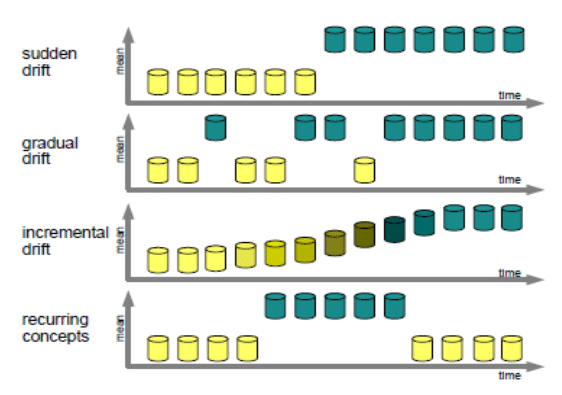
**Niestacjonarność w strumieniach danych. Algorytmy detekcji zmian w strumieniach danych.**

1. Strumień danych
   1. sekwencja danych, często powiązanych relacjami, generowanych i napływających do systemu w czasie, w stałych lub zmiennych odstępach czasu
   2. wielkość takiego strumienia w teorii jest nieskończona, jednak w praktyce zakłada się, że zbiór równy lub przekraczający sto tysięcy obiektów można uznać za strumień danych.
2. Ograniczenia sprzętowe
   1. strumienie danych mogą być bardzo duże, przez co niemożliwe do zapisania i przechowywania. Potrzebne są więc algorytmy działające w czasie zbliżonym do rzeczywistego.
3. Niestacjonarność strumieni danych
   1. stacjonarne strumienie danych nie posiadają zmienności w czasie. Inaczej mówiąc jest to taki przypadek, gdzie na przestrzeni całego strumienia nie występują żadne zmiany w rozkładzie danych.
   2. niestacjonarne są ściśle związane z pojęciem dryfu koncepcji
4. Dryf koncepcji (concept drift)
   1. po ludzku rzecz biorąc polega na tym, że dane mają tendencję do zmieniania swoich wartości w czasie. Te zmiany są często nieprzewidywalne i mogą wpłynąć na pogorszenie jakości klasyfikatora.
   2. do mierzenia zmian wartości możemy używać czego się marzy: średniej, mediany, wariancji
   3. rodzaje concept driftu:

* nagły
* stopniowy
* przyrostowy
* powtarzający się



1. Algorytmy detekcji zmian w strumieniach danych
   1. ruchome okna (sliding windows)

Najprostszym sposobem na wykrycie danych jest sprawdzanie czy dane przychodzące znacząco różnią się od poprzednich, czyli na przykład jeśli odchylenie standardowe jakiejś cechy (parametru) przekroczy ustaloną wcześniej wartość.

Chodzimy ruchomymi oknami “wzdłuż” strumienia danych i zbieramy kolejne - jeśli trafimy na wartość odstającą od reszty, informujemy o zmianie i zaczynamy tworzenie klasyfikatora od nowa.

* 1. metody komitetowe

Tworzymy jednocześnie wiele klasyfikatorów i uczymy je na blokach danych (batchach). Usuwamy nieodpowiadające nam klasyfikatory na podstawie na przykład:

* + 1. czasu istnienia - stare klasyfikatory mogą mieć naleciałości starych danych, które nie odpowiadają obecnym
    2. kryterium jakości - dodając odpowiednie kryterium jakości możemy pozbywać się nieefektywnych klasyfikatorów i polepszać w ten sposób ich “populację”
  1. SEA (przykład metody komitetowej)
     1. tworzymy zadaną liczbę klasyfikatorów i zakładamy, że powinna ona zostać stała
     2. uczone są one na małych zbiorach danych
     3. uczymy następne klasyfikatory i jeśli są lepsze (według wcześniej określonego kryterium) od najgorszego z naszego zbioru “roboczego”, to są one zamieniane tak, aby podnosić średnią wydajność zbioru
     4. algorytm dostosowuje się do dryftu danych i niejako “podświadomie” go wykrywając, podnosi jakość klasyfikatora

1. Po co to w ogóle jest
   1. bo strumienie danych się zmieniają i żeby nasz klasyfikator nie stał się przestarzały musimy wykryć zmiany, aby móc odpowiednio zareagować
   2. do rozpoznawania spamu na podstawie zawartości i/lub metadanych
   3. kontrola jakości (na przykład na podstawie zestawu czujników sprawdzamy, czy produkt nie ma wad i może iść do następnego procesu produkcji)