



METODY INŻYNIERII WIEDZY

WALIDACJA KRZYŻOWA

Adrian Horzyk

Akademia Górniczo-Hutnicza

***Wydział Elektrotechniki, Automatyki, Informatyki
i Inżynierii Biomedycznej***

Katedra Automatyki i Inżynierii Biomedycznej

Laboratorium Biocybernetyki

30-059 Kraków, al. Mickiewicza 30, paw. C3/205

horzyk@agh.edu.pl, Google: Adrian Horzyk



WALIDACJA KRZYŻOWA

k-fold CROSS-VALIDATION

- ✓ Umożliwia wykorzystanie całego zbioru danych zarówno do uczenia, jak również do walidacji modelu.
- ✓ Służy do określenia jakości modelu już w trakcie jego adaptacji / uczenia, w celu wyeliminowania problemu **przeuczenia się (*overfitting*)**.
- ✓ Polega na podziale zbioru uczącego na k równolicznych podzbiorów, z których $k-1$ jest wykorzystanych do uczenia / adaptacji modelu, a 1 podzbiór służy do walidacji modelu.



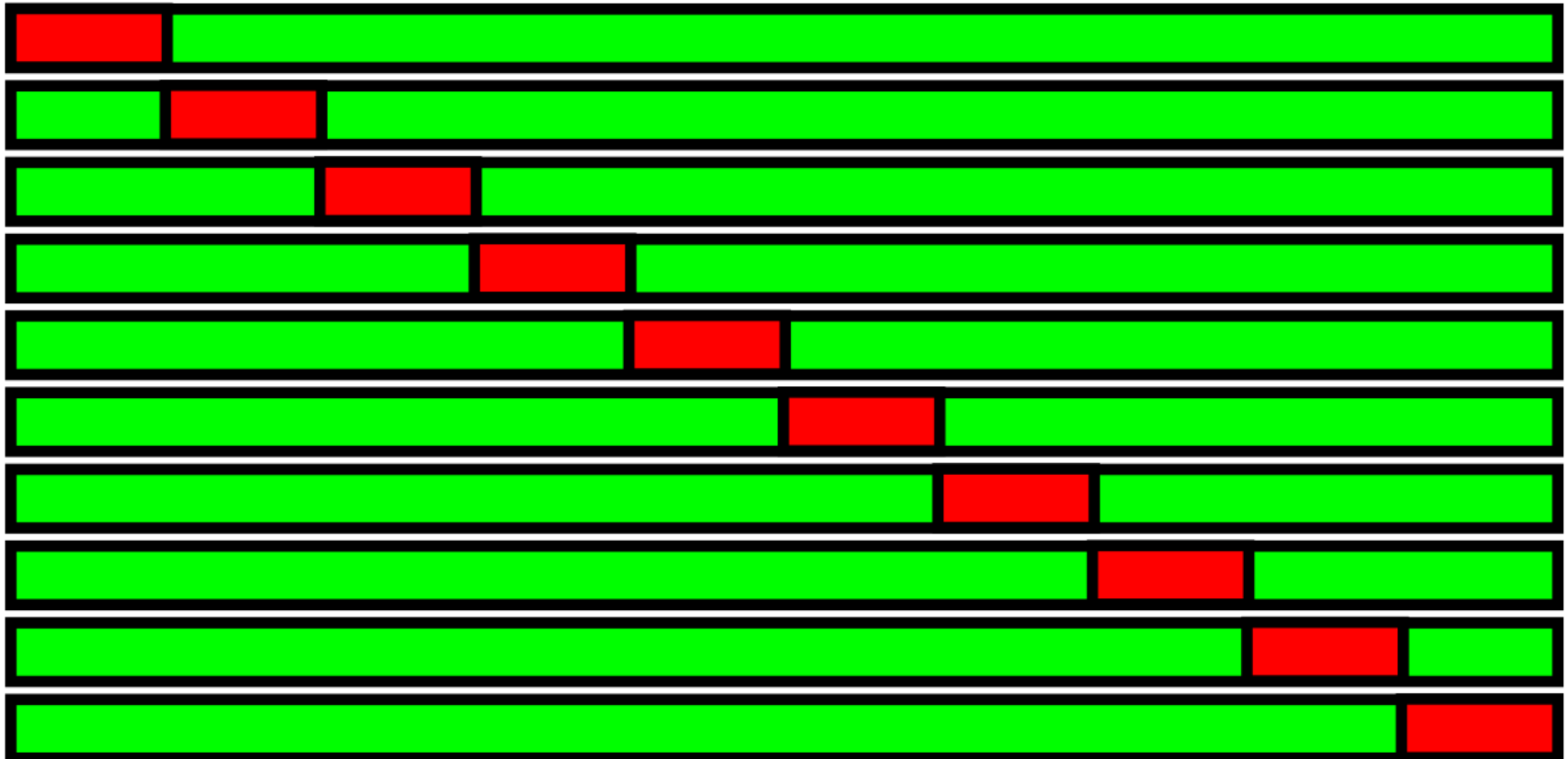
PRZYKŁAD

10-krotnej walidacji krzyżowej

CAŁY ZBIÓR DANYCH DZIELONY JEST



NA CZĘŚĆ UCZĄCĄ I WALIDACYJNĄ





PARAMETRY METODY WALIDACJI KRZYŻOWEJ

Dobór parametru k zależy jest od wielkości zbioru danych i ich rodzaju. Dla dużych zbiorów danych stosuje się $k=3$ w celu zmniejszenia ilości adaptacji modelu. Dla mniejszych zbiorów danych zwykle stosuje się większe wartości k , żeby nie uszczuplać zbioru uczącego za bardzo, co mogłoby spowodować budowę słabej jakości modeli. Najczęściej stosuje się $k=10$.

Sposób podziału zbioru danych na k podzbiorów jest niemniej istotny, gdyż jeśli wzorce są posortowane wg klas w zbiorze uczącym, wtedy wybór kolejnych podzbiorów może powodować uwzględnienie w walidacji tylko wzorców jednej klasy, a ponadto znaczne uszczuplenie wzorców uczących dla tej klasy, co jest bardzo niekorzystne z punktu widzenia budowy modelu!

Najlepiej dobierać wzorce **proporcjonalnie** do ich liczności i reprezentacji poszczególnych klas tak, aby były reprezentatywne.



4-krotnej walidacja krzyżowa z proporcjonalnym wyborem wzorców

CAŁY ZBIÓR DANYCH DZIELONY JEST PROPORCJONALNIE



DO LICZNOŚCI KLAS NA CZĘŚĆ UCZĄCĄ I **WALIDACYJNĄ**





N-krotna walidacja krzyżowa ***leave one out cross-validation***

N-krotna walidacja krzyżowa zakłada dobór $k=N$, gdzie N to ilość wzorców zbioru danych uczących.

W takim przypadku tylko 1 wzorzec jest wykluczany ze zbioru uczącego i traktowany jako walidacyjny.

Nauka jest więc powtarzana N -krotnie na zbiorach utworzonych poprzez pominięcie 1 wzorca stosowanego do walidacji.

Usunięcie jednego wzorca zwykle nie daje reprezentatywnych wyników działania odnośnie jakości modelu i jego możliwości generalizacji.



MODYFIKACJE METODY WALIDACJI KRZYŻOWEJ

W celu uzyskania lepszego działania metody, zbiór danych powinien być dzielony w taki sposób, żeby stosunek reprezentantów poszczególnych klas w zbiorze walidacyjnym był mniej więcej taki sam jak w całym zbiorze uczącym.

Z tego powodu zaleca się posortowanie zbioru danych uczących w taki sposób, żeby można było z każdej klasy wybierać kolejne podzbiory do walidacji krzyżowej w sposób **reprezentatywny** dla całego zbioru danych uczących.

Czasami stosuje się również uproszczoną metodę selekcji $1 / k$ wzorców ze zbioru uczącego do walidacji, polegającą na **losowaniu tych wzorców z całego zbioru (*random subsampling*)**, a pozostałe wykorzystywane są do adaptacji / uczenia modelu.

METODA WALIDACJI KRZYŻOWEJ W RAPIDMINERZE

RapidMiner udostępnia operator X-Validation, który umożliwia wybrać k , oraz wewnątrz umieścić walidowaną metodę automatycznie testowaną k -razy.

