Linearna regresija i njena primena u ubrzavanju SAT rešavača

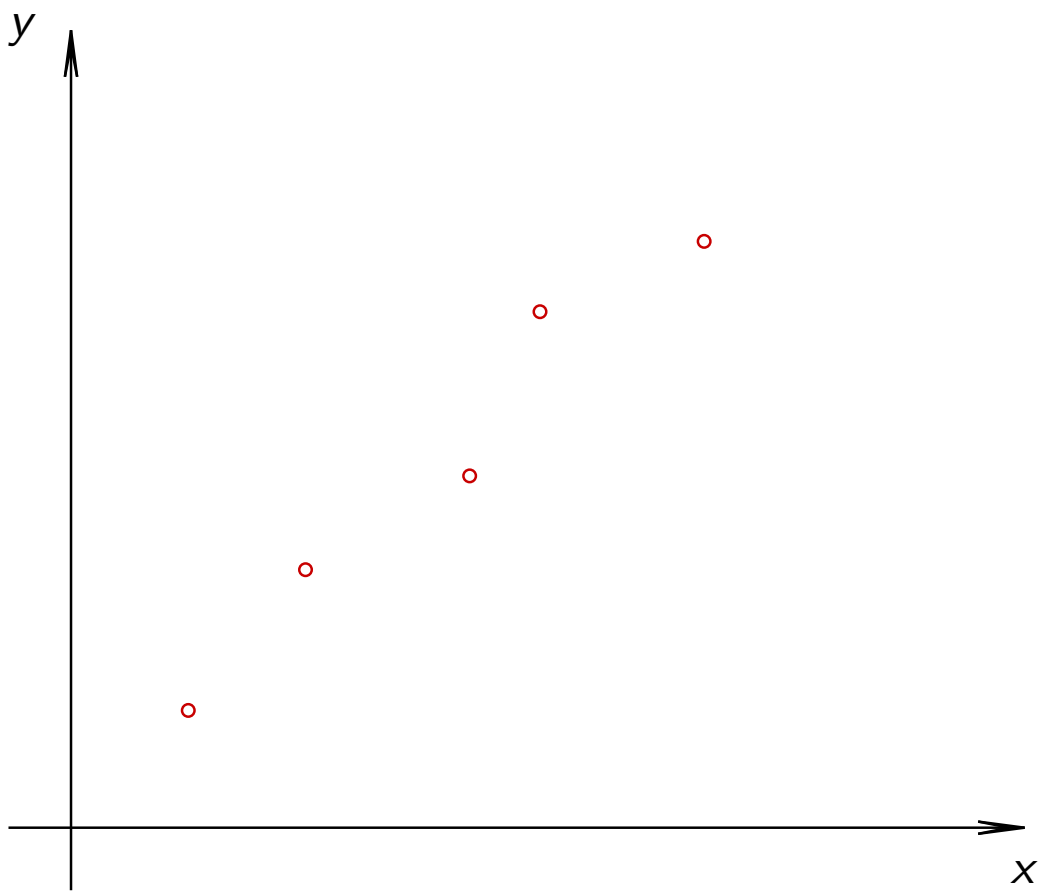
Mladen Nikolić

Linearna regresija i njene varijante

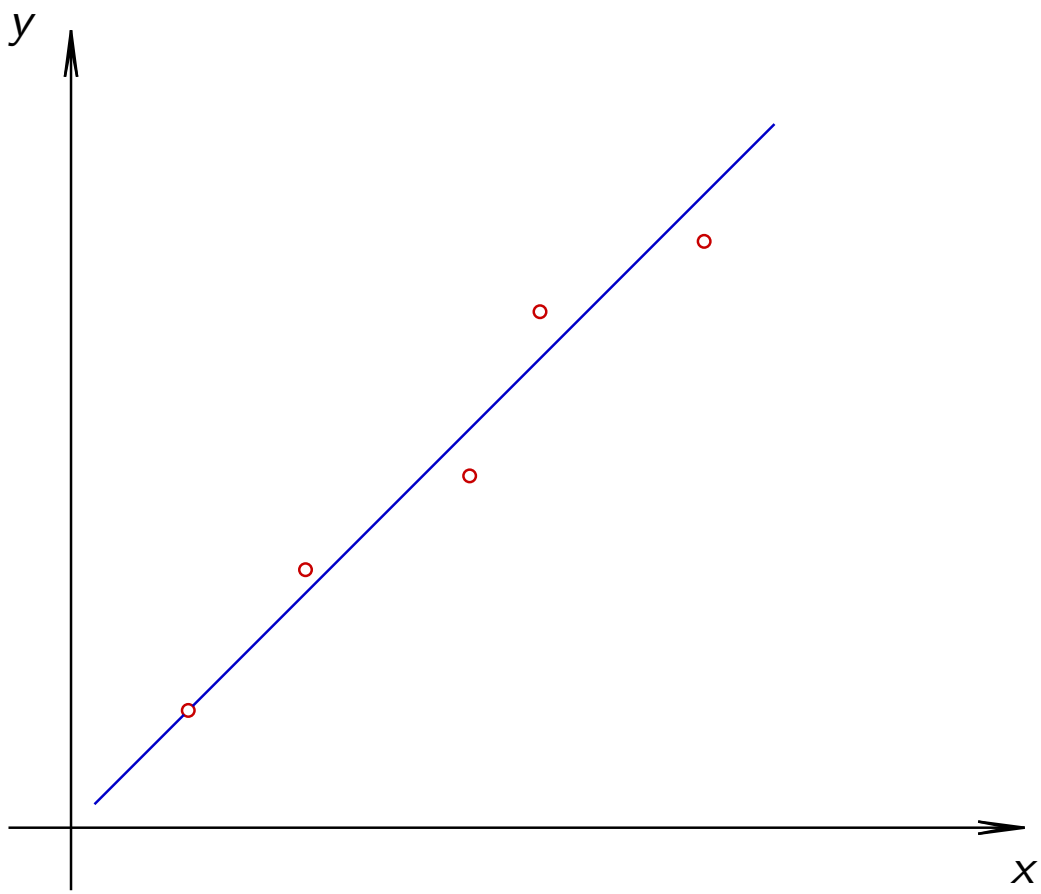
Empirijski modeli težine problema za SAT

Literatura

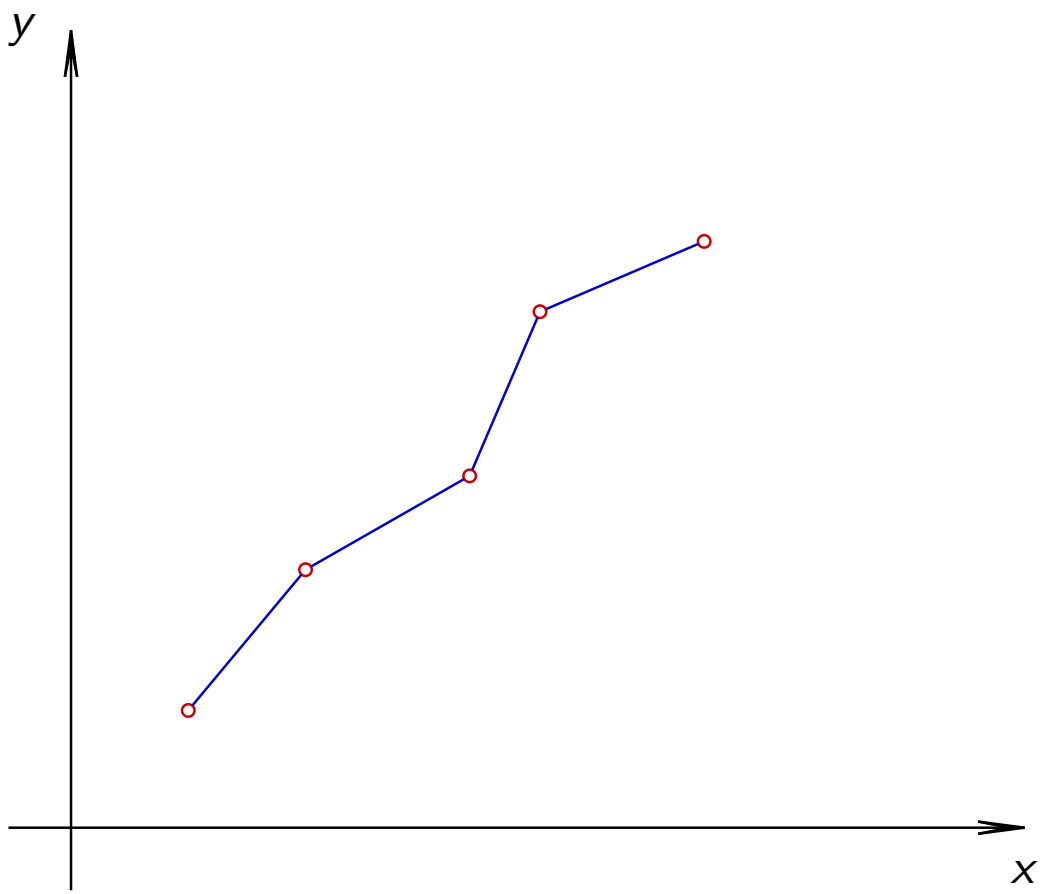
Primer



Primer



Primer



Regresiona funkcija

- X i Y su slučajne promenljive.
- Promenljiva Y uzima realne vrednosti, a promenljiva X može biti i vektorska slučajna promenljiva $X = (X_1, X_2, \ldots, X_n)$.
- Pretpostavlja se da promenljiva Y na neki način zavisi od promenljive X.
- *Regresiona funkcija* $r(x)$ koja povezuje ove dve veličine se definiše na sledeći način

$$r(x) = E(Y|X = x)$$

- Promenljiva X se naziva *prediktorom*, a Y *odzivnom promenljivom*.

Regresija

- U praksi su poznate realizovane vrednosti (X_i, Y_i) slučajnih promenljivih X i Y koje nazivamo *podacima za trening*.
- Na osnovu njih je potrebno približno odrediti regresionu funkciju r. Postupak određivanja regresione funkcije se naziva *regresijom*.
- Funkcija dobijena postupkom regresije se naziva *modelom* datih podataka.

Oblik regresione funkcije

- Regresiona funkcija može imati bilo kakav analitički oblik koji ne mora biti poznat.
- Regresiona funkcija uopšte ne mora imati analitički oblik.
- Potrebno je odrediti neku pogodnu formu regresione funkcije.
- Ova odluka se najčešće donosi na osnovu poznavanja domena iz koga potiču podaci. Na primer, opadajuća eksponencijalna funkcija je pogodna za modeliranje svojstava radioaktivnog raspada.

Izbor regresione funkcije

- Od svih dopustivih funkcija bira se jedna koja u nekom smislu najbolje odgovara podacima.
- Najčešće je skup dopustivih funkcija definisan skupom parametara čijim se fiksiranjem bira jedna konkretna funkcija.
- Potrebno je definisati metod izbora vrednosti parametara na osnovu trening podataka.

Nepreciznost podataka

- Obično se pretpostavlja da postoje problemi u merenju odzivne promenljive.
- Rezultujuće nepreciznosti se modeliraju pretpostavkom da njene vrednosti uključuju normalno raspodeljenu grešku.

Izbor prediktora

- Prediktori ne moraju biti unapred određeni. Oni se mogu birati u skladu sa mogućnostima kojima se u eksperimentu raspolaže.
- Za veći broj prediktora model bolje opisuje podatke za trening.
  - Unosi se veća količina informacije u dobijeni model.
  - Model postaje prilagodljiviji.
- Za preveliki broj prediktora model je kompleksniji i nestabilniji, a zbog velike prilagođenosti podacima za trening nepouzdan na nepoznatim podacima u toku primene.

Jednostavna linearna regresija

- Linearna regresija je postupak određivanja regresione funkcije koja je linearna po parametrima kojima se definiše.
- Osnovni slučaj je takozvana *jednostavna linearna regresija* kod koje je se predpostavlja da je veza između promenljivih X i Y oblika:

$$Y_i = \alpha_0 + \alpha_1 X_i + \epsilon_i$$

a regresiona funkcija:

$$r(x) = \alpha_0 + \alpha_1 x$$

- Pri tome se pretpostavlja da slučajna promenljiva ϵ ima normalnu raspodelu, i da je $E(\epsilon_i) = 0$ i $D(\epsilon_i) = \sigma^2$.

Multipla regresija

- U slučaju da je promenljiva X vektorska, podaci su oblika

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

gde je

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{ik})$$

- Pretpostavlja se da je veza između promenljivih X i Y oblika:

$$Y_i = \sum_{j=1}^{k} \alpha_j X_{ij} + \epsilon_i$$

a da je $X_{i1} = 1$ zbog slobodnog parametra.

- Ova jednačina se jednostavnije zapisuje u matričnom obliku

$$\tilde{Y} = \tilde{X}\alpha + \tilde{\epsilon}$$

Izbor parametara

- Parametri se biraju tako da se minimizuje razlika između predviđenih i stvarnih vrednosti.
- Obično se teži minimizovanju srednjekvadratne greške

$$E((Y - r(X))^2)$$

koja se aproksimira izrazom

$$\frac{1}{n}\sum_i (Y_i - r(X_i))^2$$

- U matričnom zapisu, minimizuje se izraz

$$\|\tilde{X}\hat{\alpha} - \tilde{Y}\|_2$$

Metoda najmanjih kvadrata

- Koeficijenti regresione funkcije za koje je srednjekvadratna greška minimalna se određuju metodom najmanjih kvadrata. Odgovarajuće rešenje je dato formulom:

$$\hat{\alpha} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y}$$

- Ovo je istovremeno i ocena parametara maksimalne verodostojnosti.
- Ovako određeni koeficijenti imaju približno normalnu raspodelu na osnovu koje se može izvesti interval poverenja za predviđene vrednosti.

Evaluacija rešenja

- Evaluacija može biti vršena na odvojenom test skupu.
- Moguće je koristiti unakrsnu validaciju.
- U slučaju linearne regresije suštinski isti rezultati kao za unakrsnu validaciju se mogu dobiti tako što se na grešku izračunatu na trening podacima doda kazna za kompleksnost modela

$$2n\hat{\sigma}^2$$

gde je n broj koeficijenata u regresionoj funkciji.

Stabilnost rešenja

- Jedan od osnovnih problema multiple regresije je to što među prediktorima može postojati linearna zavisnost ili visoka koreliranost. U tom slučaju matrica $\tilde{X}^\top \tilde{X}$ je neinvertibilna ili bar loše uslovljena.
- Ovo dovodi do nepostojanja rešenja ili njegove nestabilnosti.
- Jedan način da se ovo premosti je Tihonovljeva regularizacija.

Tihonovljeva regularizacija

- Ideja Tihonovljeve regularizacije je da se umesto izraza $\|\tilde{X}\hat{\alpha} - \tilde{Y}\|_2$ minimizuje izraz

$$\|\tilde{X}\hat{\alpha} - \tilde{Y}\|_2 + \gamma W(\tilde{X})$$

gde je W preslikavanje oblika $R^n \to R$, a $\gamma > 0$ *parametar regularizacije*.

- Čest izbor za preslikavanje W je $W(X) = \|X\|_2$. U ovom slučaju parametri regresione funkcije su dati formulom:

$$\hat{\alpha} = (\tilde{X}^\top \tilde{X} + \gamma I)^{-1} \tilde{X}^\top \tilde{Y}$$

Zašto je Tihonovljeva regularizacija efikasna?

- Dodavanjem vrednosti $W(\tilde{X})$ dobijena ocena više nije nerpistrasna. Međutim, poznato je da se srednjekvadratna greška može zapisati u obliku

$$(E(\hat{\alpha}) - \alpha)^2 + D(\hat{\alpha})$$

- U slučaju loše uslovljenih problema disperzija ocene parametara može biti vrlo velika.
- Unošenje male pristrasnosti u ocenu značajno smanjuje srednjekvadratnu grešku ako se popravi uslovljenost problema.

Empirijski modeli težine problema

- **NP**-kompletni problemi često imaju instance koje se lako rešavaju.
- Ovakve instance postoje i u praktičnim primenama.
- Zbog toga teorijski rezultati daju previše strogu procenu težine problema.
- Jedno rešenje je u korišćenju *empirijskih modela težine problema*. Oni služe za predviđanje vremena izvršavanja alogoritma ili neke druge mere težine za proizvoljnu instancu problema.

Empirijski modeli težine problema

- Trening podaci se dobijaju tako što se za svaku instancu problema računa vektor unapred određenih prediktora X, a zatim se izabranim algoritmom ona rešava jednom ili više puta i meri se odzivna promenljiva Y.
- Za svako pokretanje algoritma dodaje se u trening skup po jedan par (X, Y).
- Potom se koristi multipla linearna regresija sa regularizacijom kako bi se našla regresiona funkcija.

SATzilla

- Trenutno najuspešniji sistem za rešavanje problema SAT je SATzilla — sistem koji koristi empirijske modele težine kako bi izabrao najbolji SAT rešavač kojim bi rešio datu formulu.
- Po jedan model je treniran za svaki od 7 izabranih SAT rešavača.
- Kada se prilikom upotrebe sistema izračunaju prediktori za datu formulu, procenjuje se vreme koje će biti potrebno svakom od rešavača za rešavanje date formule i primenjuje se rešavač sa najmanjim predviđenim vremenom.

Izbor prediktora za SAT

- Broj klauza c;
- Broj promenljivih v;
- Količnik $\frac{c}{v}$;
- Statistike stepena čvorova koji odgovaraju promenljivim u grafu promenljivih i klauza: prosek, disperzija, minimum, maksimum i entropija;
- Statistike stepena čvorova koji odgovaraju klauzama u grafu promenljivih i klauza: prosek, disperzija, minimum, maksimum i entropija;
- Statistike stepena čvorova u grafu promenljivih: prosek, disperzija, minimum, maksimum i entropija;

Izbor prediktora za SAT

- Odnos pozitivnih i negativnih literala u svakoj klauzi: prosek, disperzija i entropija;
- Odnos pozitivnih i negativnih pojavljivanja svake promenljive: prosek, disperzija, minimum, maksimum i entropija;
- Udeo binarnih klauza;
- Udeo ternarnih klauza;
- Udeo Hornovih klauza;
- Broj pojavljivanja u Hornovim klauzama za svaku od promenljivih: prosek, disperzija, minimum, maksimum i entropija;

Izbor prediktora za SAT

- Broj primena pravila unit propagation procedure DPLL pri dubinama pretrage od 1, 4, 16, 64 i 256;
- Ocena veličine prostora pretrage: prosečna dubina do konflikta i ocena logaritma broja čvorova;
- Broj koraka do najboljeg lokalnog minimuma pri lokalnoj pretrazi stohastičkim SAT rešavačem SAPS pri većem broju pokretanja: prosek, medijana, 10-ti i 90-ti percentil;
- Prosek po svim pokretanjima proseka poboljšanja po koraku do najboljeg rešenja koristeći SAPS;

Izbor prediktora za SAT

- Prosek količnika poboljšanja do prvog lokalnog minimuma i ukupnog poboljšanja pri većem broju pokretanja korišćenjem stohastičkih SAT rešavača SAPS i GSAT;
- Prosek standardnih devijacija broja nezadovoljenih klauza u svakom lokalnom minimumu pri većem broju pokretanja rešavača SAPS.

Korpusi formula

- Za trening korpus sistema SATzilla korišćene su sve formule sa SAT competition takmičenja zaključno sa 2006. godinom. Korišćeno je ukupno 4811 formula.
- Sistem je testiran na takmičenju iz 2007.
- Ovi korpusi su vrlo bogati — obuhvataju formule različitih težina iz više desetina različitih familija.

Rezultati

- U vremenu u kome SATzilla uspeva da reši 92% formula, njegov najbolji komponentni rešavač rešava 72% formula.
- Prosečno vreme rešavanja formula za SATzilla sistem iznosi oko 170s dok za njegov najbolji komponentni rešavač iznosi oko 390s.
- Razlike su veće u slučaju formula iz kategorije RANDOM, ali manje za formule iz kategorije CRAFTED, a vrlo male za formule iz kategorije INDUSTRIAL.

Literatura

- Larry Wasserman, All of Statistics — A Concise Course in Statistical Inference, 2005.
- Vladimir Vapnik, Statistical Learning Theory, 1998.
- http://www.aiaccess.net/x_tut_list_lin_reg.htm
- Lin Xu, Frank Hutter, Holger H. Hoos, Kevin Leyton-Brown, SATzilla: Portfolio-based Algorithm Selection for SAT. JAIR, 2007.