Miniprojekt 2

Tomasz Mazur

1 Wprowadzenie

Poniższy raport ma na celu opisanie podjętych działań oraz otrzymanych wyników podczas doboru odpowiedniego modelu do problemu klasyfikacji otrzymanych danych. Modele brane pod uwagę to regresja logistyczna oraz naiwny klasyfikator bayesowski z wygładzeniem Laplace'a.

Warto zwrócić uwagę na interpretację klas, mianowicie jedna klasa oznacza stwierdzenie, że nowotwór jest złośliwy, a druga, że jest łagodny. Zależy nam zatem, aby model był czuły, gdyż błędy polegające na stwierdzeniu, że nowotwór jest łagodny, gdy w rzeczywistości jest on złośliwy są bardzo szkodliwe.

2 Podział i obróbka danych

Ze względu na wygodną postać danych nie była potrzeba na wprowadzenie wielu zmian. Jedyną istotnie ważną modyfikacją było zamiany wartości zmiennych objaśnianych y z 4 na 1 oraz 2 na 0. Oprócz tego w regresji logistycznej warto było standaryzować dane $x \to \frac{x-\mu}{\sigma}$, gdzie μ to średnia z kolumny oraz σ to standardowe odchylenie. Wtedy dane mają średnią 0, więc nie ma potrzeby dodawania kolumny jeydnek do zmiennych objaśniających oraz metoda gradientu stochastycznego zbiega szybciej ze względu na mniejsze wartości.

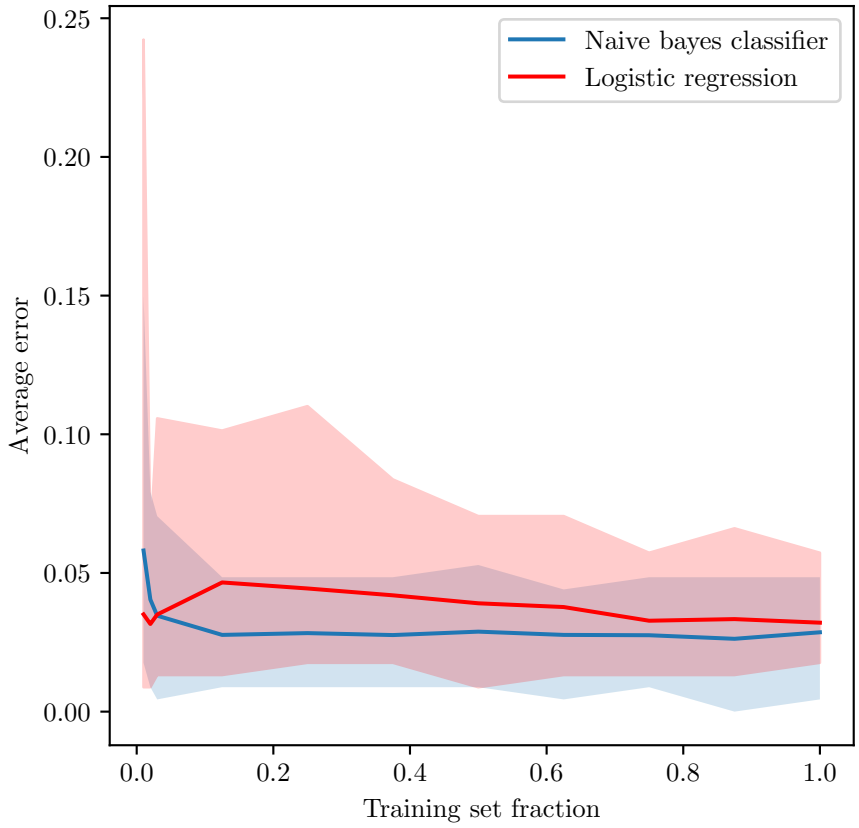
Zgodnie ze specyfikacją, zbiór danych został podzielony na zbiór treningowy i testowy w stosunku 2 : 1 w sposób losowy ze zwróceniem uwagi na to, by dane dla poszczególnych klas również były podzielone w stosunku 2 : 1.

3 Naiwny klasyfikator bayesowski

Wyniki otrzymane przy zastosowaniu naiwnego klasyfikatora bayesowskiego sugerują, że dane objaśniające nie są w istotny sposób zależne warunkowo. Poniżej znajduje się krzywa uczenia dla tego klasyfikatora, przy czym błąd to jest prawdopodobieństwo błędnej klasyfikacji, tzn. jest to

$$\frac{\text{liczba błędnych klasyfikacji na zbiorze testowym}}{\text{moc zbioru testowego}}$$

lub równoważnie 1−dokładność, przy czym ta wartość jest uśredniona po 50 losowych podziałach na zbiór treningowy i testowy. Co ciekawe, naiwny klasyfikator osiąga średnią czułość 98%.



Rysunek 1: Średni błąd dla różnych frakcji zbioru treningowego.

4 Regresja logistyczna

Przy regresji logistycznej została wykorzystana metoda gradientu stochastycznego, co może tłumaczyć zaskakujące wyniki. Dla bardzo małych frakcji zbioru treningowego (np. 0.01, 0.02) model spisuje się lepiej niż dla średnich (np. 0.375, 0.5) i dopiero przy dużych frakcjach wynik jest poprawiony. Może być to spowodowane lepszą zbieżnością metody gradientu stochastycznego dla mniejszych zbiorów.

Podobnie jak w naiwnym klasyfikatorze bayesowskim krzywa uczenia zawiera błąd uśredniony na 50 przebiegach z losowymi podziałami zbioru danych. W tym przypadku, w przeciwieństwie do naiwnego klasyfikatora bayesowskiego, zbieżność błędu jest raczej wolna i jest możliwe, że przy istotnie większym zbiorze treningowym błąd byłby również istotnie mniejszy.

Warto również wspomnieć, że regresja logistyczna osiągnęła średnią czułość 96%, co jest istotnie gorzej od naiwnego klasyfikatora bayesowskiego (2x częstsze błędy).

5 Postanowienia końcowe i wnioski

Ze względu na prostotę, wydajność oraz lepsze rezultaty, dla danego zbioru treningowego naiwny klasyfikator wydaje się lepszym wyborem. Warto jednak zwrócić uwagę, że w przeciwieństwie do naiwnego klasyfikatora bayesowskiego, regresja logistyczna poprawiła swój średni błąd pomiędzy zbiorami treningowymi o rozmiarze 200 a 400, co sugeruje, że przy większej liczbie danych można zobaczyć lepsze rezultaty, niż u naiwnego klasyfikatora bayesowskiego.

Rezultaty te są w pewnym stopniu zgodne z wynikami w pracy *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes* Andrew Nga i Michaela Jordana, gdyż jednym z głownych wiosków tej pracy jest szybsza zbieżność naiwnego klasyfikatora bayesowskiego w porównaniu z regresją logistyczną. Na danym zbiorze danych nie da się ocenić zgodności z wnioskiem o niższym progiem błędu przy regresji logistycznej.