IDA ELŐADÁS I.

Bolgár Bence

2014. október 17.

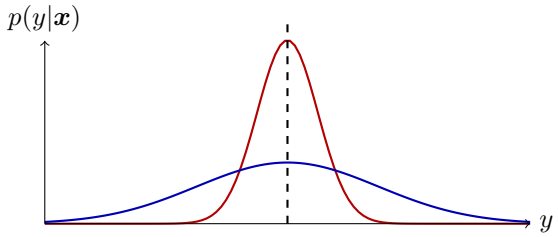
I. Generatív és diszkriminatív modellek

Korábban megismerkedtünk a felügyelt tanulással (supervised learning). Legyen adott a $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^P$ tanító halmaz, ahol rendszerint $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, d az adatok dimenziója, $y \in \mathcal{Y}$-ra pedig pl. $\mathcal{Y} = \{+1, -1\}$ (klasszifikáció) vagy $\mathcal{Y} = \mathbb{R}$ (regresszió). Célunk, hogy bármely új $\boldsymbol{x}$ mintára y-t prediktáljuk.

I.A. Generatív modellek

Az ún. *generatív* modellek esetében $p(\boldsymbol{x}, y)$-t keressük (innen a „generatív" kifejezés is: a $p(\boldsymbol{x}, y)$ együttes eloszlás ismeretében akár új mintákat is tudunk generálni). Innen $p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x},y)}{p(\boldsymbol{x})}$, azaz adott $\boldsymbol{x}$ esetén a y jósolható. Miért jobb nekünk $p(y|\boldsymbol{x})$ ismerete, mint egy egyszerű pontbecslés?

- **Konfidencia-értékek származtatása.** Bár adott $\boldsymbol{x}$-re mindkét esetben ugyanazt az értéket jósoljuk, mégis, az egyik esetben biztosabbak lehetünk az eredmény helyességében:



- **Minták eldobása.** Bizonytalanság esetén adott esetben egy-egy minta akár el is dobható.
- **Kiegyensúlyozatlan osztályozás kompenzációja.** Tegyük fel, hogy a feladat egy ritka betegség diagnosztizálása; jelölje C_b a beteg osztályt:

$$p(C_b|\boldsymbol{x}) \propto p(\boldsymbol{x}|C_b)p(C_b).$$

  Mivel az egészséges emberek sokkal nagyobb arányban fordulnak elő, a modellünk majdnem tökéletes eredményt fog elérni akkor is, ha minden pácienst egészségesnek nyilvánít. Ennek kiküszöbölésére tanítsuk a modellt egy mesterségesen kiegyensúlyozott adathalmazon. A fenti képlet alapján a kapott posterior arányos a priorral, így nincs más dolgunk, mint a kapott poszteriort leosztani a „mesterséges" priorral (azaz a kiegyensúlyozás után az adott osztályba eső minták arányával), majd visszaszorozni az eredeti populációra jellemző priorral.
- **Modellek kombinációja.** Éljünk az alábbi naiv feltételezéssel:

$$p(\boldsymbol{x}_A, \boldsymbol{x}_B|C_b) = p(\boldsymbol{x}_A|C_b)p(\boldsymbol{x}_B|C_b),$$

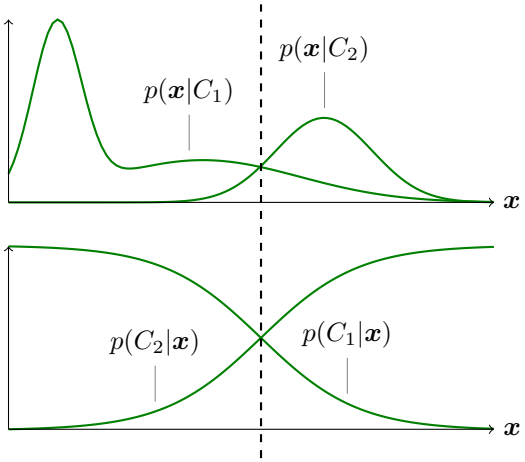
  azaz pl. a betegség diagnosztizálására két teszt is rendelkezésre áll, amelyek eredménye feltételesen (!) független egymástól. Ekkor

$$p(C_b|\boldsymbol{x}_A, \boldsymbol{x}_B) \propto p(\boldsymbol{x}_A, \boldsymbol{x}_B|C_b)p(C_b) \propto p(\boldsymbol{x}_A|C_b)p(\boldsymbol{x}_B|C_b)p(C_b) \propto \frac{p(C_b|\boldsymbol{x}_A)p(C_b|\boldsymbol{x}_B)}{p(C_b)}$$

A generatív modellek további előnyei közé tartozik a marginálisok ismerete. A $p(\boldsymbol{x})$ eloszlás például felhasználható az ún. „outlierek" (kiugró, rendellenes minták) detekciójára. A generatív modellek hátránya, hogy rendszerint sok mintát és nagy számítási teljesítményt igényelnek (mintakomplexitás, számítási komplexitás).

I.B. Diszkriminatív modellek

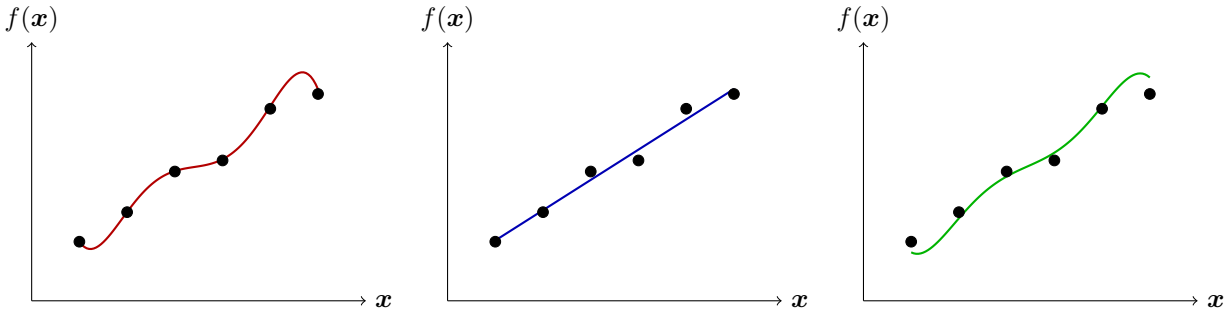
A diszkriminatív modelleknél közvetlenül $p(y|\boldsymbol{x})$-et becsüljük. Ekkor az együttes eloszlást, illetve $p(\boldsymbol{x})$-et elveszítjük, ám a fenti előnyök nagy része továbbra is megmarad. Az alábbi ábra ismét egy osztályozási feladatot mutat:



Látható, hogy $p(\boldsymbol{x}|C_1)$ baloldali módusza egyáltalán nem befolyásolja a poszteriort – az együttes eloszlás ismerete tehát nincs kihatással a predikcióra.

I.C. Diszkriminatív függvények

Itt olyan $f : \mathcal{X} \to \mathcal{Y}$, $f \in \mathcal{F}$ függvényt keresünk, amelyre $f(\boldsymbol{x}) = y$, azaz a $p(y|\boldsymbol{x})$ eloszlást is elveszítjük (annak minden előnyével együtt), y-ra csak pontbecslést kapunk. Előny viszont, hogy általában hatékonyan számítható és jó prediktív teljesítménnyel bíró eljárásokhoz jutunk (pl. SVM). Felmerül a kérdés, hogy hogyan válasszuk meg f-et?



Az ábra bal oldalán a hat adatpontunkra egy ötödfokú polinomot illesztettünk. Ez a függvény hat szabad paraméterrel bír, így $r = 0$ hibával ráilleszthető a tanítómintákra. Megmutatható, hogy ennek ellenére a modell általánosítóképessége rossz – f gyakorlatilag nem tett mást, mint „megjegyezte” a tanítómintákra adandó válaszokat. Lényegesen jobb általánosítóképesség érhető el, ha $\mathcal{F}$-et megszorítjuk, pl. a legfeljebb n-edfokú polinomokra (az ábra közepén $n = 1$, azaz lineáris regressziót végeztünk) – így a szabad paraméterek számát, más szóval a modell komplexitását csökkentjük (itt gondolhatunk például az „Occam borotvája” elvre).

A modell komplexitásának csökkentésére másik megközelítés az f függvény regularizációja. Az ábra jobb oldalán szintén ötödfokú polinomot illesztettünk, ám korlátoztuk az együtthatók nagyságát. A regularizált rizikóminimalizálás (RRM) során a hiba minimalizálása mellett a függvény komplexitásának minimalizálására törekszünk, amelyet például mérhetünk a függény valamilyen $\|f\|$ normájával. A következő szakasz egy olyan keretet tárgyal, amely magába foglalja a regularizáció kérdését, a korábban már megismert kernel módszereket, valamint elméleti garanciákat is szolgáltat.

II. Reproducing Kernel Hilbert Space (RKHS)

Tekintsük a $\mathcal{H} := \{f \mid f = \sum_i \alpha_i k(\cdot, \boldsymbol{x}_i)\}$ teret, ahol $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ egy kernel függvény (azaz szimmetrikus és pozitív definit). Ekkor a $k(\cdot, \boldsymbol{x}_i) : \mathcal{X} \to \mathbb{R}$ függvények bázist alkotnak a $\mathcal{H}$ térben. Tudjuk, f regularizációjához szükségünk lesz egy normára, valamint a legtöbb kernel gép igényel egy belső szorzatot. Definiáljunk tehát egy belső szorzatot a fenti téren a következőképpen:

1. Definíció. *Legyen $f, g \in \mathcal{H}$, $f = \sum_i \alpha_i k(\cdot, \boldsymbol{x}_i)$, $g = \sum_j \beta_j k(\cdot, \boldsymbol{x}_j)$. Ekkor*

$$\langle f, g \rangle := \sum_i \sum_j \alpha_i \beta_j k(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

1. Következmény. *A fenti választással*

$$\langle f, g \rangle = \sum_i \sum_j \alpha_i \beta_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_i \alpha_i g(\boldsymbol{x}_i) = \sum_j \beta_j f(\boldsymbol{x}_j).$$

2. Következmény. *Szintén azonnal látható, hogy*

$$\langle f, k(\cdot, \boldsymbol{x}) \rangle = \sum_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) = f(\boldsymbol{x}).$$

Ez a „reprodukáló” tulajdonság (reproducing property).

Ahhoz, hogy $\mathcal{H}$ vektortér legyen (és így az algoritmusok működjenek), be kell látnunk, hogy a fent definiált belső szorzat eleget tesz a követelményeknek.

1. Állítás. *$\langle \cdot, \cdot \rangle$ valóban belső szorzat a $\mathcal{H}$ téren.*

Bizonyítás. Az alábbi tulajdonságokat kell belátni:

- Szimmetrikus: k szimmetriájának közvetlen következménye.
- Bilineáris: pl. $\langle f + f', g \rangle = \sum_j \beta_j (f + f')(\boldsymbol{x}_j) = \sum_j \beta_j f(\boldsymbol{x}_j) + \sum_j \beta_j f'(\boldsymbol{x}_j) = \langle f, g \rangle + \langle f', g \rangle$. A $\langle \lambda f, g \rangle = \lambda \langle f, g \rangle$ eset hasonlóképpen látható; a bilinearitás a szimmetria felhasználásával következik.
- $\langle f, f \rangle \geq 0$. Tudjuk, hogy $\langle f, f \rangle = \sum_i \sum_j \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\alpha}^T K \boldsymbol{\alpha}$, ahonnan az állítás következik K pozitív definit volta miatt.
- $\langle f, f \rangle = 0 \Leftrightarrow f = 0$. Az egyik irány következik az $|f(\boldsymbol{x})|^2 = |\langle f, k(\cdot, \boldsymbol{x}) \rangle|^2 \leq k(\boldsymbol{x}, \boldsymbol{x}) \langle f, f \rangle$ egyenlőtlenségből, ahol a Cauchy-Schwarz-Bunyakovszkij egyenlőtlenséget használtuk. A másik irány triviális.

□

Emlékezzünk vissza, hogy számos algoritmus „nemlinearizálásának” alapötlete az volt, hogy az $\boldsymbol{x}_i$ mintákat egy másik (gyakran magasabb dimenziójú) térbe képeztük, majd ebben futtattuk az eredeti (lineáris) algoritmusunkat. A *kernel trükk* alkalmazása során pedig a leképezés explicit megadása helyett mintegy „lecseréltük” az euklideszi belső szorzatot a k kernel függvényre:

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \rightsquigarrow \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle = k(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

A reprodukáló tulajdonságból immár azt is tudjuk, hogyan írható fel ez a leképezés, ugyanis

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle k(\cdot, \boldsymbol{x}_i), k(\cdot, \boldsymbol{x}_j) \rangle ,$$

így látjuk, hogy a ϕ leképezés nem más, mint $\phi : \boldsymbol{x} \mapsto k(\cdot, \boldsymbol{x})$. Hogyan néz ki ϕ a gyakorlatban?

1. Példa. *Legyen a $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel függvény a következő:*

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle^2 .$$

Az egyszerűség kedvéért kétdimenziós esetet tekintünk, azaz legyen $\boldsymbol{x}_i = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$, $\boldsymbol{x}_j = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$. Ekkor

$$\begin{aligned}
\left\langle \phi\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}\right), \phi\left(\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\right) \right\rangle &= k\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}\right) = \left\langle \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \right\rangle^2 \\
&= (a_1 b_1 + a_2 b_2)^2 \\
&= a_1 b_1 a_1 b_1 + a_1 b_1 a_2 b_2 + a_2 b_2 a_1 b_1 + a_2 b_2 a_2 b_2 \\
&= a_1 a_1 b_1 b_1 + a_1 a_2 b_1 b_2 + a_2 a_1 b_2 b_1 + a_2 a_2 b_2 b_2 \\
&= \left\langle \begin{bmatrix} a_1 a_1 \\ a_1 a_2 \\ a_2 a_1 \\ a_2 a_2 \end{bmatrix}, \begin{bmatrix} b_1 b_1 \\ b_1 b_2 \\ b_2 b_1 \\ b_2 b_2 \end{bmatrix} \right\rangle .
\end{aligned}$$

Következik, hogy $\phi\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}\right) = \begin{bmatrix} a_1 a_1 \\ a_1 a_2 \\ a_2 a_1 \\ a_2 a_2 \end{bmatrix}$, azaz k rögzítésével megkaptuk a ϕ leképezést is. A fenti k a polinomiális kernelek speciális esete:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + a)^d ,$$

ahol d a kernel foka, a pedig a homogén/inhomogén tulajdonságért felel. Látjuk, hogy a másodfokú homogén polinomiális kernel ($d = 2$, $a = 0$) a feature-párok terébe képez – ez hasznos pl. képfeldolgozásnál, ahol az éldetekció történhet pixel-párok alapján. Magasabb fokú kernelek a feature-n-esek terébe képeznek, $a \neq 0$ esetén pedig inhomogén kereszt-tagok is megjelennek a reprezentációban (azaz feature 1-esek + feature 2-esek + ... + feature-n-esek).

A fenti példa jól mutatja a *kernel trükk* lényegét: míg nagy d esetén a $\phi(\boldsymbol{x})$ reprezentációk belső szorzata a nagy dimenzió miatt közvetlenül már nem kiszámítható, a k függvény segítségével mégiscsak gyorsan megkaphatjuk. Térjünk most vissza f regularizációjára és az $\|f\|$ normára.

1. Tétel. *(Kimeldorf–Wahba reprezenter tétel). Legyen adott*

- *$\Omega : \mathbb{R}^+ \to \mathbb{R}$ szigorúan monoton növő függvény,*
- *$L : (\mathcal{X} \times \mathbb{R} \times \mathbb{R})^P \to \mathbb{R}$ általános veszteségfüggvény.*

Ekkor a

$$L\left((\boldsymbol{x}_1, y_1, f(\boldsymbol{x}_1)), \ldots, ((\boldsymbol{x}_P, y_P, f(\boldsymbol{x}_P))\right) + \Omega\left(\|f\|\right)$$

regularizált rizikó minden minimalizátora a következő alakban írható:

$$f(\boldsymbol{x}) = \sum_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}).$$

2. Példa. *Kétosztályos SVM.*

- $L\left((\boldsymbol{x}_1, y_1, f(\boldsymbol{x}_1)), \ldots, ((\boldsymbol{x}_P, y_P, f(\boldsymbol{x}_P))\right) = \frac{1}{P} \sum_i \max\left(0, 1 - y_i f\left(\boldsymbol{x}_i\right)\right)$
- $\Omega\left(\|f\|\right) = \frac{\lambda}{2} \|f\|^2$

Ez a felírás ekvivalens a következővel:

$$\begin{aligned}
\min \quad & \frac{1}{2} \|f\|^2 + C \sum_i \xi_i \\
s.t. \quad & y_i f(\boldsymbol{x}_i) \geq 1 - \xi_i, \quad \xi_i \geq 0,
\end{aligned}$$

ami a kétosztályos SVM primálja, ha figyelembe vesszük, hogy $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ (ezt nem bizonyítjuk; a Riesz reprezentációs tétel következménye). A Kimeldorf–Wahba tétel garantálja, hogy a megoldás is a kívánt alakot veszi fel. Figyeljük meg, hogy mivel $f = \sum_i \alpha_i k(\cdot, \boldsymbol{x}_i)$, $\|f\|^2$ minimalizálása ekvivalens az α_i együtthatók korlátozásával (nagyon hasonlóan az előző szakaszban látott polinom-illesztéshez). Korábban azt is láttuk, hogy a kétosztályos SVM duál feladata a következő:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_i \alpha_i \\ s.t. \quad & 0 \leq \alpha_i \leq C, \end{aligned}$$

ahol a kényszerfeltétel szintén az α_i együtthatók korlátozását jelenti. Magas C esetén a korlát magasabb, ami gyengébb regularizációt jelent – komplexebb modelleket és esetlegesen túlilleszkedést eredményezve.