# Big Data Analitycs - Spring 2014 - Homework #2

1. Consider the following series of measurements of the unknown value $x$:

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

where $y_i$ are measurement results, $a_i$ are known coefficients, and $\varepsilon_i$ represent random error of measurement and are independent identically distributed (i.i.d.) with zero mean and variance $\sigma^2$:

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

(a) What function of $y_1, \dots, y_n$ (and of $a_1, \dots, a_n$) would you use as a good estimate $\widehat{x}$ for $x$? $\widehat{x} = ?$

(b) Is this estimate optimal in any sense?

(c) Is it a biased or an unbiased estimate?

(d) What is its variance (expressed through $\sigma^2$)? $\mathrm{Var}(\widehat{x}) = ?$

(e) How would you estimate $\sigma^2$ if it is unknown? $\widehat{\sigma^2} = ?$

(f) What would you use as an estimate for $\mathrm{Var}(\widehat{x})$ if $\sigma^2$ is unknown? $\widehat{\mathrm{Var}(\widehat{x})} = ?$

(g) Suppose that the variance $\sigma^2$ is known. What "canonical information" would be sufficient to extract from the series of observations

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

in order to compute the estimate $\widehat{x}$, and its variance $\mathrm{Var}(\widehat{x})$?

(h) Suppose that the variance $\sigma^2$ is NOT known. What "canonical information" would be sufficient to extract from the series of observations in order to compute $\widehat{x}$, $\widehat{\sigma^2}$, and $\widehat{\mathrm{Var}(\widehat{x})}$?

(i) How should we update such "information" when a new observation $(y_{n+1}, a_{n+1})$ arrives?

(j) How should we "combine" (merge) two pieces of "canonical information"?

Please do not try to use general formulas, but develop as much as possible from scratch.

2. Write a program which illustrates simple linear regression (or a more general variant of linear regression) and implements accumulation of canonical information.

   (a) For some fixed parameters $a$ and $b$ (or, in a more general case, $a_1, \dots, a_m$) generate a sequence of "observations" $(x_i, y_i)$:

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

   where

   $$f(x) = a + bx \quad \text{or} \quad f(x) = a_1 + a_2 x + a_3 x^2 + \cdots + a_m x^{m-1}$$

   $\varepsilon_i$ are i.i.d. with zero mean and $\mathrm{E}\varepsilon_i^2 = \sigma^2$. Values $x_i$ can be generated randomly with some mean and variance.

   (b) Accumulate canonical information, i.e., at each step, when a new observation $(x_i, y_i)$ is produced, update canonical information.

   (c) Illustrate $\widehat{f(x)}$.

   (d) Illustrate $\mathrm{Var}(\widehat{f(x)})$, assuming that $\sigma^2$ is known.

   (e) Illustrate $\widehat{\mathrm{Var}(\widehat{f(x)})}$, assuming that $\sigma^2$ is NOT known.

   In your report present the source code and a few (around 3) nice graphs showing estimations for "small", "intermediate", and "large" number of observations.