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

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

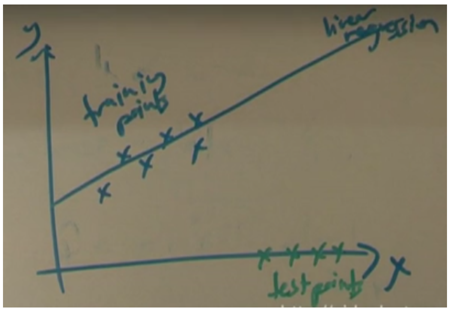
Normal maksimum olurlukta bilindigi gibi olasiliklarin carpimi maksimize edilir, burada maksimize ettigimiz "kosulsal" olasiliklarin carpimi.

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 ustteki 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 ustteki 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.



Ustteki 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 kosulsal model ogrenilebilir.