Log Lineer Modeller ve Kosulsal Rasgele Alanlar (Log Linear Models and Conditional Random Fields)

Ders 2

Charles Elkan ders notlari

Kosulsal Olurluk (Conditional Likelihood)

Diyelim ki elimizde egitim verisi olarak ikili $< x, y >$ veri noktalari var. O zaman y'nin x'e kosulsal olarak bagli (conditional on) bir dagilimi oldugunu soyleyebiliriz.

$$y \sim f(x; \theta)$$

Yani her x icin farkli bir y dagilimi ortaya cikabilir. Ve tum bu farkli dagilimlarin ortak noktasi θ parametresidir. Kosulsal olasilik yani soyle yazilabilir,

$$P(Y = y|X = x; \theta)$$

Usttekiler Y icin bir model ortaya koydu, peki elimizde X'in dagilimi icin bir olasilik modelimiz var mi? Cevap hayir. Niye? Dusunelim, $p(y, x)$ nedir ?

$$p(x, y) = p(x)p(y|x)$$

Ustte $p(y|x)$'i tanimlayacak (θ uzerinden) bir olasilik demeti / ailesi tanimladik, fakat elimizde $p(x)$ dagilimini verecek bir model yok, o zaman $p(x, y)$'yi tanimlayacak bir model de yok.

Fakat bu dunyanin sonu degil. Belki de Makine Ogrenimi bransinin bir slogani su olmali: "Ogrenmen gerekmeyen seyi ogrenme". Ustteki ornekte $p(y|x)$'i ogrenebiliriz, ama $p(x)$'i illa ogrenmemiz gerekir mi?

Siniflayici (classifier) ve takip edilen (supervised) ogrenim durumunu dusunursek, bize egitim amacli olarak $< x, y >$ ikili veri noktalari saglanacak. x kaynak veri, y tahmin edilecek (ya da basta egitim hedefi olan) etiket olacak. y icin bir model ortaya cikartiyoruz, cunku test zamaninda y olmayacak, fakat x hep olacak. Yani y'nin modellenmesi mecburi, cunku "genelleyerek" onun ne oldugunu bulacagiz, ama x hep verili.

Kosulsal Olurluk Maksimum Olurluk Prensibi

Egitim verisi $< x_1, y_1 >, ..., < x_n, y_n >$ icin, θ'yi soyle sec

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} p(y_i|x_i; \theta)$$

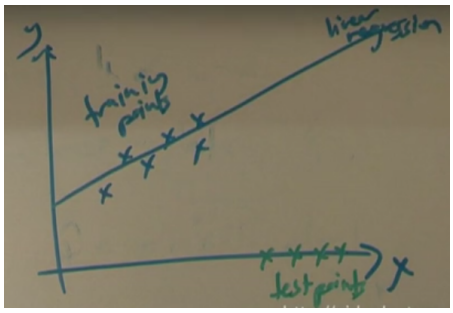
Normal maksimum olurlukta bilindigi gibi olasiliklarin carpimi maksimize edilir, burada maksimize ettigimiz "kosulsal" olasiliklarin carpimi.

Burada onemli bir soru su: bildigimiz gibi maksimum olurluk hesabi her veri noktasinin bir digerinden bagimsiz oldugunu farzeder [cunku her olurluk hesabini bir diger ile carpiyoruz, baska ek carpim, toplama, vs yapmiyoruz], bu faraziye dogru bir faraziye midir? Bu soru ve ona verilecek cevap cok onemli. Evet, eger egitim noktalari birbirinden bagimsiz degilse maksimum olurluk kullanmamaliyiz. Bagimsizligi da iyi tanimlamak gerekiyor tabii, eger usttteki durumda x_i *verildikten sonra* y_i'larin birbirinden bagimsiz olmasi yeterli.

Bu model klasik Istatistik'te cokca kullanilan bir yaklasimdir, hatta lineer regresyon'un temeli usttteki faraziyedir.

$$y = \alpha + \bar{\beta}\bar{x} + N(0, \sigma^2)$$

Bu standart lineer regresyon modeli, ve bu modelde her y ona tekabul eden x'e bagli, bu sayede x'ler biliniyorsa y'ler birbirinden kosulsal olarak bagimsiz hale geliyor, boylece x'ler birbirine bagimli olsa bile α ve β'nin bulunmasi mumkun oluyor.



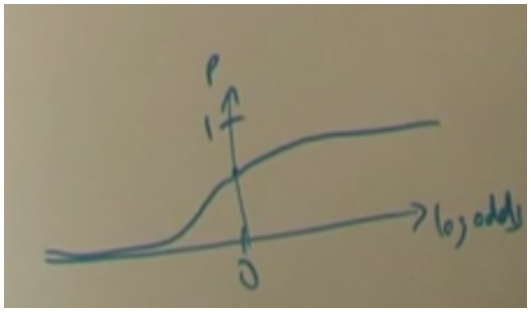
Usttteki resimde egitim noktalari (training points) mavi olsun, test noktalari yesil olsun (hemen altinda). Bazi Yapay Ogrenim yaklasimlari diyebilir ki egitim x'lerinin dagilimi test x'lerinin dagilimindan farkli, bu veri seti ogrenilemez (yani genellenemez, modellenemez). Fakat klasik Istatistik buna bakar ve der ki x'lerin verildigi durumda y'ler bagimsizdir, bu sekilde bir kosulsal model ogrenilebilir.

Lojistik Regresyon ayni sekilde isler (lojistik regresyon, log lineer modellerin ozel bir halidir, Kosulsal Rasgele Alanlar ayni sekilde). Burada da ogrenilen bir

$$p = p(y|x; \alpha, \beta)$$

modeli vardir ve y degerleri sadece 0 ve 1 olabilir. Tahmin edilen olasilik ise y'nin 1 olma olasiligidir. Bu model Rasgele Gradyan Cikisi ile egitilir [detaylar icin *Lojistik Regresyon* notlarimiza bakabilirsiniz].

$$log\frac{p}{1-p} = \alpha + \sum_j \beta_j x_j$$



p log sansinin monotonik bir fonksiyonudur, ve ters yonden bakarsak, log sans p'nin monotonik bir fonksiyonudur. Yani lineer bir fonksiyon (sag taraf) ne kadar buyurse, olasilik / log sans o kadar buyuyecektir. Bu buyume durumu mesela β_j katsayisini veri analizi baglaminda yorumlanabilir hale getirir. Diyelim ki β_4 katsayisi pozitif, o zaman diger tum sartlarin esit oldugu durumda (with all else being equal) x_4 ne kadar buyurse 1 olma olasiligi o kadar artar.

Lojistik modellerin onemli bazi avantajlari var, ki bu avantajlar log lineer modellere de sirayet ediyor (bu iyi).

1) Degiskenler arasi ilinti (correlation) probleme yol acmaz: Bu fayda aslinda daha once belirttigimiz x'lerin birbirine bagimli olabilmesi ile alakali. Bagimsizlik onsarti aranmadigi icin istedigimiz kadar x'i problemin uzerine atabiliriz, egitici algoritma bunlardan cikartabildigi kadar iyi bir model bulacaktir.

Kiyasla mesela Naive Bayes boyle degildir, eger bir NB siniflayicisini egitiyorsak, ve ogelerin (feature) arasinda ilinti var ise, siniflayicinin dogrulugu (accuracy) azalabilir.

2) LR ile "1 olma olasiligini", yani "bir sayisal skoru", elde ediyoruz, bu sadece 1/0 degerinden daha fazla bir bilgi demektir.

3) Bu skor, anlami olan bir olasiliksal degerdir: Sonucta SVM siniflayicilari da $-\infty$ ve $+\infty$ arasinda degerler dondururler, ve bu degerler siralama (ranking) amacli kullanilabilir, fakat olasilik matematigi acisindan anlami olan bir degerin olmasi bundan bile iyidir. Naive Bayes 0 ve 1 arasinda deger dondurebilir, fakat bu degerlerin de olasiliksal olarak aslinda anlami yoktur, pratikte goruldu ki bu degerler cok uc noktalarda, ya sifira cok yakin, ya bire cok yakin. Literaturde NB skorlarinin "iyi kalibre edilmis olmadigi" soylenir.

$X_1, ..., X_n$ test ornekleri ve tahmin edilen olasiliklar $P(Y = 1|x_i) = v_i$ olsun. Diyelim ki $s = \sum_i v_i$ ve t sayisi $1, .., n$ tane ogenin icinden $y = 1$ degerini tasiyan ogelerin sayisi olsun. Ornek, elimizde 100 tane egitim noktasi var, bunlarin 60'i 1 degerinde. Bu durumda s yaklasik 60 olacaktir (rasgele gurultuyu hesaba katarsak tabii), yani $E[t] = s$ denebilecektir ve bu sadece eger olasiliklar iyi kalibre edilmisse soylenebilir.

4) Dengesiz egitim verisi kullanilabilir: pek cok egitim setinde mesela 1 degeri tasiyan degerleri 0 degeri tasiyanlardan cok daha fazla. Lojistik regresyon bu tur veriyle rahatca calisabilir.

Ders 3

Lojistik regresyon icin log olurlugun (LCL) turevini almak lazim. Once basitlestirme amacli $\alpha = \beta_o$, ve $x_0 = 1$. O zaman log sansin eski hali (altta esitligin sol tarafi) soyle yazilabilir (sag taraf), daha derli toplu bir formul olur,

$$\alpha + \sum_j \beta_j x_j = \sum_{j=0}^{d} \beta_j x_j$$

Bulmak istedigim her j icin $\frac{d}{d\beta_j} LCL$ lazim

$$\frac{d}{d\beta_j} LCL = \sum_{i:y_i=1} \frac{d}{d\beta_j} \log p(1|..) + \sum_{i:y_i=0} \frac{d}{d\beta_j} \log p(0|..) \quad (3)$$

Eger ustteki bir bolumu p digerine $1-p$ dersem, yani soyle

$$= \sum_{i:y_i=1} \frac{d}{d\beta_j} \underbrace{\log p(1|..)}_{p} + \sum_{i:y_i=0} \frac{d}{d\beta_j} \underbrace{\log p(0|..)}_{1-p}$$

O zaman

$$= \sum_{i:y_i=1} \frac{d}{d\beta_j} \log p + \sum_{i:y_i=0} \frac{d}{d\beta_j} \log(1-p)$$

Biliyoruz ki

$$\frac{d}{d\beta_j} \log p = \frac{1}{p} \frac{d}{d\beta_j} p \quad (1)$$

$$\frac{d}{d\beta_j} \log(1-p) = \frac{1}{1-p}(-1) \frac{d}{d\beta_j} p \quad (2)$$

Ustteki son iki formulun her ikisinde de $d/d\beta_j p$ kismi olduguna dikkat.

Notasyon

$$e = \exp[-\sum_{j=0}^{n} \beta_j x_j]$$

$$p = \frac{1}{1+e}$$

$$1 - p = \frac{1+e-1}{1+e} = \frac{e}{1+e}$$

Simdi $d/d\beta_j p$'e donelim, ve p'nin ustteki gibi oldugundan hareketle,

$$\frac{d}{d\beta_j}p = (-1)(1+e)^{-2}\frac{d}{d\beta_j}e$$

$$= (-1)(1+e)^{-2}(e)\frac{d}{d\beta_j}(x_j)$$

$$= \frac{1}{1+e}\frac{e}{1+e}x_j = p(1-p)x_j$$

Son ifade kodlama icin oldukca uygun, $d/d\beta_j p$ hesabini yine icinde p iceren bir ifadeye bagladik, ayrica turev x_j ile orantili.

Bu hesapla aslinda (1) icindeki $d/d\beta_j p$ kismini hesaplamis olduk. Eger yerine koyarsak,

$$\frac{d}{d\beta_j}\log p = \frac{1}{p}p(1-p)x_j$$

p'ler iptal olur

$$= (1-p)x_j$$

Ayni sekilde (2) icin

$$\frac{d}{d\beta_j}\log(1-p) = \frac{1}{1-p}(-1)p(1-p)x_j$$

$$= -px_j$$

Ustteki turevler tek bir egitim veri noktasi icin. Tum egitim veri setinin turevi her noktanin turevlerinin toplami olacak, (3)'de goruldugu gibi.

$$\frac{d}{d\beta_j}LCL = \sum_{i:y_i=1}(1-p_i)x_{ij} + \sum_{i:y_i=0} -p_i x_{ij} \quad (4)$$

x_{ij} notasyonunda j, j^{inci} oge / ozellik anlamina geliyor. Simdi notasyonel bir numara kullanacagim,

$$= \sum_{tum\ i}(y_i - p_i)x_{ij}$$

Bunu niye yaptim? (4) formulunde esitligin sag tarafi, birinci terim icinde 1 sayisi var, sonraki terimde 1 yok. Eger 1 olup olmamasi yerine y_i kullanirsam, ki zaten 1'in olup olmamasi y_i'nin 1 olup olmamasina bagli, tek bir terimde isi halledebilirim. $y_i = 1$ oldugu zaman usttteki ifade $1 - p_i$ olacaktir, olmadigi zaman $-p_i$ olacaktir.

Eristigimiz sonucu analiz etmemiz gerekirse, nihai formul gayet basit ve temiz cikti.

[24:10] kalibrasyonla alakali bir yorum

Rasgele Gradyan Cikisi (Stochastic Gradient Ascent)

Fikir: turevi egitim noktasi basina hesapla, ve modeli hemen guncelle.

Egitim noktalari $< x, y >$ olarak gelsinler. Her nokta icin, ve her β_j icin

$$\frac{d}{d\beta_j} p(y|x;\beta) = g_j$$

hesapla.

$$\beta_j := \beta_j + \alpha g_j$$

Gradyanin ne oldugunu hatirlayalim, bir fonksiyonun maksimumuna "dogru" olan bir gidis yonunu gosterir, ve bu gidis yonu o fonksiyonu olusturan degiskenlerin (parcali turevleri) uzerinden belirtilir. O zaman elimizdeki gradyan o ic degiskenlerin maksimum yondeki degisim seklini bize tarif eder.

Algoritmanin tamami: alttaki formul icin

$$\frac{d}{d\beta_j} p(y|\bar{x};\bar{\beta}) = (y - p)x_j$$

Her x icin

- O anki modele gore p'yi hesapla

- Her $j = 0, .., d$ icin

  - $\beta_j := \beta_j + \alpha \underbrace{(y - p)x_j}_{kismi\ turev}$ hesapla

Peki metotun ismindeki "rasgele (stochastic)" tanimi nereden geliyor? Iyi bir soru bu cunku metotta rasgele sayi uretimi gibi seyler gormuyoruz. Cevap, metot yine de rasgele, cunku her noktayi ayri ayri isliyoruz, ve bu noktalarin egitim algoritmasini gelisi bir nevi "veriyi orneklemek" gibi sanki, ek olarak veriyi egitime almadan once rasgele sekilde karistirmak ta iyi olabilir.

Bazi Tavsiyeler (Heuristics)

1) Her ozellik (feature) x_j'i olceklemek, yani ayni ortalama (mean) ve varyansa sahip olacak sekilde tekrar ayarlamak. Yani mesela 0 ile 100 arasinda olabilecek "yas" gibi

bir ozelligi, 0 ve 1 arasinda degisen ozellikler ile ayni ortalama ve varyansa sahip olacak sekilde ayarlamak. Bunun sebebi guncelleme hesabindaki λ'nin tek bir sabit olmasi, ve bu sabit her j icin aynidir, o sebeple λ'nin her ogeye "ayni sekilde" uygulanabilmesi icin ogelerin birbirine yakin olmasi iyidir. Ek olarak, genellikle egitim verisinde 0 ile 1 arasinda ikisel turden ogeler vardir, o sebeple bu sekilde olmayan diger ogeleri 0 ve 1 arasinda cekmek daha uygun ve kolay olur.

2) Veriyi rasgele sekilde siralamak. Terminoloji: egitim veri seti uzerinden bir gecis yapmak bir "cag" (epoch) olarak bilinir.

3) λ'yi deneme / yanilma yontemi ile bulun (bu sabiti bulmanin sistemik bir yontemi yok). Belki verinin icinden alinan daha ufak bir orneklem uzerinde bu deneme / yanilma islemi yapilabilir.

4) Deneme yanilma islemini soyle yapabilirsiniz: buyuk bir λ ile ise baslarsiniz, ve her cagda λ degerini azaltabilirsiniz (mesela her cag sonunda $1/2$ ile carparak).

Ders 4

Log Lineer Modeller

Bu modeller lojistik regresyonun yapiya sahip (structured) girdiler ve ciktilar icin genellenmis halidir. Lojistik regresyonda girdi $\bar{x} \in \mathbb{R}^d$ ve cikti $y \in 0, 1$ idi, yani cikti ikiseldi. Fakat biz bundan daha genel makine ogrenimi problemlerini cozmek istiyoruz, yani istedigimiz $x \in \mathbb{X}$, ki $\mathbb{X}$ herhangi bir uzay olabilmeli, ve $y \in \mathbb{Y}$ ki $\mathbb{Y}$ ayni sekilde herhangi bir uzay olabilmeli.

Mesela x bir cumle olabilmeli, diyelim ki $x =$ "he sat on the mat", tercumesi "adam paspasin uzerinde oturdu". Buna karsilik olan y ise mesela soyle olabilmeli, $y =$ "pronoun verb article noun", yani her kelimenin hangi gramer ogesi oldugunu gosteren bir ibare. Mesela "sat" yani oturmak, bir fiil (verb), "mat" paspas, bir isim (noun), ve y icinde gelen egitim verisinde bunlar olabilmeli (ustteki ornekte ikinci oge), sadece 0/1 degerleri degil.

Bu tabii ki takip edilen (supervised) bir egitim sekli olacak. Fakat dikkat bazi makine ogrenimi uygulamalarinda "cok siniftan gelen" ama tek bir deger vardir, mesela $y \in 1, 2, 3$ olabilir, 3 sinifli bir cikti yani. Bazen cikti gercek sayi (real number) olabilir, ama yine de tek bir y degeri vardir. Ustteki durum boyle degildir. Potansiyel olarak y'nin buyuklugu x ile birebir ayni bile olmayabilir. Bu tur bir karisik eslemeden bahsediyoruz. Tek sinirlamamiz $\mathbb{Y}$'nin sonlu (finite) olmasi.

Model soyle (notasyonu biraz degistirdik, β yerine w kullaniyoruz mesela, w modelin "agirliklarini (weights)" temsil ediyor.

$$p(y|x; w) = \frac{\exp\left[\sum_j w_j F_j(x, y)\right]}{Z(x, w)}$$

Yakindan bakarsak model LR modeline benziyor. Bir lineer fonksiyonun exp'si aliniyor ve bu deger olasilik hesabinda kullaniliyor. Ileride zaten gorecegiz ki LR ustteki yaklasimin bir "ozel durumu", yani ustteki model daha genel bir tanim.

Aklimiza bircok soru geliyor herhalde, mesela "Z nedir?" ya da "F_j nasil hesaplanir?" gibi. Z soyle tanimlanir

$$Z(x, w) = \sum_{y'} \exp\,[\sum_{j} w_j F_j(x, y')]$$

Tum y''lere bakiliyor, yani tum mumkun $\mathbb{Y}$ degerleri teker teker y' uzerinden toplamda kullaniliyor. $\mathbb{Y}$'nin sonlu olma faraziyesi burada onemli hale geliyor, toplami sonsuz bir kume uzerinden yapamayiz.

Z normalizasyon icin kullaniliyor, cunku olasilik teorisinde eger elimizde coklu bir hedef var ise, bu hedeflere olan olasilik degerlerinin toplami 1 olmalidir. Z iste bunu garantiler, bu sebeple bolen (denominator) bolumun (nominator) toplami olmalidir.

Her $F_j(x, y)$ bir ozellik fonksiyonudur (feature function). Niye? Cunku elimdeki x'ler illa bir vektor olmayabilir, yani x_j "vektorunu" alip w_j "vektoru" ile carpamam, bu sebeple once bir fonksiyon ile bir numerik deger uretmem gerekiyor. Kume olarak

$$F_J : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$$

Eger $F_j(x, y) > 0$ ve $w_i > 0$ ise, o zaman $F_j(x, y) = 0$'a kiyasla $p(y|x; w)$ artar. Sezgisel olarak tarif edersek ozellik fonksiyonun soyledigi sudur, eger agirlik pozitif ise ozellik fonksiyonunun degeri ne kadar buyurse elimizdeki y, x ile o kadar "uyumludur" (tabii ki belli bir ozellik yani j icin). Negatif ilinti bunun tam tersi olurdu.

Egitim w_j agirliklarini bulmamizi saglar. F onceden tanimlidir (yani egitime bile baslamadan once), bu fonksiyonun ne olacagi "secilir". Secilirken tabii ki x, y arasindaki ilintiye gore fazla / az sonuc geri getirebilecek sekilde secilmelidir.

Kelime ornegine geri donersek, bir F soyle olabilir,

$F_{15}(x, y)$ = "eger ikinci kelimenin bas harfi buyuk ve ikinci etiket isim (noun)". Ozellik fonksiyonlari reel degerlidir. Bunun ozel durumu 0/1 degeri veren ozellik fonksiyonlaridir. Biraz onceki ornek mesela 0/1 donduruyor.

Ya da $F_{14}(x, y)$ diyelim ki soyle "ilk kelimenin bas harfi buyuk, ve ilk etiket bir isim". Tahmin edebiliriz ki egitim setimizde ilk kelimesinin bas harfi buyuk *olan* ama o kelimesi isim olmayan pek cok ornek olacaktir. Bu durumda w_{14} kucuk olur.

Dedigimiz gibi F reel degeri olabilir, mesela

$$F_{16}(x, y) = lengh(y) - lengh(x)$$

yani bu fonksiyonda x'nin uzunlugunu y'nin uzunlugundan cikartiyoruz. Bu ne ise yarar? Diyelim ki otomatik tercume yapmasi icin bir yapay ogrenim programi yaziyoruz, x, y egitim noktalari birbirinin tercumesi olan Ingilizce/Fransizca cumleler. Cogunlukla Fransizca cumleler tekabul ettikleri Ingilizce cumlelerden cok daha

uzun oluyorlar, yani ustteki cikarma cogunlukla pozitif sonuc verecek. Degisik bir acidan bakarsak, pozitif bir sonuc, bir tercumenin dogru oldugu yonunde bir isaret olarak kabul edilebilir, ve ustteki ozellik fonksiyonu uzerinden egitim algoritmasi bunu kullanir.