**University of Wrocław: Data Science** 09.07.2020

Math and coding problems

1. A palindromic number reads the same both ways. Find the largest 13 digit palindromic **prime** number.

2. A number is 'sevenfold' if it contains the sequence of 7 consecutive 7's (for instance 1237777777 is sevenfold, and 77077077077 is not). What is the number of 10-digits sevenfold prime numbers?

3. Assume we sample (uniformly at random, with replacement) $n = 130$ times from a set {A, C, G, T}. What is the expected number of occurences of the pattern ACA?

   Note that overlapping is allowed, *e.g.,* ACACA contains two overlapping occurences of the pattern.

4. For some dataset $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$ it seems reasonable to assume that the regression line passes through the origin of the coordinate system. Thus the model becomes

   $$Y_i = \beta x_i + U_i, \quad \text{for } i = 1, 2, .., n,$$

   where $U_1, U_2, ..., U_n$ are independent random variables with zero expectation and finite variance.

   **(i)** Find the least squares estimator $\hat{\beta}$ of $\beta$.

   **(ii)** For the dataset:

   $$(1.50, 8.034), (3.45, 18.099), (5.46, 30.378), (7.20, 39.737), (8.45, 42.570),$$

   $$(14.00, 73.951), (10.05, 52.818), (7.34, 38.973), (2.68, 17.373), (9.49, 50.426)$$

   give the value of least squares estimator $\hat{\beta}$.

5. Take a sequence of 1000 pseudo-random numbers $x_1, \ldots, x_{1000}$ generated using the recurrence

   $$\begin{cases} x_0 &= 0 \\ x_i &= (x_{i-1} \cdot 1664525 + 1013904223) \bmod 2^{32}, i = 1, \ldots, 1000 \end{cases}$$

   Let $y_i = \dfrac{x_i}{2^{32}}$ such that $y_i$ are within $[0, 1)$.

   We will quantize these numbers by rounding them to a number from the set $Z = \{0, 0.1, 0.2, \ldots, 1.0\}$. Let $z_i$ be the quantized value. The quantization error is defined to be the difference $y_i - z_i$.

(a) suppose that each point i is quantized independently from the other ones, by setting $z_i$ to the element of set $Z$ that is closest to $y_i$. Find the root mean square error (RMSE), i.e. $\sqrt{\sum_{i=1}^{1000}(z_i - y_i)^2}$.

(b) now suppose we want to limit the number of changes of value in the sequence $z_i$: the quantization of point i is no longer performed independently from other points

- Assume that the value of $z_i$'s can only change once, i.e. there is a $k$ such that $z_1 = z_2 = \ldots = z_k$ and $z_{k+1} = z_{k+2} = \ldots = z_{1000}$ and $z_k \neq z_{k+1}$. Find the error of the optimal quantization in this case.
- Assume that the quantized sequence can change exactly 10 times, find the error of the optimal quantization. What is the runtime complexity of your algorithm?
- Assume the quantized sequence can change exactly 333 times, now the optimal quantization is quite expensive to find (how expensive is the computation of the exact solution?), find a good quantization with possible low RMSE that abides by the constraint and explain how your heuristic algorithm works and what its computational complexity is.