Prosta Linearna Regresija(Deo dokumentacije!!!)

Aleksa Tešić

Decembar 2017.

1 Uvod

U ovom dokumentu će biti prikazana implementacija proste linearne regresije u programskom jeziku c++. Zapravo, na ovaj dokument se najviše može gledati kao na dokumentaciju za implementiranu mini-bibliteku.
Takođe biće dato i uprošćeno objašnjenje samog pojma proste linearne regresije, a za detaljnija objašnjenja biće dati odgovarajući linkovi.

2 Pojam Proste Linearne Regresije

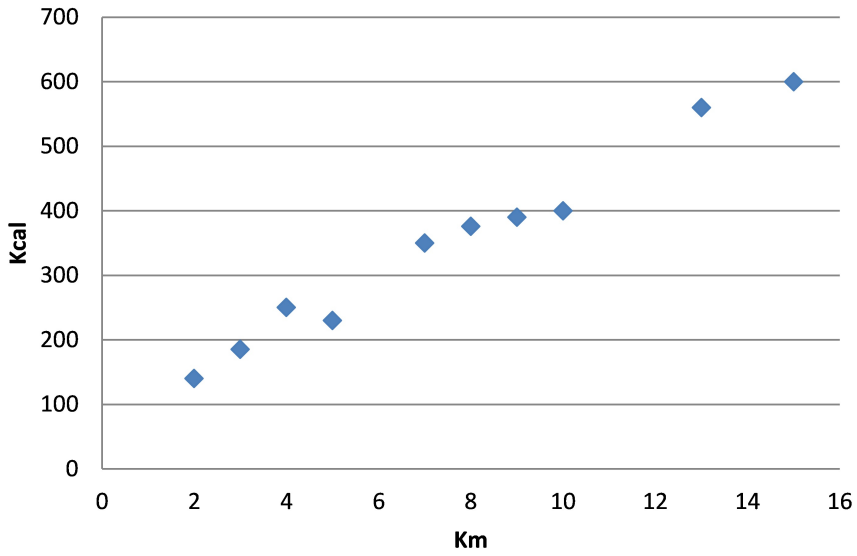
Zamislimo da se, na primer, bavimo trčanjem i 10 dana, svakog dana, istrčimo određeni broj kilometara. I na kraju svakog dana nam je poznato koliko kalorija smo tog dana potrošili na trčanje.
I sada na osnovu podataka izmerenih u toku tih 10 dana želimo da vidimo u kakvoj su vezi broj pretrčanih kilometra i broj potrošenih kalorija da bismo pokušali da predvidimo koliko kalorija očekujemo da ćemo potrošiti ako pretrčimo određen broj kilometara.
Intuitivno možemo primetiti da nam broj kalorija direktno zavisi od broja pretrčanih kilometara. Pa ćemo zato reći da nam je broj kilometara **nezavisna promenljiva**, dok će nam broj kalorija biti **zavisna promenljiva**.
Malo formalniji zapis ovoga bi bio:

$$y = f(x) \tag{1}$$

Ovo znači da se naša zavisna promenljiva y može predstaviti kao nekakva funkcija od nezavisne promenljive x tj. postoji nekakva linearna veza između njih. Još ispravniji oblik ove jednačine bi bio:

$$y = k * x + n \tag{2}$$

Primećujemo da je ova jednačina zapravo jednačina prave. Što će nam uskoro i imati smisla.
Na Slici 1 jedan je iscrtan grafik koji pokazuje odnos između naše 2 promenljive:



Slika 1 Odnos između dve promenljive

Primećujemo da kroz tačke na grafiku skoro da možemo povući pravu liniju. Ali kako tačke ipak nisu potpuno poravnate, moramo na neki način da pronađemo pravu koja najbolje prolazi između naših tačaka, odnosno koordinata. Tu pravu ćemo nazivati **najboljim fitom**.
Upravo za nalaženje najboljeg fita se koristi prosta linearna regresija (Slika 2).

3 Implementacija

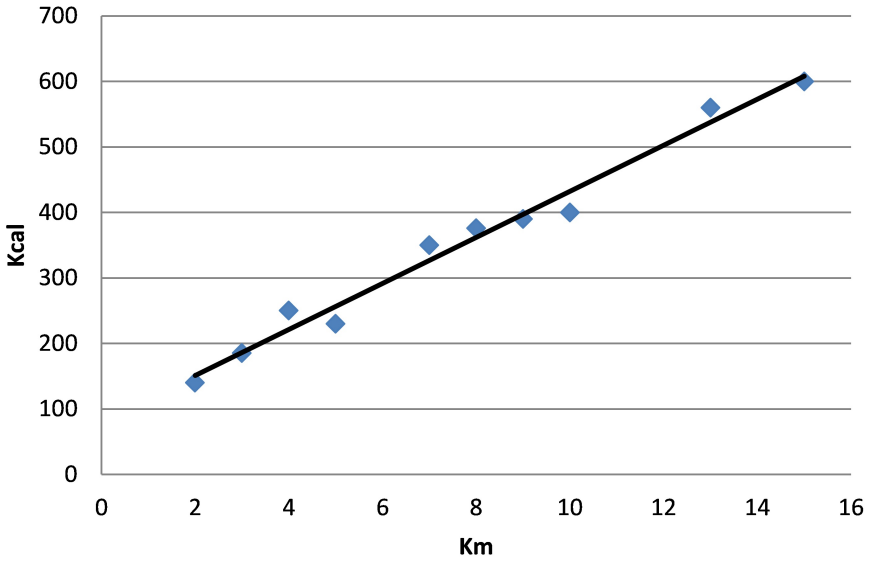
Na početku ćemo prvo ponovo malo modifikovati našu jednačinu prave u oblik:

$$y = b_1 * x + b_o \tag{3}$$

Ovo smo uradili zato što se u literaturi ovo standardne oznake kada se govori prostoj linearnoj regresiji.

3.1 Izračunavanje ukupne kvadratne greške

Pretpostavimo da znamo samo vrednosti zavisnih promenljivih, onda bi najbolji i jedini način za bilo kakvu dalju procenu bio jednostavno nalaženje srednje vrednosti svih promenljivih. Ta srednja vrednost bi se uzimala kao procena za svaku sledeću promenljivu.
Kada smo dobili srednju vrednost želimo da vidimo koliko se ta procena razlikuje od naših 10 unapred poznatih vrednosti (y). Kako namerno hoćemo da



Slika 2 Tražena prava koja je pronađena prostom linearnom regresijom

istaknemo veće greške, oduzećemo vrednosti y od srednje vrednosti i onda ćemo to kvadrirati. Upravo suma svih kvadrata je naša ukupna kvadratna greška - **SST** (eng.*Sum Square Total*).
Mi želimo da nađemo pravu kod koje je ova greška najmanja.

3.2 Izračunavanje centroida

Centroid je tačka unutar koordinatnog sistema u koji možemo postaviti naš grafik čija x koordinata je jednaka sumi nezavisnih (x) promenljivih, odnosno čija y koordinata jednaka sumi zavisnih (y) promenljivih. I za centroid važi da najbolji fit **mora** prolaziti kroz njega.

3.3 Izračunavanje b_1 i b_0

Sada je potrebno da nađemo parametre b_1 i b_0.
Kao što je već rečeno u delu 3.1 mi želimo da nađemo pravu kod koje je greška najmanja. Za tu pravu važi da će suma razlika kvadrata y i y_p biti najmanja, y je poznata vrednost zavisne promenljive, y_p vrednost sa prave. To jest, nama je y_p zapravo ona već pomenuta predviđena vrednost. Parametar b_1 nalazimo pomoću sledeće formule:

$$b_1 = \frac{\sum_{i=0}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=0}^{n}(x_i - \overline{x})^2} \tag{4}$$

U formuli nam je x_i i-ta poznata nezavisna promenljiva, y_i i-ta poznata zavisna promenljiva, dok su nam $\overline{x}$ i $\overline{y}$ srednje vrednosti nezavisnih, odnosno zavisnih promenljivih.
Kako sada imamo b_1 parametar b_0 nalazimo na sledeći način:

$$b_0 = \overline{y} - b_1 * \overline{x} \tag{5}$$

Važe iste oznake za parametre kao i kod nalaženja b_1. Pogledajmo ponovo formulu:

$$y = b_1 * x + b_o \tag{6}$$

kada u nju stavimo naše parametre, odnosno broj kalorija i broj pretrčanih kilometara. Naša jednačina nam kazuje da ćemo na svaki pređeni 1 kilometar potrošiti b_1 kalorija, dok je b_0 broj kalorija koji potrošimo ako pređemo 0 kilometara i b_0 **ne mora** imati smisla kod nekog stvarnog problema.