Осень 2016 - БОЛЬШИЕ ДАННЫЕ - Домашнее задание №2

1. Рассмотрим следующую серию измерений неизвестной величины x:

$$y_i = a_i x + \varepsilon_i, \qquad i = 1, \dots, n,$$

где y_i - результаты измерений, a_i - известные коэффициенты и случайные величины ε_i, представляющие ошибки измерения, независимы и одинаково распределены с нулевым средним и дисперсией σ^2:

$$\mathsf{E}\varepsilon_i = 0, \qquad \mathsf{D}\varepsilon_i = \mathsf{E}\varepsilon_i^2 = \sigma^2, \qquad i = 1, \dots, n,$$

(a) Какую функцию от $y_1, \dots, y_n$ и $a_1, \dots, a_n$ Вы бы использовали в качестве хорошей оценки $\widehat{x}$ для x? $\widehat{x} =?$

(b) Является ли эта оценка оптимальной в каком-либо смысле?

(c) Является ли она несмещенной оценкой?

(d) Какова ее дисперсия (выраженная через σ^2)? $\mathsf{D}\widehat{x} =?$

(e) Как бы Вы оценили σ^2 если она неизвестна? $\widehat{\sigma^2} =?$

(f) Что можно использовать в качестве оценки для $\mathsf{D}\widehat{x}$ если σ^2 неизвестна? $\widehat{\mathsf{D}\widehat{x}} =?$

(g) Предположим, что дисперсия σ^2 известна. Какую "каноническую информацию" было бы достаточно извлечь из серии данных

$$(y_1, a_1), \dots, (y_n, a_n), \qquad i = 1, \dots, n$$

для того, чтобы вычислить оценку $\widehat{x}$ и ее дисперсию $\mathsf{D}\widehat{x}$?

(h) Предположим, что дисперсия σ^2 НЕ известна. Какую "каноническую информацию" было бы достаточно извлечь из серии данных для того, чтобы вычислить $\widehat{x}$, $\widehat{\sigma^2}$, and $\widehat{\mathsf{D}\widehat{x}}$?

(i) Как следует обновлять такую "информацию" когда поступает новое "наблюдение" (y_{n+1}, a_{n+1})?

(j) Как следует "объединять" информацию в канонической форме?

Пожалуйста, не пытайтесь использовать общие формулы, а получите все, насколько это возможно, с нуля.

2. Напишите программу, которая иллюстрирует линейную регрессию и реализует накопление канонической информации.

(a) Для некоторых фиксированных параметров $a_1, \dots, a_m$ сгенерируйте последовательность "наблюдений" (x_i, y_i):

$$y_i = f(x_i) + \varepsilon_i,$$

где

$$f(x) = a_1 + a_2 x + a_3 x^2 + \dots + a_m x^{m-1}$$

ε_i независимы и одинаково распределены с нулевым средним и дисперсией $\mathsf{E}\varepsilon_i^2 = \sigma^2$. Значения x_i также могут быть сгенерированы случайным образом.

(b) Осуществите накопление канонической информации, т.е. на каждом шаге, когда производится новое наблюдение (x_i, y_i), обновляйте каноническую информацию.

(c) Проиллюстрируйте истинную функцию $f(x)$ и ее оценку $\widehat{f(x)}$.

(d) Проиллюстрируйте $\mathsf{D}\widehat{f(x)}$, при условии, что σ^2 известно.

(e) Проиллюстрируйте $\widehat{\mathsf{D}\widehat{f(x)}}$, при условии, что σ^2 НЕ известно.

В своем отчете представьте исходный код и несколько (около 3) графиков, показывающих оценки для "малого", "среднего" и "большого" числа наблюдений.