Uczenie Maszynowe
Laboratorium 6: Gaussian Processes

1 Cele laboratorium

- Praktyczne zapoznanie się z Procesami Gaussowskimi w kontekście przewidywania szeregów czasowych
- Automatyczne doposowywanie parametrów funkcji jądra do danych treninowych poprzez maksymalizację ujemnej zlogarytmowanej funkcji wiarygodności (gradientowe metody optymalizacji)

2 Literatura

- Peter Roelants, *Understanding Gaussian processes*, Github
- Ch. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- D.P. Kingma, J. Ba, *Adam: A method for stochastic optimization*, 2014.

3 Przykładowe dane

`http://home.agh.edu.pl/~czech/ml/`

(a) Pomiary koncentracji ppm cząsteczek CO_2 w atmosferze w obserwatorium Manua Loa od 1958 do 2021 roku

(b) Notowania indeksu *WIG20* (Data, Otwarcie, Najwyższy, Najniższy, Zamknięcie, Wolumen) od 16.04.1991 do 10.12.2020

(c) Notowania spółki *Comarch* (Data, Otwarcie, Najwyższy, Najniższy, Zamknięcie, Wolumen) od 10.03.1999 do 10.12.2020

(d) Notowania spółki *CD Projekt* od 02.08.1994 do 10.12.2020

4 Przydatne biblioteki i narzędzia

1. Google Colab files
2. Tensorflow, Tensorflow Probability
3. Numpy
4. Pandas
5. Bokeh (`https://bokeh.org`)

5 Przewidywanie koncentracji CO_2 w długim horyzoncie czasowym

1. Załaduj zbiór danych (a) (`pandas.read_csv()`), wybierz z niego kolumnę 3 oraz 4 (data, koncentracja CO_2 [ppm])
2. Dokonaj wstępnego przetworzenia danych usuwając rekordy z brakującym pomiarem koncentracji oraz rekordy z wartościami NaN
3. Dokonaj wizualizacji całości danych (wykres [ppm] dla kolejnych lat, aż do roku 2021)
4. Podziel dane na część treningową i testową (dane testowe od początku roku 2012)
5. Zdefiniuj funkcję wartości średniej, jako funkcję stałą określoną przez średnią wartość pomiaru w zbiorze treningowym (`tensorflow.constant`)
6. Zdefiniuj sparametryzowaną złożoną funkcję kowariancji (jądra) będącą sumą następujących funkcji jądra:
   - Exponential Quadratic kernel (`tfp.ExponentiatedQuadratic`)
   - Local Periodic kernel (iloczyn `tfp.ExpSinSquared` oraz `tfp.ExponentiatedQuadratic`)
   - RationalQuadratic kernel (`tfp.RationalQuadratic`)
   - White Noise kernel

$$\begin{aligned} k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \theta_0 \exp\left(-\frac{\theta_1}{2}\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2\right) + \\ \theta_2 \exp\left(-\frac{2}{\theta_3}\sin^2\left(\pi\frac{\|\boldsymbol{x}_a - \boldsymbol{x}_b\|}{\theta_4}\right)\right)\exp\left(-\frac{\theta_5}{2}\left\|\boldsymbol{x}_a - \boldsymbol{x}_b\right\|^2\right) + \\ \theta_6\left(1 + \frac{\left\|\boldsymbol{x}_a - \boldsymbol{x}_b\right\|^2}{2\theta_7\theta_8}\right)^{-\theta_7} + \theta_9 \end{aligned} \tag{1}$$

7. Zdefiniuj ujemną zlogarytmowaną funkcję wiarygodności, której argumentami są obserwacje ze zbioru treningowego, złożona funkcja jądra oraz funkcja wartości średniej (`tfp.GaussianProcess`)

8. Zainicjalizuj i uruchom optymalizator Adam (`tf.keras.optimizers.Adam`): *learning_rate=0.001, batch_size=128, nb_interations=10001* w celu minimalizacji ujemnej zlogarytmowanej funkcji wiarygodności

9. Po zakończeniu obliczeń narysuj wykres pokazujący jak zmieniała się (malała) wartość optymalizowanej funkcji w kolejnych iteracjach procesu (wartość obliczona dla wszsytkich danych treningowych oraz wartości dla podzbiorów 128 punktów)

10. Wyświetl w tabeli wartości znalezionych parametrów θ_i

11. Korzystając wyznaczonych parametrów funkcji jądra oraz wcześniej wyznaczonej funkcji wartości średniej, utwórz model regresji typu posterior pozwalający na wykonywanie predykcji dla zbioru testowego (`tfp.GaussianProcessRegressionModel`)

12. Dla zbioru testowego narysuj predykcję koncentracji CO_2 (począwszy od 2012 roku). Jako niepewność predykcji zaznacz dwa odchylenia standardowe.

13. Skomentuj uzyskane wyniki. W jakim horyzoncie czasowym wyniki przewidywania koncentracji CO_2 mieszczą się w przedziale niepewności $\pm 2\sigma$? Jaki trend można zaobserwować dla predykcji i rzeczywistych wartości od około 2016 roku?

6 Przewidywanie notowań giełdowych whoryzoncie 7 i 14 dni

1. Korzystając ze zbioru danych (b) oraz jednego z dwóch pozostałych (c) albo (d) wykonaj eksperyment podobny jak w poprzednim zadaniu, przewidując kurs w horyzoncie 7 i 14 dni. Dla każdego z dwóch zbiorów wybierz 2 różne punkty podziału na zbiór treningowy i zbiór testowy. Jeden punkt podziału w 2019, a drugi w 2020 roku. Przedstaw wizualizację wyników. Skomentuj dokładność predykcji.

2. * W jaki sposób możemy zapobiegać przeuczeniu, w sytuacji agresywnej optymalizacji parametrów funkcji kowariancji na zbiorze treningowym? Zaproponuj i zaimplementuj rozwiązanie.

3. * W jaki sposób zmienia się błąd NRMSE w przewidywaniu kursu na 7 dni w przód przy przesuwaniu punktu podziału na zbiór treningowy / testowy 10 razy o 7 dni?

4. * Porównaj uzyskane rezultaty predykcji z wynikami uzyskanymi dla modelu Prophet (`https://facebook.github.io/prophet/`).