Ciljevi knjige.

Moć data mining-a je otkriće za veliki broj kompanija. Data mining jeste izvlačenje informacija iz bitnih podataka koji se generišu svakog momenta u svakom delu našeg života. U svakodnevnom radu sa podacima mogu se pronaći skriveni šabloni i značenja. Ova knjiga je rezultat detaljnih proučavanja podataka i prikazuje naučene lekcije kada se suočavamo sa podacima i kada ih koristimo da učinimo stvari boljim. Postoji mnogo trikova u trgovini koji osiguravaju efektivne rezultate. Statistička analiza koja je uključena u data mining ima karakteristike koje su drugačije od drugih tipova statistike.

Unutar knjige, pominjemo tipične probleme u prodaji i marketingu, kao što su pronalazak 10% korisnika koji će najverovatnije kupiti specijalan proizvod. Sadržaj se fokusira na prodaju i marketing zato što svi imaju osnovno znanje iz tih oblasti. Svi ljudi su obuhvaćeni kao korisnici i mušterije, tako da se svi možemo pronaći u problemima prodaje i marketinga.

Definicija Data Mining-a

Data mining pokriva širok opseg aktivnosti. On teži da obezbedi odgovor na pitanja kao što su:

* Šta postoji u prikupljenim podacima
* Koje vrste šablona se mogu razlikovati u lavirintu podataka
* Kako se svi ovi podaci mogu koristiti za benefit u budućnosti

Populacija i uzorkovanje

U data mining-u setovi podataka mogu biti ogromni – može postojati na milione slučajeva. Različiti tipovi industrije međutim mnogo se razlikuju u broju slučajeva koje prouzrokuju poslovni procesi. Web aplikacije, naprimer, mogu prikupljati podatke od miliona cookie-ja, dok druge aplikacije, kao što su loyality klubovi ili CRM programi mogu imati više ograničenih slučajeva.

Data mining koristi naučne metode istraživanja i primene. Prezentuje se veliki broj podataka koje u jednom slučaju smatramo za čitavu populaciju. Drugim rečima koristimo sve informacije na raspolaganju. U drugom slučaju, naš set podataka može da se smatra kao veliki uzorak. Ako radimo sa manjom količinom podataka (do 10 000 slučeva), onda se možemo odlučiti za rad sa čitavim setom podataka. Ako radimo sa velikim setom podataka, možda ćemo se odlučiti za podset podataka radi lakše manipulacije. Ako se analiza vrši na uzorku, implikacija je da će rezultati biti reprezentativni za čitavu populaciju. Drugim rečima, rezultati analize nad uzorkom su relevantni za čitavu populaciju.

Ovo znači da uzorak mora biti dobar. Dobar uzorak je onaj koji je reprezentativan i nepristrasan. Uzorkovanje je samo po sebi predmet za izučavanje. Kako uglavnom radimo sa velikim populacijama i možemo priuštiti velike uzorke, možemo uzeti proizvoljan uzorak u kome će svi članovi populacije imati jednake šanse da budu izabrani. Vratićemo se kasnije na praktične probleme u vezi uzorkovanja u drugoj sekciji knjige. Takođe možemo podeliti set podataka u nekoliko uzoraka na kojima možemo testirati naše rezultate. Ako imamo mali set podataka, onda možemo ponovo uzorkovati taj set, uzimajući proizvoljne podsetove u okviru istog uzorka. Ovo se naziva bootstrapping. Tada možemo uzeti u obzir način na koji proveravamo da su rezultujući uzorci reprezentativni.

Ponekad, uzimamo u obzir samo deo populacije, za odredjenu analizu, na primer, mi možda bude zainteresovani samo za ponašanje kupaca oko Božića ili u toku letnjih meseci. U ovom slučaju, podset se referiše kao ram uzorka (sampling frame), i samo će se iz ovog podseta na dalje birati uzorci.

Priprema podataka

Priprema podataka za data mining je vitalni korak koji je ponekad zanemaren. Od najranijih godina, učeni smo da su 2 i 2 jednako 4. Brojevi se smatraju konkretnim, opipljivim, čvrstim, neizbežnim alatom koji se koristi da se izmeri sve. Međutim brojevi imaju inherentne varijacije, naprimer, dva proizvoda se mogu prodati na isti dan, ali njihova cena može biti drugačija. Interpretacije napravljene po nominalnoj vrednosti možda nisu tačne. Neki poslovi koriste podatke za pravljenje odluka bez provere da su podaci značajni, bez prethodne transformacije podataka u znanje i kasnije u inteligenciju. ‘Inteligencija’ dolazi od podataka kojima je potvrđena validnost kroz prethodno iskustvo i opisano je iz razmatranje njegovog konteksta.

Nadgledani i nenadgledani metodi

Data mining je proces koji koristi raznolike metode analize podataka da bi otkrio nepoznate, neočekivane i relevantne šablone i veze među podacima koji se mogu koristiti da bi se napravila tačna i precizna predviđanja. Generalno, postoje dve metode analize podataka: nadgledane (supervised) i nenadgledane (unsupervised). Vidi sliku 2.1 i 2.2. U oba slučaja, uzorak posmatranih podataka se zahteva. Ovi podaci se mogu nazvati kao trening uzorak (training sample). Trening uzorak se koristi u data mining-u kako bi se otkrili šabloni u podacima.

Nadgledane analize podataka se koriste za procenu nepoznatih zavisnosti iz poznatih ulazno izlaznih podataka. Ulazne promenljive mogu uključiti kvanititet različitih artikala koje je kupio određeni kupac, datum kada su obavili kupovinu, lokaciju i cenu koju su platili. Izlazne promenljive mogu uključiti indikaciju od toga da li je kupac odgovorio na kampanju prodaje ili ne. Izlazne promenljive su poznate i kao targeti u data mining-u. U nadgledanim okruženjima, uzorak ulaznih promenljivih se prosledjuje sistemu za učenje i naredni izlaz sistema za učenje se poredi sa izlaznim promenljivim uzorka. Drugim rečima, pokušavamo da predvidimo ko će odgovoriti na prodajnu kampanju. Razlika između izlaza sistema za učenje i izlaza uzorka se može smatrati za signal greške. Signali greške (Error signals) se koriste pri prilagođavanju sistema učenja. Ovaj proces se izvršava više puta sa podacima iz uzorka i sistem učenja se prilagođava dok se izlaz ne dođe do minimalnog praga greške. Isti proces se koristi za završno štimovanje tek kupljenog klavira. Fino podešavanje se može izvršiti uz pomoć eksperta ili korišćenjem nekog elektronskog instrumenta. Ekspert obezbeđuje note za trening uzorak, a tek kupljeni klavir je sistem za učenje. Melodija je usavršena kada se vibracije žica klavira podudaraju sa vibracijama u uhu eksperta.

Unsupervised data analysis does not involve any fine tuning.