

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 $\langle x, y \rangle$ veri noktaları 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 dagilimlerin 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 ol-mali: “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 amaclı olarak $\langle x, y \rangle$ ikili veri noktaları saglanacak. x kaynak veri, y tahmin edilecek (ya da basta egitim hedefi olan) etiket olacak. y icin bir model ortaya cikartiyoruz, cunku test zamanında 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 $\langle x_1, y_1 \rangle, \dots, \langle x_n, y_n \rangle$ icin, θ 'yi soyle sec

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

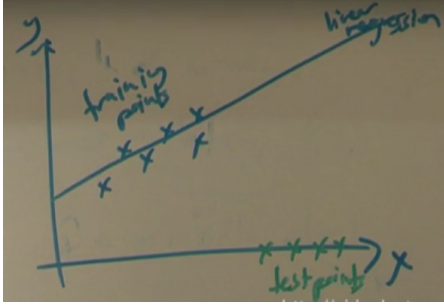
Normal maksimum olurlukta bilindigi gibi olasiliklerin carpimi maksimize edilir, burada maksimize ettigimiz “kosulsal” olasiliklerin carpimi.

Burada önemli bir soru su: bildiğimiz gibi maksimum olurluk hesabı her veri noktasının bir diğerinden bağımsız olduğunu farzeder [çünkü her olurluk hesabını bir diğer ile çarpıyoruz, başka ek çarpım, toplama, vs yapmıyoruz], bu faraziye doğru bir faraziye midir? Bu soru ve ona verilecek cevap çok önemli. Evet, eğer eğitim noktaları birbirinden bağımsız değilse maksimum olurluk kullanmamalıyız. Bağımsızlığı da iyi tanımlamak gerekiyor tabii, eğer üstteki durumda x_i verildikten sonra y_i 'lerin birbirinden bağımsız olması yeterli.

Bu model klasik İstatistik'te çokça kullanılan bir yaklaşımdır, hatta lineer regresyon'un temeli üstteki faraziyedir.

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

Bu standart lineer regresyon modeli, ve bu modelde her y ona tekabül eden x 'e bağlı, bu sayede x 'ler biliniyorsa y 'ler birbirinden koşulsal olarak bağımsız hale geliyor, böylece x 'ler birbirine bağımlı olsa bile α ve β 'nin bulunması mümkün oluyor.



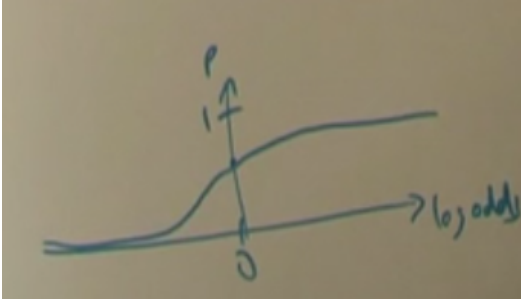
Üstteki resimde eğitim noktaları (training points) mavi olsun, test noktaları yeşil olsun (hemen altında). Bazı Yapay Öğrenim yaklaşımları diyebilir ki eğitim x 'lerinin dağılımı test x 'lerinin dağılımından farklı, bu veri seti öğrenilemez (yani genellenemez, modellenemez). Fakat klasik İstatistik buna bakar ve der ki x 'lerin verildiği durumda y 'ler bağımsızdır, bu şekilde bir koşulsal model öğrenilebilir.

Lojistik Regresyon aynı şekilde işler (lojistik regresyon, log lineer modellerin özel bir halidir, Koşulsal Rasgele Alanlar aynı şekilde). Burada da öğrenilen bir

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

modeli vardır ve y değerleri sadece 0 ve 1 olabilir. Tahmin edilen olasılık ise y 'nin 1 olma olasılığıdır. Bu model Rasgele Gradyan Çikisi ile eğitilir [detaylar için *Lojistik Regresyon* notlarımıza 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 linear 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 bulacaktır.

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) amaclil 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 sifra cok yakin, ya bire cok yakin. Literaturde NB skorlarinin “iyi kalibre edilmiş olmadığı” soyleneir.

X_1, \dots, 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 $1, \dots, n$ tane ogenin icinden $y = 1$ degerini tasiyan ogelerin sayisi olsun. Ornek, elimizde 100 tane egitim noktasi var, bunlari 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 edilmissse 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 için log olurlugun (LCL) turevini almak lazim. Once basitlestirme amaclı $\alpha = \beta_o$, ve $x_0 = 1$. O zaman log sansin eski hali (altta esitligin sol tarafı) soyle yazilabilir (sag taraf), daha derli toplu bir formül olur,

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

Bulmak istedigim her j için $\frac{d}{d\beta_j} LCL$ 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 formülün her ikisinde de $d/d\beta_j p$ kısmi olduguna dikkat.

Notasyon

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

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

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

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

$$\begin{aligned}\frac{d}{d\beta_j} p &= (-1)(1+e)^{-2} \frac{d}{d\beta_j} e \\ &= (-1)(1+e)^{-2} (e) \frac{d}{d\beta_j} (x_j) \\ &= \frac{1}{1+e} \frac{e}{1+e} x_j = p(1-p)x_j\end{aligned}$$

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

$$\begin{aligned}\frac{d}{d\beta_j} \log(1-p) &= \frac{1}{1-p} (-1)p(1-p)x_j \\ &= -px_j\end{aligned}$$

Ustteki turevler tek bir egitim veri noktası 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 anlamına geliyor. Simdi notasyonel bir numara kullanacagim,

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

Bunu niye yaptım? (4) formülünde eşitliğin sağ tarafı, birinci terim içinde 1 sayısı var, sonraki terimde 1 yok. Eğer 1 olup olmaması yerine y_i kullanırsam, ki zaten 1'in olup olmaması y_i 'nin 1 olup olmamasına bağlı, tek bir terimde işi halledebilirim. $y_i = 1$ olduğu zaman üstteki ifade $1 - p_i$ olacaktır, olmadığı zaman $-p_i$ olacaktır.

Eristigimiz sonucu analiz etmemiz gerekirse, nihai formül gayet basit ve temiz çıktı.

[24:10] kalibrasyonla alakalı bir yorum

Rasgele Gradyan Çikisi (Stochastic Gradient Ascent)

Fikir: türevi eğitim noktası başına hesapla, ve modeli hemen güncelle.

Eğitim noktaları $\langle x, y \rangle$ olarak gelsinler. Her nokta için, ve her β_j için

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

hesapla.

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

Gradyanın ne olduğunu hatırlayalım, bir fonksiyonun maksimumuna “doğru” olan bir gidis yönünü gösterir, ve bu gidis yönü o fonksiyonu oluşturan değişkenlerin (parçalı türevleri) üzerinden belirtilir. O zaman elimizdeki gradyan o iç değişkenlerin maksimum yandaki değişim şeklini bize tarif eder.

Algoritmanın tamamı: alttaki formül için

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

Her x için

- O anki modele göre p 'yi hesapla

- Her $j = 0, \dots, d$ için

- $\beta_j := \beta_j + \alpha \underbrace{(y - p)x_j}_{\text{kısmi türev}}$ hesapla

Peki metotun ismindeki “rasgele (stochastic)” tanımı nereden geliyor? İyi bir soru bu çünkü metotta rasgele sayı üretimi gibi şeyler görmüyoruz. Cevap, metot yine de rasgele, çünkü her noktayı ayrı ayrı işliyoruz, ve bu noktaların eğitim algoritmasını gelişi bir nevi “veriyi örneklemek” gibi sanki, ek olarak veriyi eğitime almadan önce rasgele şekilde karıştırmak da iyi olabilir.

Bazı Tavsiyeler (Heuristics)

1) Her özellik (feature) x_j 'i ölçeklemek, yani aynı ortalama (mean) ve varyansa sahip olacak şekilde tekrar ayarlamak. Yani mesela 0 ile 100 arasında olabilecek “yas” gibi

bir özelliği, 0 ve 1 arasında değişen özellikler ile aynı ortalama ve varyansa sahip olacak şekilde ayarlamak. Bunun sebebi güncelleme hesabındaki λ 'nin tek bir sabit olması, ve bu sabit her j için aynıdır, o sebeple λ 'nin her öğeye “aynı şekilde” uygulanabilmesi için öğelerin birbirine yakın olması iyidir. Ek olarak, genellikle eğitim verisinde 0 ile 1 arasında ikisel türden öğeler vardır, o sebeple bu şekilde olmayan diğer öğeleri 0 ve 1 arasında çekmek daha uygun ve kolay olur.

2) Veriyi rasgele şekilde sıralamak. Terminoloji: eğitim veri seti üzerinden bir geçiş yapmak bir “çag” (epoch) olarak bilinir.

3) λ 'yi deneme / yanılma yöntemi ile bulun (bu sabiti bulmanın sistemik bir yöntemi yok). Belki verinin içinden alınan daha ufak bir örneklem üzerinde bu deneme / yanılma işlemi yapılabilir.