Zaawansowane metody uczenia maszynowego
Projekt 1

Mikołaj Małkiński
malkinskim@student.mini.pw.edu.pl

20 Apr 2019

1 Wstęp

2 Przygotowanie danych

2.1 Brakujące dane

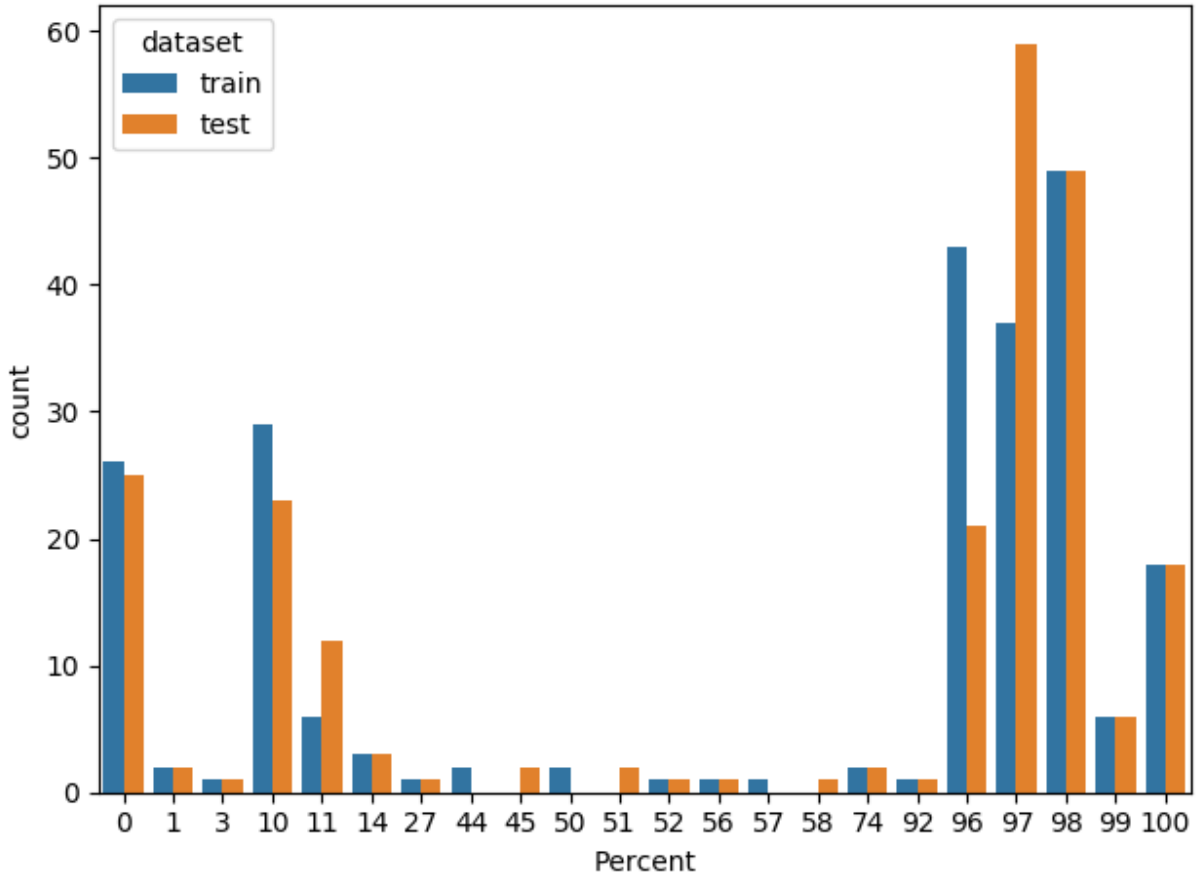
Zbiór danych zawiera wiele kolumn które nie są w pełni wypełnione danymi. Analiza wykazała, że tylko 20 kolumn posiadały wszystkie wartości. Wykres 1 przedstawia zależność między stopniem braku danych a liczbą kolumn. Istnieje kilka możliwych podejść które można zastosować w tym przypadku. Po pierwsze, można kompletnie zignorować i wybrać model który jest w stanie sam odpowiednio obsłużyć luki w zbiorze. Przykładem takiego modelu jest *XGBoost*. Jednakże, nie wszystkie algorytmy do klasyfikacji są przygotowane na braki w danych, Dlatego, aby móc porównać działanie kilku modeli na tym samym zbiorze podjęto decyzję o wypełnieniu brakujących danych.

Niektóre z kolumn posiadały nawet ponad 90% brakujących danych. Nie mając żadnych informacji o zbiorze danych oraz nie wiedząc co dana kolumna reprezentuje, ciężko stwierdzić z czego wynika taki duży brak. Może to oznaczać zwyczajnie brak pomiaru, wartość ważną równie dobrze jak ta która istnieje w danej kolumnie lub w danym przypadku wartość w tej kolumnie może nie mieć żadnego znaczenia dla konkretnego wiersza. Ze względu na duży rozmiar zbioru danych, podjęto decyzję o kompletnym usunięciu części z takich kolumn, które mają więcej braków niż dany procent. Resztę poddano imputacji.

Kolumny w zbiorze można podzielić na 2 rodzaje: *numeryczne* i *kategoryczne*. Aby wypełnić pierwszy z nich, brakujące dane w każdej kolumnie wypełniono jej medianą. Oczywiście w innych przypadkach mogłyby zostać także inne funkcje, takie jak średnia lub moda. W przypadku kolumn kategorycznych, dodano nową kategorię: *unknown*, która wypełniła braki w tych kolumnach.

2.2 Unikalność danych

Następnie, analizie poddano liczbę unikalnych wartości dla każdej z kolumn. Zbiór zawierał kolumny wypełnione tylko jedną tą samą wartością. Takie kolumny nie niosą ze sobą



Rysunek 1: Brakujące dane

żadnej wartości więc zostały one usunięte Dodatkowo, część kolumn kategorycznych, posiadały bardzo dużo unikalnych wartości. Można przypuszczać że są to dane tekstowe, któ©e także nie przyniosą pozytywnych efektów w klasyfikacji. Te kolumny zostały także usunięte.

2.3 Kolumny kategoryczne

Niektóre algorytmy klasyfikacji wymagają aby zbiory na których zostaną użyte posiadały tylko cechy numeryczne. Z tego powodu, cechy kategoryczne musiały zostać przekształcone w liczbowe. Najprostszym sposobem jest zamienienie każdej z kategorii na unikalną liczbę. Jednakże to implikuje pewien porządek w tej kolumnie, który tak na prawdę nie występuje. Dlatego każdą kolumnę przekształcono w n nowych kolumn, gdzie n to liczba kategorii dla danej kolumny. Proces ten nazywany jest *One-hot encoding*.