6 Instrukcja dla ćwiczenia nr 6: Aproksymacja liniowa i nieliniowa.

6.1 Wstęp

1. **Terminologia – aproksymacja liniowa vs. nieliniowa**.
   Przypominam **warunki liniowości przekształcenia** $x \to f(x)$:

   - addytywność $f(x + y) = f(x) + f(y)$,
   - jednorodność $f(cx) = c \cdot f(x)$.

   Używa się pojęcia *funkcja liniowa* dla określenia przekształcenia $f(x) = ax + b$. Taka funkcja **nie jest przekształceniem liniowym** - lecz afinicznym.

   W literaturze poświęconej zagadnieniom aproksymacji można napotkać różne rozumienie pojęcia aproksymacji liniowej:

   - wąskie - aproksymacja funkcją liniową (a dokładniej - afiniczną) zmiennej x,
   - szersze - aproksymacja liniową kombinacją dowolnych funkcji zmiennej x.

   Uwaga: Funkcję jednej zmiennej można tu zastąpić funkcją n zmiennych $[x_1, x_2, \ldots, x_n]$, ale w tym ćwiczeniu ograniczymy się do funkcji jednej zmiennej.

   W dalszej części będziemy używać pojęcia aproksymacji w szerszym znaczeniu, czyli w przypadku aproksymacji liniowej, funkcja aproksymującą ma postać

   $$f(x) = a_0 f_0(x) + a_1 f_1(x) + \ldots + a_m f_m(x)$$

   Funkcje $f_i(x)$, $i = 0, 1, \ldots, m$ nie mają parametrów (albo mają, ale są ustalone i nie podlegają zmianom w czasie wyznaczania funkcji aproksymującej). Funkcje te muszą być liniowo niezależne, a więc tworzą $(m + 1)$-wymiarową "bazę", np.:

   - $f_i(x) = x^i$,
   - $f_i(x) = x^{-i}$,
   - $f_i(x) = e^{ix}$.

   W większości zadań tego ćwiczenia funkcją aproksymującą będzie wielomian.

2. **Ćwiczenie nr 6 obejmuje zadania aproksymacji dyskretnej (a nie ciągłej)** .
   W przypadku aproksymacji dyskretnej, wartosci zmiennej zależnej y znane są jedynie dla skończonej liczby wartości zmiennej niezależnej x.[1] Czyli dane są współrzędne $n+1$ punktów $P_j(x_j, y_j)$ $j = 0, 1, \ldots, n$.

   Zadnie aproksymacji polega na znalezieniu funkcji najbliższej zadanym punktom.

[1] W teorii funkcji mówi się o aproksymacji, w statystyce - o regressji. W statystyce używa się bardziej-sudestywnych - moim zdaniem - określeń: y to zmienna objaśniana, a x to zmienna objaśniająca

3. **Trzy decyzje poprzedzające aproksymację**:

   a. Wybór klasy funkcji aproksymującej, tj bazy funkcji $f_i(x)$, $i = 0, 1, \ldots, m$.

   b. Wybór rozmiaru przestrzeni generowanej przez tę bazę, tj. wybór wartości m.

   c. Wybór miary odległości funkcji od zadanych punktów.

4. **Znajdowanie optymalnych wartości parametrów funkcji aproksymującej**:

   - Rozwiązując układ równań normalnych (równań powstałych z przyrównania pochodnych miary odległości względem parametrów do zera) - jeżeli układ równań normalnych jest liniowy. Odwołuję się tutaj do wykładu Aproksymacja2020.pdf umieszczonego w materiałach do ANiSS w Wirtualnym dziekanacie - ramki 19 - 27.
   - W przeciwnym przypadku - stosując ogólne metody optymalizacji. Ogólnie o tych zagadnieniach w materiałach do wykładu Optymalizacja2020.pdf w j.w.

5. **Problem nadmiernego dopasowania**
   O walidacji krzyżowej – ramka 36 w przentacji Aproksymacja2020.pdf

6.2 Plan ćwiczenia nr 6

1. **Aproksymacja liniowa danych dokładnych**

   Dane aproksymowane będą niezaburzonymi wartościami pewnej nieznanej funkcji.

   Celem jest liniowa aproksymacja tych danych.

   Ten punkt ćwiczenia ma dwa etapy:

   A. wybór najlepszego algorytmu numerycznego,

   B. wybór najlepszej funkcji aproksymującej (w określonej z góry klasie funkcji).

   Dla spełninia założenia o aproksymacji liniowej:

   - Wybieramy funkcję aproksymującą liniowo zależną od jej parametrów, np. $f_i(x) = x^i$, czyli wielomian stopnia m, $f(x) = \sum_{i=0}^{m} a_i x^i$.
   - Wybieramy średniokwadratową miarę odległości punktów od funkcji.

   Ad A. **Porównanie algorytmów numerycznych wyznaczania parametrów w aproksymacji liniowej**

   Nieznaną zależnością w danych użytych w tej części ćwiczenia jest funkcja wykładnicza $y = \exp(x)$.

Realizacja:

– Skrypt `algorytmy.m` wykonuje trzy numeryczne algorytmy MATLABa wyznaczania optymalnych wartości parametrów funkcji aproksymującej:
  (a) z macierzą pseudoinwersji Moore'a-Penrose'a i obliczaniem macierzy odwrotnej,
  (b) z wykorzystaniem matlabowego „lewego dzielenia" przez macierz współczynników $X^T X$ (czyli zastąpienie inwersji macierzy algorytmem rozwiązywania układu równań),
  (c) z wykorzystaniem bibliotecznej funkcji `polyfit` dla aproksymacji wielomianowej (opis w pomocniku MATLABa np. `help polyfit`).

  Przykładowe dane - punkty leżące na wykresie funkcji wykładniczej - są aproksymowane wielomianami stopnia od 1 do 23. Skrypt wypisuje błędy aproksymacji `d1, d2, d3` mierzonych normą RMSE wg 3 ww. algorytmów dla wielomianów kolejnych stopni od 1 do 23. Odsłonięcie komentarza umożliwi rysowanie wykresów danych i trzech wielomianów - tego samego stopnia, ale o współczynnikach obliczanych wg innych algorytmów.
– Należy (jakościowo) porównać błędy numeryczne badanych algorytmów – od najgorszego do najlepszego.

Ad B. **Wybór stopnia wielomianu aproksymującego dane niezaburzone**

Skoro dane są dokładne, to można byłoby sądzić, że im funkcja aproksymująca jest bliższa danym, tym aproksymacja jest lepsza, tj. aproksymacja lepiej przybliża nienaną zależność $y(x)$. Czy tak jest rzeczywiście i ewentualnie do jakich granic.

Realizacja:

– Dane generujemy skryptem `generator_danych_1.m`
  Nieznaną zależnością jest jeden okres funkcji sinus.
  Liczba punktów `l_pom` jest wstępnie ustalona na 50.
– **Obliczanie optymalnych wartości współczynników wielomianów różnych stopni.** Do obliczania współczynników funkcji aproksymującej wybieramy najdokładniejszy dostępny algorytm numeryczny. Stopień wielomianu jest zmieniany w pętli od 1 do n_{max} (wstępnie
  ustawione na 15). W każdej iteracji pętli są rysowane 2 wykresy:
  * górny - funkcji aproksymującej na tle punktów aproksymowanych,
  * dolny - wykres róznicy między funkcją aproksymującą, a zależnością aproksymowaną (sinusem).

  Równolegle z rysowaniem wykresów, w oknie Command window, pojawia sie informacja o n - aktualnym stopniu wielomianu aproksymujacego oraz błędzie średniokwadratowym - blad.
– Należy zaobserwować jak zmienia się błąd aproksymacji gdy stopień wielomianu aproksymującego rośnie.

2. **Aproksymacja liniowa danych z losowym błędem**
W poprzednim punkcie dane były dokładne. Teraz dane zawierają składnik zaburzający faktyczną zależność $y(x)$. Zatem zbliżanie funkcji aproksymującej do danych nie zawsze oznacza zbliżanie do funkcji $y(x)$ – a to jest celem aproksymacji. Możemy przewidywać, że po przekroczeniu pewnej granicy, funkcja aproksymująca będzie dążyła do odwzorowania nie tylko $y(x)$, lecz także także błędu pomiarowego. Jeżeli ten błąd jest losowy, to odwzorowywanie go na podstawie jednej serii pomiarów nie jest racjonalne.
Należy teraz poszukać nowej granicy w polepszaniu odwzorowywania $y(x)$, gdy funkcja aproksymująca zaczyna "dopasowywać się" do zaburzeń.

Ilustracja funkcji aproksymującej na przykładzie abstrakcyjnym.

- Skrypt `generator_danych_2.m` generuje wartości z jednego okresu funkcji sinus, ale dodaje do nich składnik losowy reprezentujący błąd pomiaru. Skrypt rysuje nieznanej („ukrytej w pomiarach") zależności faktycznej - sinusa oraz wykres zaburzonych danych - każde uruchomienie skryptu daje inny rozkład błędu (otrzymywanego z generatora liczb pseudolosowych).
- Decyzje o wyborze klasy funkcji aproksymującej (wielomianu) i miary odległości punktów od funkcji - jak w poprzednim punkcie ćwiczenia.
- Skrypt wykonuje iteracje pętli dla kolejnych stopni wielomianu aproksymującego, w których rysuje wykres wielomianu i wyprowadza wartości miary odległości danych od funkcji aproksymującej. Czy ta odległość zawsze maleje ze wzrostem stopnia wielomianu?
- Czy kryterium odległości jest właściwe dla wyboru stopnia wielomianu?

Przykład z danymi realnymi.

- Temperatura powietrza mierzona co godzinę w dniu 22 marca 2016 roku w Badenii-Wirtembergii – są to oficjalne dane meteorologiczne prezentowane zgodnie z wytycznymi UE, a gromadzone dla celów optymalizacji pracy elektrowni solarnej.
- Czy są to dane dokładne. Należałoby zadać pytanie – co to znaczy dokładne? Tepmperatura w całym kraju związkowym (choć niewielkim) nie jest dokładnie taka sama, a więc możnna dopuścić niewielkie odległości fubnkcji aproksymującej od danych. Przyjmijmy, że ważniejsze od odległości od danych jest „odtworzenie kształtu" zależności temperatury od czasu.
- Skrypt `dane_meteo.m` zawiera dane, oraz rysuje na ich tle kolejne wykresy wielomianów aproksymujące kolejnych stopni.
- Na podstawie obseracji wykresu należy wskazać: Wielomian którego stopnia najlepiej przybliża zmiany temperatury w czasie dnia? Należy uwzględnić (wyważyć) odległość wielomianu od danych oraz kształt wykresu (np. czy krótkookresowe wahania temperatury są odzwierciedleniem faktycznego, powtarzalnego i regularnego zjawiska).

3. **Aproksymacja liniowa – Rozpoznawanie nadmiernego dopasowania**

   - Wracamy do przykładu „sztucznego” - nieznanej zależności $y(x)$ i danych pomiarowych zaburzonych losowym błędem. W poprzednim punkcie ćwiczenia podstawą wskazania najlepszego wielomianu była obserwacja wykresów wielomianu aproksymującego. Celem tego punktu jest znalezienie bardziej obiektywnego i wymiernego ilościowo kryterium wyboru najlepszej aproksymacji.
   - Zadanie: znaleźć ilościową miarę najlepszego, ale nie nadmiernego dopasowania funkcji aproksymującej do danych zaburzonych.
   - Skrypt `over_fitting.m` generuje dane tak jak skrypt poprzedni `gegerator_-danych_2`. Dane aproksymowane dzieli na dwa podzbiory: dane treningowe (zaznaczone na wykresie kółkami) i testowe (zaznaczone krzyżykami).
   - Kolejne wielomiany aproksymujące wyznacza na postawie tylko podzbioru treningowego - oblicza parametry (współczynniki wielomianu minimalizując odległość od punktów oznaczonych kółkami, a ignorując krzyżyki).
   - Odległość każdego wielomianu od zbioru testowego (krzyżyków) też jest obliczana (ale nie brana pod uwagę przy wyznaczaniu wielomianu).
   - Obie odległości (`RMSE_cw` i `RMSE_test`)są wyprowadzane na ekran dla każdego wielomianu (w każdej iteracji pętli). Po zakończeniu pętli jest rysowany wykres przebiegu tych odległości dla zmieniającego się (rosnącego) stopnia kolejnych wielomianów,
   - Skrypt należy wykonać kilka razy (za każdym razem wynik będzie inny z uwagi na losowość błędu pomiaru danych). Na podstawie obserwacji wykresów przebiegu odległości wielomianu od danych treningowych i testowych należy określić właściwy stopień wielomianu aproksymującego.
   - Czy te obserwacje nasuwają pomysł algorytmu, który bez udziału „czynnika ludzkiego” wskazywałby stopień najlepszego wielomianu aproksymującego?

4. **Wstęp do aproksymacji nieliniowej – Problem wyboru klasy funkcji aproksymującej**.

   - Dane aproksymowane pochodzące z pomiaru przepływu wody $[m^3/s]$ w Rabie w Stróży w ciągu 68 godzin (od godziny 10 17 listopada 2015r do godziny 6 20 listopada 2015r) są w pliku `dane_przeplywu.m`.
     Dane te obejmują pojedynczą falę wezbraniową po dużym opadzie deszczu w zlewni Raby.
   - Należy zaproponować klasę funkcji aproksymującej przebieg opadania przepływu - pomiary od 18. do 69. Czy wybór wielomianu byłby najlepszą decyzją? Jakie są inne możliwości?

6.3 Skrypty

- `algorytmy.m` - oblicza współczynniki wielomianu aproksymacyjnego trzema algorytmami.
- `generator_danych_1.m` - aproksymacja danych niezaburzonych (dokładnych).
- `generator_danych_2.m` - aproksymacja danych zaburzonych – dane symulowane, fizycznie zmierzone – model zależności jest znany.
- `dane_meteo.m` - aproksymacja danych zaburzonych – dane fizycznie zmierzone – model zależności nie jest znany.
- `over_fitting.m` - ilustracja u nadmiernego dopasowania.
- `dane_przeplywu.m` - przykład danych pomiarowych dla aproksymacji aproksymacji innymi funkcjami niż wielomiany – wstęp do aproksymacji nieliniowej.

6.4 Punkty sprawozdania

1. Aproksymacja liniowa - dane dokładne (niezaburzone błędem losowym:
    - Który algorytm wyznaczania parametrów funkcji aproksymującej jest najdokładniejszy?
    - Który algorytm należy wybrać dla aproksymacji liniowej ale funkcją inną niż wielomian?
    - Wielomian którego stopnia daje najmniejszy błąd aproksymacji?
    - Czy im wyższy stopień wielomianu tym mniejszy błąd? Czy błąd aproksymacji zmaleje do zera?
2. Aproksymacja liniowa - dane z błędem losowym
    - Czy odległość punktów (danych) od wielomianu aproksymującego zawsze maleje ze wzrostem stopnia wielomianu? Krótkie uzasadnienie.
    - Aproksymacja liniowa - dane realne.
      Wielomian którego stopnia najlepiej przybliża zmiany temperatury w czasie dnia? Czy decyduje wielkość błedu aproksymacji? (z krótkim uzasadnieniem).
    - Czym się kierować przy doborze stopnia wielomianu aproksymującego gdy nie mamy wiedzy o aproksymowanej zależności?
3. Jak wykryć nadmierne dopasowanie
    - Jaki stopień wielomianu należy wybrać widząc rezultat kilku eksperymentów (uruchomień skryptu)? Odpowiedź nie musi być liczbą – może być przedziałem.

- Czy zastosowany w ćwiczeniu podział danych na zbiór ćwiczebny i testowy jest zgodny z regułami?
- Ewentualnie: Dodać więcej danych i zaproponować lepszy podział.
- Ilościowe rozpoznawanie nadmiernego .
  Jak sformułować prosty algorytm właściwego doboru stopnia wielomianu - a ogólniej - liczby parametrów funkcji aproksymującej? Uwaga: użyłem słowa „prosty” dla zaznaczenia, że algorytm może porównywać wartości liczbowe, a nie np. kształty wykresów.
- Na czym polega eksperyment walidacji krzyżowej? - jednym zdaniem.

4. Problem wyboru klasy funkcji aproksymującej

   - Czym się kierować w przypadku danych z pliku `dane_przeplywu.m`?
   - Które funkcje można brać pod uwagę:
     a). $\exp(at)$,
     b). $a + \exp(bt)$,
     c). $a \cdot \exp(bt - c)$,
     d). $\frac{\sum_{i=0}^{n} a_i t^i}{\sum_{i=0}^{m} b_i t^i}$,
     e). $\sum_{k=0}^{n} a_k \exp(kt)$,
     f). $\sum_{k=0}^{n} a_k t^{-k}$,
     g). $\sum_{k=0}^{n} a_k \exp(b_k t)$,
     h). $a_0 + \sum_{k=1}^{n} (a_k \sin(kt) + b_k \cos(kt))$,
     i). inne propozycje...
   - Które z powyższych można zastosować w aproksymacji liniowej?
     Ew. jakie zmiany w algorytmie w stosunku do algorytmu dla wielomianów?
   - Czy możliwe jest użycie funkcji wymiernej w zadaniu aproksymacji liniowej?
     Ew. z jakimi ograniczeniami (tzn. czy każdą funkcję wymierną)?
   - Aproksymacja nieliniowa - zalety i wady w stosunku do liniowej (krótkie wyliczenie)