brak
500	0,01	90,21%	0,9	0,91	nie	brak
1000	0,01	90,21%	0,9	0,93	nie	brak
100	0,1	46,15%	0,63	0,5	nie	brak
300	0,1	88,11%	0,89	0,89	nie	brak
500	0,1	88,11%	0,88	0,89	nie	brak
1000	0,1	89,51%	0,89	0,9	nie	brak

(Krok – współczynnik kroku w algorytmie gradient descent,

Iteracje – liczba iteracji algorytmu gradient descent.

Wszystkie pomiary wykonywane są dla parametru seed równego 69, pomiary tabeli „usunięcie kolumn” wykonywane są dla parametrów liczby iteracji = 500, mnożnika długości kroku = 100 oraz znormalizowanych danych. W odniesieniu do tego też przypadku obliczano deltę – zmianę wartości)

Wyniki algorytmu – normalizacja danych

Iteracje	Krok	Dokładność	F1	AUROC	Normalizacja	Usunięte kolumny
100	0,0001	66,43%	0,45	0,83	tak	brak
300	0,0001	66,43%	0,45	0,83	tak	brak
500	0,0001	66,43%	0,45	0,83	tak	brak
1000	0,0001	67,83%	0,49	0,83	tak	brak
100	0,001	67,83%	0,49	0,83	tak	brak
300	0,001	69,93%	0,54	0,84	tak	brak
500	0,001	70,63%	0,55	0,84	tak	brak
1000	0,001	73,43%	0,61	0,85	tak	brak
100	0,01	73,43%	0,61	0,85	tak	brak
300	0,01	78,32%	0,72	0,87	tak	brak
500	0,01	79,02%	0,75	0,88	tak	brak
1000	0,01	81,12%	0,78	0,89	tak	brak
100	0,1	81,12%	0,78	0,89	tak	brak
300	0,1	84,62%	0,83	0,9	tak	brak
500	0,1	86,01%	0,85	0,91	tak	brak
1000	0,1	83,92%	0,83	0,92	tak	brak
100	1	83,92%	0,83	0,92	tak	brak
300	1	86,71%	0,86	0,93	tak	brak
500	1	88,11%	0,87	0,94	tak	brak
1000	1	88,11%	0,87	0,96	tak	brak
100	10	88,11%	0,87	0,96	tak	brak
300	10	89,51%	0,88	0,97	tak	brak
500	10	92,31%	0,92	0,98	tak	brak
1000	10	92,31%	0,92	0,98	tak	brak
100	100	87,41%	0,87	0,96	tak	brak
300	100	91,61%	0,92	0,98	tak	brak
500	100	93,01%	0,93	0,99	tak	brak
1000	100	93,01	0,93	0,99	tak	brak
100	1000	83,92%	0,85	0,87	tak	brak
300	1000	90,91%	0,91	0,94	tak	brak
500	1000	93,71%	0,93	0,96	tak	brak
1000	1000	94,41%	0,94	0,97	tak	brak
100	10000	86,71%	0,87	0,88	tak	brak
300	10000	92,31%	0,92	0,94	tak	brak
500	10000	93,71%	0,94	0,93	tak	brak
1000	10000	91,61%	0,92	0,92	tak	brak

Wyniki algorytmu – usunięcie kolumn

Dokładność	Δ Dokładności	F1	Δ F1	AUROC	Δ AUROC	Usunięte kolumny
90,91%	-2,10%	0,91	-0,02	0,98	-0,01	radiuses
90,91%	-2,10%	0,91	-0,02	0,98	-0,01	areas
91,61%	-1,40%	0,91	-0,02	0,97	-0,02	areas, radiuses
93,71%	0,70%	0,93	0	0,98	-0,01	areas, radiuses, textures, smoothnesses
86,01%	-7,00%	0,86	-0,07	0,98	-0,01	areas, radiuses, textures, smoothnesses, perimeters, compactnesses
91,61%	-1,40%	0,92	-0,01	0,99	0	perimeters, compactnesses
86,71%	-6,30%	0,86	-0,07	0,95	-0,04	wszystkie wartości kończące się cyfrą 1
90,91%	-2,10%	0,91	-0,02	0,99	0	wszystkie wartości kończące się cyfrą 2
90,21%	-2,80%	0,9	-0,03	0,97	-0,02	wszystkie wartości kończące się cyfrą 3
76,22%	-16,79%	0,71	-0,22	0,84	-0,15	wszystkie wartość i kończące się cyfrą 1 lub 2
81,12%	-11,89%	0,82	-0,11	0,9	-0,09	wszystkie wartości kończące się cyfrą 1 lub 3
85,31%	-7,70%	0,85	-0,08	0,95	-0,04	wszystkie wartości kończące się cyfrą 2 lub 3
92,31%	-0,70%	0,91	-0,02	0,99	0	textures, smoothnesses
89,51%	-3,50%	0,89	-0,04	0,98	-0,01	wszystkie wartości poza areas i radiuses
79,72%	-13,29%	0,79	-0,14	0,89	-0,1	wszystkie wartość i poza areas
79,02%	-13,99%	0,76	-0,17	0,84	-0,15	wszystkie wartości poza radiuses
54,55%	-38,46%	0,66	-0,27	0,84	-0,15	wszystkie wartości poza textures
53,85%	-39,16%	0	-0,93	0,63	-0,36	wszystkie wartości poza smoothness
58,74%	-34,27%	0,23	-0,7	0,56	-0,43	wszystkie wartości poza compactness
87,41%	-5,60%	0,87	-0,06	0,94	-0,05	wszystkie wartośći poza concavity
81,12%	-11,89%	0,77	-0,16	0,9	-0,09	wszystkie wartości poza concave_points
53,15%	-39,86%	0,65	-0,28	0,78	-0,21	wszystkie wartości poza symmetry
77,62%	-15,39%	0,78	-0,15	0,84	-0,15	wszystkie wartości poza fractal_dimension
82,52%	-10,49%	0,79	-0,14	0,93	-0,6	wszystkie wartości poza concave_points i concavity

Wnioski

Analizując dane pozyskane z pomiarów bez ingerencji w dane, zgodnie z przewidywaniami, poza jednym wyjątkiem, widzimy pozytywny wpływ większych liczb iteracji w algorytmie gradient descent na przewidywania algorytmu. Wyższe wartości iteracji pozwalają na dogłębniejsze zbadanie oraz naukę na podanym materiale treningowym. Źródło wspomnianej anomalii – dokładności 61,54% dla 1000 iteracji oraz 90,21% dla 500 iteracji dla współczynnika kroku równego 0,001 – ciężko jednoznacznie określić. Osobiście uważam, że przez fakt, iż operujemy na aż 30 różnych cechach, algorytm gradient descent może w określonych przypadkach wpadać w pewnego rodzaju pułapki, ze względu na to, że optymalizuje on funkcje biorąc pod uwagę wartość całkowitą funkcji, a nie indywidualnych 30 wartości. Drugą opcją, która niekoniecznie wyklucza pierwszą, może być także fakt, że dla tego niefortunnego przypadku algorytm wyuczył się materiału zbyt dobrze, doszukując się następnie znalezionych korelacji w materiale testowym, których tam nie zastał.

Tej samej korelacji ciężko doszukać się jednak w przypadku kolumny „Krok” - mnożnika długości kroku, w tym przypadku wartości wydają się bardziej indywidualnie wpływać na działanie programu i należy je dobrać do określonego celu (jakim może być maksymalizacja dokładności, wyniku f1 lub wartości AUROC).

Biorąc pod lupę pomiary po normalizacji, widzimy drastyczne różnice m.in. w optymalnych wartościach kroku, które tak jak można było się domyślać biorąc pod uwagę sposób obliczania gradientu, potrzebowały znaczącej inkrementacji by „nadrobić” przeskalowanie wartości cech. Zmiany te są jednak jeszcze większe niż się spodziewałem, doprowadzając do tego, że program przewiduje zadowalające (>90%) wartości nawet przy współczynniku kroku równym 10, 100, a nawet 1000 i 10000.

Poza zmianami w wartościach współczynnika kroku pomiary te pokazują także wzrost zarówno maksymalnych dokładności jak i AUROC oraz F1. Świadczyć to może o tym, że funkcja ucząca dużo szybciej oraz skuteczniej dobiera gradienty dla wag wartości mających podobne rzędy wielkości, niż dla zestawienia wartości równych kilka tysięcy razem z niewielkimi ułamkami.

Rozpatrując ostatnią, trzecią serię pomiarów, dotyczącą usuwania oraz wybierania poszczególnych kolumn, możemy zauważyć m.in. to, iż każda z cech, mimo rozbieżności w skali w jakiej to czyni, zdaje się korelować z tym czy guz można ocenić jako złośliwy lub nie. Wniosek ten wypływa w głównej mierze z faktu, że zostawiając każdy typ cech jako jedyne wartości wejściowych danych, wartość AUROC nigdy nie spadła do 0,5. Najmniejsza jej wartość wyniosła 0,56 dla kolumny compactness (zwartość guza), co oznacza minimalny, lecz jednak istniejący wpływ tej cechy na jego złośliwość. Po drugiej stronie „rankingu" widzimy natomiast zmienne areas (pola) i radiuses (promienie), które będąc jedynymi pozostawionymi cechami pozwoliły na AUROC wynoszący aż 0,98, oraz, mniej intuicyjnie, indywidualne wartości concavity (wklęsłość) i concave_points (liczba punktów wklęsłości). Dla tych wartości współczynnik AUROC wynosił odpowiednio 0,94 oraz 0,9, jednocześnie nie skutkując zbyt drastycznym spadkiem dokładności ani F1.

Kolejnym ciekawym i dość niespodziewanym spostrzeżeniem jest także fakt, że pozostawienie niektórych wartości, mimo pozytywnego wpływu na wyniki gdy pozostawiamy je „indywidualnie", może prowadzić do gorszych rezultatów gdy pozostawimy je razem. Mimo braku pewności co do przyczyny tego zjawiska, zakładam, że może być to m.in. bardzo widoczna korelacja tych cech pomiędzy sobą. Przykładem może być promień oraz pole czy wklęsłość i liczba punktów wklęsłości. Posiadanie dwóch cech, które w praktyce prawie zawsze będą ze sobą silnie korelować może więc prowadzić do błędnego oraz przesadnego dopasowania wyników do tychże wartości. Może to być również kwestia tego, że wartości te mają tak duży wpływ na to czy guz jest złośliwy, że dodanie drugiej zwyczajnie niepotrzebnie utrudnia optymalizację. Domyślam się także, że ze względu na to jak wartość „error" wpływa na wartości gradientów wag, zjawisko to zostałoby pogłębione znacząco gdy cechy te korelowałyby „odwrotnie" do przewidywania. Byłby to przypadek gdy dwie silnie powiązane cechy są do siebie wprost proporcjonalne, lecz takowych nie udało mi się znaleźć w danych na których operujemy w tym zadaniu.