

LOGISTIK REGRESSIYA MODELI

Raximberdiyev Sanjabek Alisher o'g'li
Toshkent axborot texnologiyalari universiteti
Urganch filiali Kompyuter injineriingi fakulteti
Axborot xavfsizligi 4-bosqich talabasi

Annotasiya: Logistik regressiya - bu statistika sohasidan mashinani o'rganish orqali olingan yana bir usul. Bu ikkilik tasniflash muammolari ikki sinf qiymati bilan bog'liq muammolar uchun asosiy usuldir. Ushbu postda siz mashinani o'rganish uchun logistik regressiya algoritmini orqali aniqlanadigan kutilgan va kutilayotgan natijalarni va harakatlarni aniqlashda foydalaniladi.

Kalit so'zlar: Regressiya → Modellashtirish, Logistik regressiya → mashinani o'rganish, $EXP()$ → expandenta funksiyasi, Mashinali o'qitish → bu Suniy intellektual yondashuv asosida o'qitish.

Regressiya -- Statistik modellashtirishda regressiya tahlili - bu bog'liq o'zgaruvchi (ko'pincha "natija" yoki "javob" o'zgaruvchisi deb ataladi) va bir yoki bir nechta mustaqil o'zgaruvchilar (ko'pincha "prediktorlar", "kovariatlar" deb ataladi) o'rtasidagi munosabatlarni baholash uchun statistik jarayonlar to'plami. "tushuntiruvchi o'zgaruvchilar" yoki "xususiyatlar"). Regressiya tahlilining eng keng tarqalgan shakli chiziqli regressiya bo'lib, unda ma'lum bir matematik mezon bo'yicha ma'lumotlarga eng mos keladigan chiziq (yoki yanada murakkab chiziqli birikma) topiladi. Masalan, oddiy eng kichik kvadratlar usuli haqiqiy ma'lumotlar va bu chiziq (yoki giperplan) o'rtasidagi kvadrat farqlar yig'indisini minimallashtiradigan noyob chiziqni (yoki giperplanni) hisoblaydi. Muayyan matematik sabablarga ko'ra (chiziqli regressiyaga qarang), bu tadqiqotchiga mustaqil o'zgaruvchilar berilgan qiymatlar to'plamini olganida, bog'liq o'zgaruvchining shartli kutilishini (yoki aholining o'rtacha qiymatini) baholashga imkon beradi. Regressiyaning kamroq tarqalgan shakllari muqobil joylashuv parametrlarini (masalan, kvantli regressiya yoki Kerakli shartlar tahlili[1]) baholash yoki chiziqli bo'lmagan modellarning kengroq to'plami (masalan, parametrik bo'lmagan regressiya) bo'yicha shartli kutishni baholash uchun biroz boshqacha tartiblardan foydalanadi.

Regressiya tahlili asosan ikkita kontseptual maqsadda qo'llaniladi. Birinchidan, regressiya tahlili bashorat qilish va bashorat qilish uchun keng qo'llaniladi, bu erda undan foydalanish mashinani o'rganish sohasi bilan sezilarli darajada mos keladi. Ikkinchidan, ba'zi hollarda regressiya tahlili mustaqil va qaram o'zgaruvchilar o'rtasidagi sabab-oqibat munosabatlarini aniqlash uchun ishlatilishi mumkin. Muhimi shundaki, regressiyalar o'z-o'zidan faqat qat'iy ma'lumotlar to'plamidagi qaram o'zgaruvchi va mustaqil o'zgaruvchilar to'plami o'rtasidagi munosabatlarni ochib beradi. Regressiyalarni bashorat qilish yoki sabab-oqibat munosabatlarini xulosa qilish uchun ishlatish uchun tadqiqotchi nima uchun mavjud munosabatlar yangi kontekst uchun bashorat qilish kuchiga ega ekanligini yoki



nima uchun ikkita o'zgaruvchi o'rtasidagi munosabat sabab-oqibat talqiniga ega ekanligini diqqat bilan asoslashi kerak. Tadqiqotchilar kuzatuv ma'lumotlaridan foydalangan holda sabab-oqibat munosabatlarini baholashga umid qilganda, ikkinchisi ayniqsa muhimdir

Logistik regressiya - bu statistika sohasidan mashinani o'rganish orqali olingan yana bir usul. Bu ikkilik tasniflash muammolari (ikki sinf qiymati bilan bog'liq muammolar) uchun asosiy usuldir. Ushbu postda siz mashinani o'rganish uchun logistik regressiya algoritmini topasiz.

Ushbu maqolani o'qib chiqqandan so'ng, siz quyidagilarni bilib olasiz: Logistik regressiyani tavsiflashda ishlatiladigan ko'plab nomlar va atamalar (masalan, log odds va logit). Logistik regressiya modeli uchun ishlatiladigan vakillik.

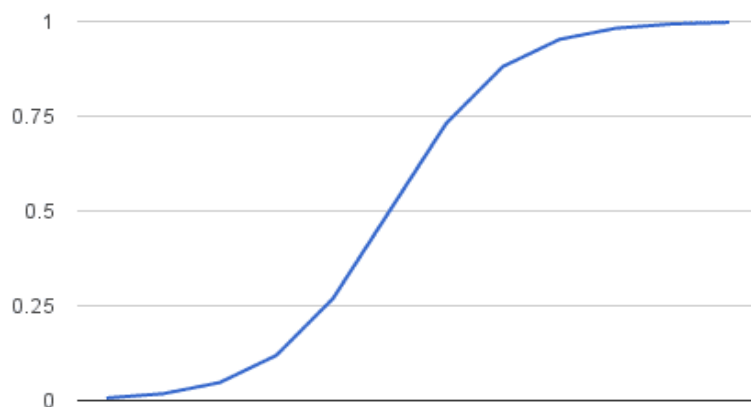
Ma'lumotlardan logistik regressiya modelining koeffitsientlarini o'rganish uchun foydalaniladigan texnikalar. O'rganilgan logistik regressiya modelidan foydalanib, qanday qilib haqiqatda bashorat qilish kerak. Bir oz chuqurroq qazishni istasangiz, qo'shimcha ma'lumot uchun qayerga murojaat qilishingiz kerak. Ushbu post amaliy mashinalarni o'rganishga, xususan, bashoratli modellashtirishga qiziqqan ishlab chiquvchilar uchun yozilgan. Chiziqli algebra yoki statistika bo'yicha ma'lumotga ega bo'lishingiz shart emas. Loyihangizni “Mashina o'rganish algoritmlari” nomli yangi kitobim, jumladan bosqichma-bosqich qo'llanmalar va barcha misollar uchun Excel elektron jadval fayllari bilan boshlang.

Logistika funktsiyasi:

Logistik regressiya usulning o'zagida qo'llaniladigan funktsiya, logistika funktsiyasi uchun nomlanadi. Sigmasimon funktsiya deb ham ataladigan logistik funktsiya statistiklar tomonidan ekologiyada populyatsiyaning o'sishi, tez o'sishi va atrof-muhitning ko'tarish qobiliyatini maksimal darajada oshirish xususiyatlarini tavsiflash uchun ishlab chiqilgan. Bu S shaklidagi egri chiziq bo'lib, u har qanday haqiqiy sonni qabul qilishi va uni 0 dan 1 gacha bo'lgan qiymatga keltirishi mumkin, lekin hech qachon bu chegaralarda aniq emas.

$$1 / (1 + e^{-\text{vauble}})$$

Bu erda e - tabiiy logarifmlarning asosi (Eyler raqami yoki elektron jadvalingizdagi $\text{EXP}()$ funktsiyasi) va qiymat siz o'zgartirmoqchi bo'lgan haqiqiy raqamli qiymatdir. Quyida logistik funktsiyadan foydalangan holda 0 va 1 diapazoniga aylantirilgan -5 dan 5 gacha bo'lgan raqamlar syujeti keltirilgan.



Logistik funktsiya nima ekanligini bilganimizdan so'ng, keling, logistik regressiyada qanday ishlatilishini ko'rib chiqaylik.

Logistik regressiya uchun ishlatiladigan vakillik:

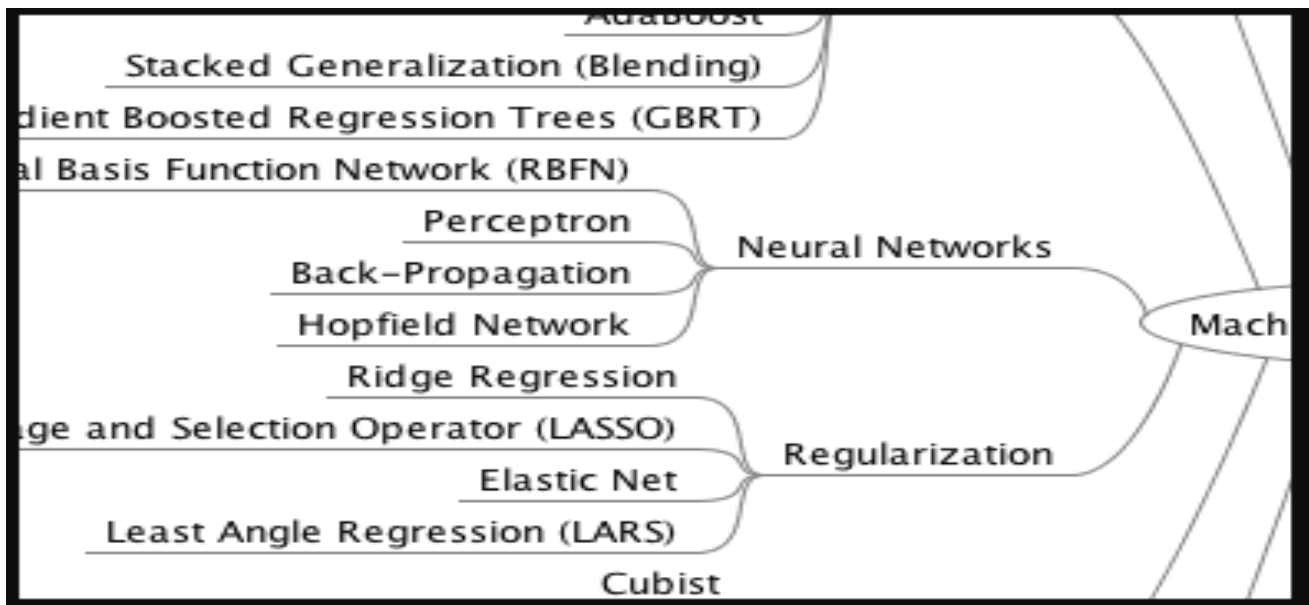
Logistik regressiya chiziqli regressiya kabi tenglamadan vakillik sifatida foydalanadi.

Kirish qiymatlari (x) chiqish qiymatini (y) bashorat qilish uchun og'irliklar yoki koeffitsient qiymatlari (yunoncha bosh harf Beta deb ataladi) yordamida chiziqli tarzda birlashtiriladi. Chiziqli regressiyadan asosiy farq shundaki, modellashtirilayotgan chiqish qiymati raqamli qiymat emas, balki ikkilik qiymatlar (0 yoki 1) hisoblanadi.

Quyida logistik regressiya tenglamasining namunasi keltirilgan:

$$y = e^{(b_0 + b_1 \cdot x)} / (1 + e^{(b_0 + b_1 \cdot x)})$$

Bu erda y prognoz qilingan chiqish bo'lsa, b_0 - egilish yoki kesish muddati va b_1 - bitta kirish qiymati (x) uchun koeffitsient. Kirish ma'lumotlaringizdagi har bir ustunda o'quv ma'lumotlaringizdan o'rganish kerak bo'lgan bog'liq **b** koeffitsienti (doimiy real qiymat) mavjud. Siz xotirada yoki faylda saqlaydigan modelning haqiqiy ko'rinishi tenglamadagi koeffitsientlardir (beta qiymati yoki b).



Logistik regressiya ehtimolliklarni bashorat qiladi (texnik oraliq)

Logistik regressiya standart sinf (masalan, birinchi sinf) ehtimolini modellashtiradi.

Misol uchun, agar biz odamlarning jinsini ularning bo'yi bo'yicha erkak yoki ayol sifatida modellashtirsak, unda birinchi sinf erkak bo'lishi mumkin va logistik regressiya modeli erkakning bo'yi bo'yicha ehtimollik sifatida yozilishi mumkin yoki rasmiyroq:

$$P(\text{sex}=\text{male}|\text{height}):$$

Written another way, we are modeling the probability that an input (X) belongs to the default class (Y=1), we can write this formally as:

$$P(X) = P(Y=1|X)$$

Biz ehtimolliklarni bashorat qilyapmizmi? Logistik regressiyani tasniflash algoritmi deb o'yladimmi?



E'tibor bering, ehtimollik bashorati haqiqatda bashorat qilish uchun ikkilik qiymatlarga (0 yoki 1) aylantirilishi kerak. Bu haqda keyinroq bashorat qilish haqida gapirganda.

Logistik regressiya chiziqli usuldir, ammo bashoratlar logistik funktsiya yordamida o'zgartiriladi. Buning ta'siri shundaki, biz bashoratlarni endi chiziqli regressiya bilan, masalan, yuqoridan davom etgan holda, kirishlarning chiziqli birikmasi sifatida tushuna olmaymiz, modelni quyidagicha ifodalash mumkin:

$$p(X) = e^{(b_0 + b_1 * X)} / (1 + e^{(b_0 + b_1 * X)})$$

Men matematikaga ko'p kirishni xohlamayman, lekin biz yuqoridagi tenglamani quyidagicha aylantirishimiz mumkin (esda tutingki, biz bir tomondan e ni tabiiy logarifmni (\ln) ikkinchisiga qo'shish orqali olib tashlashimiz mumkin):

$$\ln(p(X) / 1 - p(X)) = b_0 + b_1 * X$$

Bu foydali, chunki biz o'ngdagi chiqishni hisoblash yana chiziqli ekanligini ko'rishimiz mumkin (xuddi chiziqli regressiya kabi), chapdagi kirish esa standart sinf ehtimoli jurnali.

Chapdagi bu nisbat standart sinfnings koeffitsientlari deb ataladi (biz koeffitsientlardan foydalanamiz, masalan, ot poygalarida ehtimollikdan ko'ra koeffitsientlar qo'llaniladi). Koeffitsientlar hodisa ehtimolining hodisa bo'lmasligi ehtimoliga bo'lingan nisbati sifatida hisoblanadi, masalan. $0,8/(1-0,8)$ koeffitsienti 4 ga teng. Shuning o'rniga biz yozishimiz mumkin

$$\ln(\text{odds}) = b_0 + b_1 * X$$

Koeffitsientlar jurnalga aylantirilganligi sababli, biz bu chap tomonni log-kod yoki probit deb ataymiz. O'zgartirish uchun boshqa turdagi funktsiyalardan foydalanish mumkin (bu ko'lamdan tashqarida, lekin shuning uchun chiziqli regressiya tenglamasini ehtimollar bilan bog'laydigan transformatsiyani bog'lanish funktsiyasi sifatida ko'rsatish odatiy holdir, masalan, probit bog'lanish funktsiyasi.

Biz predmentni harakatlantira olamiz va o'z holatiga qaytara olamiz va bu shunday amalga oshiriladi:

$$\text{odds} = e^{(b_0 + b_1 * X)}$$

Bularning barchasi bizga model haqiqatan ham kirishlarning chiziqli birikmasi ekanligini tushunishga yordam beradi, lekin bu chiziqli kombinatsiya standart sinfnings log-kodlari bilan bog'liq.

Logistik regressiya modelini o'rganish:

Logistik regressiya algoritmining koeffitsientlari (Beta qiymatlari b) sizning ta'lim ma'lumotlaringizdan hisoblanishi kerak. Bu maksimal ehtimollikni baholash yordamida amalga oshiriladi. Maksimal ehtimollikni baholash - bu turli xil mashinalarni o'rganish algoritmlari tomonidan qo'llaniladigan umumiy o'rganish algoritmi, garchi u sizning ma'lumotlaringizni taqsimlash haqida taxminlar qilsa ham bu haqda ko'proq ma'lumotlaringizni tayyorlash haqida gapiramiz. Eng yaxshi koeffitsientlar standart sinf uchun 1 ga (masalan, erkak) va boshqa sinf uchun 0 ga (masalan, ayol) juda yaqin qiymatni



bashorat qiladigan modelga olib keladi. Logistik regressiya uchun maksimal ehtimollik sezgisi shundan iboratki, qidiruv protsedurasi model tomonidan bashorat qilingan ehtimollar xatosini ma'lumotlardagilarga (masalan, agar ma'lumotlar asosiy bo'lsa, 1 ehtimolli) minimallashtiradigan koeffitsientlar (Beta qiymatlari) qiymatlarini qidiradi. sinf).

Biz maksimal ehtimollik matematikasiga kirmoqchi emasmiz. Ta'lim ma'lumotlaringiz uchun koeffitsientlarning eng yaxshi qiymatlarini optimallashtirish uchun minimallashtirish algoritmidan foydalanilganligini aytish kifoya. Bu ko'pincha amalda samarali raqamli optimallashtirish algoritmidan foydalangan holda amalga oshiriladi (Kvazi-Nyuton usuli kabi).

Logistikani o'rganayotganda, uni o'zingiz noldan ancha sodda gradient tushish algoritmi yordamida amalga oshirishingiz mumkin.

Logistik regressiya yordamida bashorat qilish

Logistik regressiya modeli bilan bashorat qilish logistik regressiya tenglamasiga raqamlarni kiritish va natijani hisoblash kabi oddiy.

Keling, buni aniq misol bilan aniqlaylik:

Aytaylik, bizda odamning bo'yi bo'yicha erkak yoki ayol ekanligini bashorat qila oladigan model mavjud (butunlay xayoliy). 150 sm balandlikda erkak yoki ayol bo'ladi. Biz $b_0 = -100$ va $b_1 = 0,6$ koeffitsientlarini bilib oldik. Yuqoridagi tenglamadan foydalanib, 150 sm yoki undan ko'proq rasman P (erkak|bo'yi = 150) bo'lgan erkakning ehtimolini hisoblashimiz mumkin. Biz e uchun EXP() dan foydalanamiz, chunki siz ushbu misolni elektron jadvalingizga kiritsangiz, undan foydalanishingiz mumkin:

$$y = e^{(b_0 + b_1 \cdot X)} / (1 + e^{(b_0 + b_1 \cdot X)})$$

$$y = \exp(-100 + 0.6 \cdot 150) / (1 + \exp(-100 + 0.6 \cdot X))$$

$$y = 0.0000453978687$$

Yoki bu odamning erkak bo'lish ehtimoli nolga yaqin.

Amalda biz to'g'ridan-to'g'ri ehtimolliklardan foydalanishimiz mumkin. Bu tasnif bo'lgani uchun va biz aniq javobni xohlaymiz, ehtimolliklarni ikkilik sinf qiymatiga bog'lashimiz mumkin, masalan:

$$0, \text{ if } p(\text{male}) < 0,5$$

$$1, \text{ if } p(\text{male}) \geq 0,5$$

Logistik regressiya yordamida bashorat qilishni bilganimizdan so'ng, texnikadan maksimal darajada foydalanish uchun ma'lumotlarimizni qanday tayyorlashni ko'rib chiqaylik. Logistik regressiya uchun ma'lumotlarni tayyorlang Logistik regressiya orqali ma'lumotlaringizdagi taqsimot va munosabatlar haqidagi taxminlar chiziqli regressiyada qilingan taxminlar bilan deyarli bir xil.

Ushbu taxminlarni aniqlash bo'yicha ko'p tadqiqotlar olib borildi va aniq ehtimollik va statistik tildan foydalanildi. Mening maslahatim shundan iboratki, ulardan ko'rsatmalar yoki qoidalar sifatida foydalanish va turli xil ma'lumotlarni tayyorlash sxemalari bilan tajriba qilish. Oxir oqibat, bashoratli modellashtirish mashinalarini



o'rganish loyihalarida siz natijalarni sharhlashdan ko'ra aniq prognozlar qilishga qaratilgansiz. Shunday qilib, model mustahkam va yaxshi ishlashi sharti bilan siz ba'zi taxminlarni buzishingiz mumkin.

Ikkilik chiqish o'zgaruvchisi: Bu allaqachon aytib o'tganimizdek aniq bo'lishi mumkin, ammo logistik regressiya ikkilik (ikki sinf) tasniflash muammolari uchun mo'ljallangan. U 0 yoki 1 tasnifiga birlashtirilishi mumkin bo'lgan standart sinfga tegishli bo'lgan namunaning ehtimolini bashorat qiladi.

Shovqinni olib tashlash: Logistik regressiya chiqish o'zgaruvchisida (y) xatolik yo'qligini ko'zda tutadi, o'quv ma'lumotlaridan chet va ehtimol noto'g'ri tasniflangan misollarni olib tashlashni ko'rib chiqing.

Gauss taqsimoti: Logistik regressiya chiziqli algoritmdir (chiqarishda chiziqli bo'lmagan transformatsiyaga ega). U kirish o'zgaruvchilari bilan chiqish o'rtasida chiziqli munosabatni nazarda tutadi. Ushbu chiziqli munosabatlarni yaxshiroq ko'rsatadigan kirish o'zgaruvchilaringizning ma'lumotlarini o'zgartirish aniqroq modelga olib kelishi mumkin. Masalan, ushbu munosabatlarni yaxshiroq ochish uchun log, root, Box-Cox va boshqa bir o'zgaruvchan transformatsiyalardan foydalanishingiz mumkin.

Tegishli kirishlarni olib tashlash: Chiziqli regressiya kabi, agar sizda bir nechta yuqori korrelyatsiya qilingan kirishlar bo'lsa, model ortiqcha moslashishi mumkin. Barcha kirishlar orasidagi juft korrelyatsiyalarni hisoblash va yuqori korrelyatsiya qilingan kirishlarni olib tashlashni ko'rib chiqing.

Birlasha olmasligi: Koeffitsientlarni o'rganadigan kutilgan ehtimollikni baholash jarayoni yaqinlashmasligi mumkin. Agar ma'lumotlaringizda juda ko'p o'zaro bog'liq bo'lgan kirishlar mavjud bo'lsa yoki ma'lumotlar juda siyrak bo'lsa (masalan, kiritilgan ma'lumotlaringizda juda ko'p nollar bo'lsa) bu sodir bo'lishi mumkin.

FOYDALANILGAN ADABIYOTLAR:

1. 100 days of Machine learning Siraj Raval
2. Data-science engineer software Tim Menzies
3. Technical Strategy for AI Engineer Andrew NG
4. Artificial intelligence Kurzweil -- 2017 yil
5. Learning GANs Roadmap O'reilly Media tomonidan yigilgan
6. Machine Learning for dummies Mueller, John Paul, Massaron
7. Thoughtful Machine Learning O'reilly media Matthew Krik
8. Rules of Machine Learning O'reilly Media tomonidan yigilgan
9. Nuclei T+ Tony O'brien

