Deferred Learning 2

Michał Pastuszka

Idea

Znaleźć najbardziej “podobny” zbiór w bazie i wybrać najlepszy dla niego model

Definicja podobieństwa

Dla każdej kolumny naszego zbioru znaleźć w porównywanym zbiorze kolumnę o największym wskaźniku podobieństwa.

Wskaźnik podobieństwa dla zmiennych numerycznych

$$uniqueness = e^{-|unique(x)/length(x) - unique(y)/length(y)|}$$

$$missingness = e^{-|missing(x)/length(x) - missing(y)/length(y)|}$$

$$p = 1 - |q1(x) - q1(y)| + |\overline{x} - \overline{y}| + |med(x) - med(y)| + |q3(x) - q3(y)|/4$$

$$similarity = uniqueness * missingness * p$$

Wskaźnik podobieństwa dla zmiennych kategorycznych

$$categoryQuantity = e^{-|numberOfCategories(x) - numberOfCategories(y)|}$$

$$categoryRatio = e^{-|\frac{maxCategorySize(x) - minCategorySize(x)}{length(x)} - \frac{maxCategorySize(y) - minCategorySize(y)}{length(y)}|}$$

$$similarity = categoryQuantity * categoryRatio$$

Wynik

```
## openml_ozone-level-8hr
##                     79
```

Najlepszy model w bazie: `randomForestSRC`