Lineární regrese – opakování

- Cíl: aproximovat funkci $f(x)$ pomocí lineární funkce

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{n} x_j \hat{\beta}_j$$

- minimalizujeme součet čtverců reziduí (RSS – residual sum squares), tj.

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 = (y - X\beta)^T (y - X\beta)$$

- Není–li $X^T X$ singulární, dostaneme jednoznačné řešení

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- a odhad $\hat{y}$ pro dané x_i je $\hat{y}(x_i) = x_i^T \hat{\beta}$.

IBL=NN – k nejbližších sousedů

- uchovám data
- pro predikci vyberu k nejbližších, predikuji jejich průměr (resp. nejčastější hodnotu při klasifikaci)

IB3 – vylepšené IBL

- instanci zahodit, pokud
  $u_{inst} = \hat{p}_{inst} + 1,15 \cdot \sqrt{\frac{\hat{p}_{inst}(1-\hat{p}_{inst})}{\hat{N}_{inst}}} < l_{apri} = \hat{p}_j - 1,15 \cdot \sqrt{\frac{\hat{p}_j(1-\hat{p}_j)}{\hat{N}}}$
- predikovat vážený průměr k nejbližších splňujících
  $u_{apri} = \hat{p}_j + 1,645 \cdot \sqrt{\frac{\hat{p}_j(1-\hat{p}_j)}{\hat{N}}} < l_{inst} = \hat{p}_{inst} - 1,645 \cdot \sqrt{\frac{\hat{p}_{inst}(1-\hat{p}_{inst})}{\hat{N}_{inst}}}$
- ostatní nechat cvičně predikovat a počítat $\hat{p}_{inst} = \frac{s_{inst}}{N_{inst}}$.

Dva scénáře

1. Trénovací data každé třídy jsou generována dvourozměrným gaussovským rozložením s nekorelovanými komponentami a různými středními hodnotami
2. Trénovací data každé třídy pocházejí ze směsi deseti gaussovských distribucí s malým rozptylem; střední hodnoty těchto distribucí jsou opět gaussovsky rozloženy.

V prvním případě je lineární model téměř optimální a potřebuje daleko méně dat.

Ve druhém je lineární model dost problematický, NN při dost datech daleko lepší.

Statistická teorie rozhodování

Začneme spojitou cílovou veličinou Y, tj. regresí.

- Mějme vektor $X \in \Re^n$ vstupních veličin, $Y \in \Re$ výstupní veličinu, pravděpodobnostní rozložení veličin je $P(X, Y)$.
- Hledáme funkci $f(X)$ predikující Y pro daný vstup X.
- Potřebujeme zvolit **chybovou funkci (loss function)** $L(Y, f(X))$, která penalizuje chyby v predikci
- zdaleka nejčastější je **kvadratická chybová funkce**: $L(Y, f(x)) = (Y - f(X))^2$
- Snažíme se minimalizovat **očekávanou chybu predikce**

$$EPE(f) = E(Y - f(X))^2 = \int (y - f(x))^2 P(x, y) dx dy$$

- Sdruženou pravděpodobnost/chybu podmíníme X–em:

$$EPE(f) = E(Y - f(X))^2 = E_X E_{Y|X}([Y - f(X)]^2|X)$$

- tedy stačí minimalizovat EPE bodobě:

$$f(x) = arcmin_c E_{Y|X}([Y - c]^2|X = x)$$

- což řeší podmíněné očekávání, tzv. **regresní funkce**

$$f(x) = E(Y|X = x)$$

- tj. nejlepší predikce Y v libovolném bodě $X = x$ je podmíněný **průměr**, *při kvadratické chybové funkci*.

Nejbližší sousedi (IB3)

- většinou v daném bodě $X = x$ máme nejvýše jedno pozorování
- proto průměr bereme přes určité okolí (k nejbližších sousedů)
- a očekávaný průměr odhadneme průměrem na (nejbližších) trénovacích datech.
- za mírných podmínek na $P(X, Y)$ to pro $N, k \to \infty$ tž. $\frac{k}{N} \to 0$ odhad konverguje k hledané $f(x)$.
- Proč hledat dál? ... Na rychlosti konvergence také záleží!
  S rostoucí dimenzí X se drasticky snižuje (při stejně velkých datech).

Lineární regrese

- Předpokládá lineární regresní funkci $f(x) \approx x^T \beta$
- dosazením do $EPE(f) = E(Y - f(X))^2$ a derivací dostaneme:

$$\beta = [E(XX^T)]^{-1} E(XY)$$

- tj. *nepodmiňuji* X–em, pro libovolnou predikci používám všechna trénovací data
- vhodnější pro "málo" datech, "málo" záleží na dimenzi n.
- Pokud linearita f neplatí, mohu být hodně mimo.
- Mnohé modely "zobecňují" lineární regresi, např. **aditivní modely** předpokládají $f(X) = \sum_{j=1}^{n} f_j(X_j)$.

Proč *kvadratická* chyba?

- Když vezmeme chybu: $L_1 = E|Y - f(X)|$
- řešení je **medián**: $\hat{f}(x) = median(Y|X = x)$
- medián je robustnější než průměr
- jenže L_1 nemá spojité derivace, takže se s ní špatně počítá.
- Zdaleka nejpoužívanější je proto **kvadratická chybová funkce**, pěkně se derivuje a hledá minimum.

Klasifikace

- Pro **kategoriální** cílovou veličinu G paradigma zůstává, jen potřebujeme jinou **chybovou funkci**.
- Pro K tříd G je chybová funkce reprezentovaná maticí L rozměru $K \times K$, s nulami na diagonále a nezáporná všude, $L(k, \ell)$ je cena za klasifikaci pozorování g_k jako g_ℓ.
- Očekávaná chyba predikce a rozpis po podmínění:

$$EPE = E[L(G, \hat{G}(X))] = E_X \sum_{k=1}^{K} L[g_k, \hat{G}(X)] P(g_k|X).$$

- Opět řešíme bodově:

$$\hat{G}(x) = argmin_{g \in G} \sum_{k=1}^{K} L(g_k, g) P(g_k|X = x).$$

Bayesovský klasifikátor

- pro 0–1 chybu, tj. všude kromě diagonály v L jedničky, dostaneme:

$$\hat{G}(x) = argmin_{g \in G}[1 - P(g|X = x)]$$

- tj. pro dané x predikujeme nejčetnější třídu, $\hat{G}(X) = g_k$ pro které $P(g_k|X = x) = max_{g \in G} P(g|X = x)$.
- Toto řešení se nazývá **bayesovský klasifikátor**, z něj se dá odvodit bayesovsky optimální rozhodovací hranice a bayesovsky optimální (bayesovská) chyba – za daného zadání nelze očekávanou chybu snížit pod bayesovsky optimální.

Nejbližší sousedi, regrese

- k–nejbližších sousedů můžeme použít pro klasifikaci; opět bod aproximujeme okolím a nejčastěji očekávanou třídu nejčastější třídou v datech
- Pro $K = 2$ dvouhodnotovou klasifikaci můžeme třídy kódovat reálnými čísly 0,1 a tuto proměnnou Y predikovat regresí $\hat{f}(X)$. Pak:

$$\hat{f}(X) = E(Y|X) = P(G = g_1|X)$$

  kde g_1 odpovídá $Y = 1$.
- Vrátíme se k tématu později (SVM, rozhodovací stromy, možná bude i logistická regrese aj.).

Prokletí dimenzionality

- Budeme precizovat **prokletí dimenzionality** = pro velký počet vstupních proměnných mohu mít dat kolik chci, a stejně jich je málo.
- Předpokládejme dimenzi p, data rovnoměrně rozložená v hyperkrychli. Chceme zachytit r–tinu pozorování $(0 < r < 1)$ nejblíže danému bodu, tj. r –tinu objemu hyperkrychle. Objem krychle je 1, proto očekávaná délka hrany "okolí" bude $e_p(r) = r^{\frac{1}{p}}$.
- $e_{10}(0.01) = 0.63$, $e_{50}(0.0001) = 0.83$, $e_{15}(\frac{1}{10000000}) = 0.34$
- pro pokrytí desetitisíciny prostoru v 50ti dimenzích musím z každé proměnné vzít zhruba 83 procent hodnot, tedy zdaleka ne jen blízké hodnoty.

Blízkost kraje

- Mějme N příkladů v p dimenzionální jednotkové kouli.
- Medián vzdálenosti nejbližšího soused ke středu je

$$d(p, N) = \left(1 - \frac{1}{2}^{\frac{1}{N}}\right)^{\frac{1}{p}}$$

- pro $N = 500$, $p = 10$, $d(p, N) \approx 0,52$.

Málo dat

V jednom rozměru je $N_1 = 100$ rozumné množství příkladů.
V deseti dimenzích stejné hustotě odpovídá $N_{10} = 100^{10}$ vzorků.

Příklad – mnoharozměrná boule

- Vyrobíme umělá data, 1000 příkladů rovnoměrně rozložených na $[-1;1]^p$. Y je dáno přesně, bez šumu, vztahem:

$$Y = f(x) = e^{-8||X||^2}$$

- Pomocí nejbližšího souseda predikujeme hodnotu y_0 v bodě $x_0 = 0$.
- V málo dimenzích vpohodě,
- pro $p = 10$ dimenzí pro 99% příkladů je nejbližší soused dále než $0,5$ od počátku, tj. odhad jde k nule.

Rozklad chyby na vychýlení a rozptyl

Bias variance decomposition

- Chybu v předchozím příkladu můžeme rozepsat:

$$\begin{aligned} MSE(x_0) &= E_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 \\ &= E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0)]^2 + [E_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2 \\ &= Var_{\mathcal{T}}(\hat{y}_0) + Bias^2(\hat{y}_0) \end{aligned}$$

- ve více dimenzích vychýlení (bias) roste, rozptyl zůstává (v tomto příkladu dokonce klesá)
- vychýlení (bias) roste u složitějších funkcí často - zde jde o interakci všech proměnných.
- Pokud funkce závisí jen na jedné (málo) dimenzi, může být rozptyl horší než vychýlení.

Složitost modelu a vychýlení x rozptyl

- Nejbližší sousedi mají parametr k.
- Velké $k \rightarrow$ "hladší funkce", menší efektivní počet parametrů (u NN je $\frac{N}{k}$)
- Velké $k \rightarrow$ větší vychýlení, menší rozptyl, neboť:
  pro $Y = f(X) + \epsilon$, $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$

$$\begin{aligned} EPE_k(x_0) &= E[(Y - \hat{f}_k(x_0))^2 | X = x_0] \\ &= \sigma^2 + [Bias^2(\hat{f}_k(x_0)) + Var_{\mathcal{T}}(\hat{f}_k(x_0))] \\ &= \sigma^2 + [f(x_0) - \frac{1}{k}\sum_{\ell=1}^{k} f(x_{(\ell)})]^2 + \frac{\sigma^2}{k} \end{aligned}$$

kde ℓ probíhá nejbližší sousedy.

Lineární regrese, IBL nebo něco jiného?

obrázek, 500 příkladů

- Pokud chceme zobecňovat z viděných dat na nová, vždy musíme přijmout nějaké předpoklady o procesu, jak jsou data generována.
  - IBL předpokládá lokálně konstantní funkci
  - lineární model předpokládá globálně lineární funkci
  - bez předpokladů nejsme schopni říct vůbec nic o x, které se nevyskytlo mezi příklady.
- pokud máme velkou znalost procesu (či přijímáme silné předpoklady o procesu, učíme málo parametrů), pak stačí "málo" dat
- pokud má model hodně efektivních parametrů, potřebujeme dat nesrovnatelně více

Lineární model má $n + 1$ parametrů, efektivní počet parametrů k–NN je N/k, což je obecně větší než n (tvoříme N/k prostůrků a pro každý učíme

střed).

Statistické modely

- U regrese většinou předpokládáme deterministickou funkci f a nezávislý identicky rozložený šum ϵ, $E(\epsilon) = 0$, tj.

  $$f(x) = E(Y|X = x)$$

  $P(Y|X)$ závisí na X jen přes podmíněný průměr $f(x)$.

- Mnohá klasifikace se dá představit jako obarvená mapa, tj. deterministická funkce s šumem.

- Pokud predikujeme pravděpodobnost cílových tříd (při klasifikaci), naše cílová funkce $p(x)$ je přímo podmíněná hustota $p(x) = P(G|X)$.
  Při 0–1 kódování je $E(Y|X = x) = p(x)$, ale rozptyl není konstantní: $Var(Y|X = x) = p(x)[1 - p(x)]$.

Aproximace "doladěním" parametrů θ

- Často na aproximaci můžeme nahlížet jako na doladění parametrů θ u předem zvoleného modelu.
- Např. u lineární regrese model $f(x) = x^T\beta$ má $\theta = \beta$.
- Lineární expanze báze $f_\theta(x) = \sum_{k=1}^{K} h_k(x)\theta_k$
  kde h_k jsou zvolené funkce x, např. $x_1^2, x_1x_2^2, cos(x_1)$ apod.
- Parametry doladíme minimalizací RSS
  $RSS(\theta) = \sum_{i=1}^{N}(y_i - f_\theta(x_i))^2$.
- Neuronové sítě používají nelineární aproximaci
  $h_k(x) = \frac{1}{1+exp(-(x^T\beta_k+\beta_0))}$ a učí gradientně.

Maximálně věrohodný odhad
Maximum likelihood estimation

- Hledáme model, ve kterém je (log)pravděpodobnost pozorovaných dat maximální, tj. maximalizujeme $L(\theta) = logP_\theta(y_i)$.
- Modelu s aditivní chybou $Y = f_\theta(X) + \epsilon$, $\epsilon \approx N(0, \sigma^2)$ odpovídá maximalizaci věrohodnosti $P(Y|X, \theta) = N(f_\theta(X), \sigma^2)$, tj.

$$L(\theta) = -\frac{N}{2}log(2\pi) - Nlog\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f_\theta(x_i))^2)$$

  kde jen poslední výraz obsahuje θ, čili maximalizujeme zápornou konstantou vynásobenou kvadratickou chybu.
- Pro klasifikaci s modelem $p_{k,\theta}(x) = P(G = \mathcal{G}_k|X = x)$, $k = 1, \ldots, K$ maximalizujeme $L(\theta) = \sum_{i=1}^{N} \log p_{g_i,\theta}(x_i)$.

Strukturované regresní modely

- Regrese – predikujeme spojitou veličinu
- zpravidla máme spojitý vstup (v některých dimenzích), tedy nekonečně hodnot, tedy jednu nebo žádnou hodnotu pro většinu vstupů
- minimalizaci RSS splňuje nekonečně mnoho funkcí interpolujících naměřené hodnoty
- to ale nebývá vhodné pro predikci (velká očekávaná a testovací chyba).
- Omezíme přípustné funkce, ve zvolené třídě jednoznačné řešení.
- Chceme ”jednoduché” funkce, které nejsou divoké na malých okolích ve vstupním prostoru.
- Velikost okolí bývá parametrem, čím větší, tím větší restrikce.

Penalizace za složitost

- Míru chyby RSS přímo upravíme penaltou $J(f)$ za složitost modelu

$$PRSS(f;\lambda) = RSS(f) + \lambda J(f)$$

- např. Hřebenová (ridge) regrese penalizuje nenulové složky β

$$\hat{\beta}^{ridge} = argmin_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

- kubický vyhlazující splajn pro jednorozměrný vstup

$$PRSS(f;\lambda) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx.$$

- $\lambda \geq 0$ řídí velikost penalty, $\lambda = 0$ vede k interpolaci, $\lambda = \infty$ dovolí jen lineární funkce x.

Jádrové funkce a lokální regrese

- Nejbližší sousedi tvoří "skokovou" aproximující funkci, není to hezké a je to zbytečné
- specifikujme okolí jádrovou funkcí $K_\lambda(x_0, x)$ určující, nakolik je x v okolí x_0, např. gausovské jádro

$$K_\lambda(x_0, x) = \frac{1}{\lambda} exp\left[-\frac{||x - x_0||^2}{2\lambda}\right]$$

- y pak odhadneme

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}$$

- nebo ”naučíme” parametrickou funkci f_θ minimalizací RSS

$$RSS(f_\theta, x_0) = \sum_{i=1}^{N} K_\lambda(x_0, x_i)(y_i - f_\theta(x_i))^2$$

- $f_\theta = \theta_0$ vede na odhad viz výše (Nadaraya–Watson)
- $f_\theta = \theta_0 + \theta_1 x$ dává lokálně lineární regresní model.
- Nejbližší sousedi mají jádrovou funkci závislou na datech, je 1 ve vzdálenosti menší či rovné vzdálensti $k$tého souseda, jinak nula.

Báze funkcí a "slovníkové" metody

- vezmeme bázi funkcí $\{h_m(x)\}$ a modelujeme f jako jejich lineární kombinaci, θ značí parametry:

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m h_m(x)$$

- vejde se sem lineární a polynomiální expanze
- splajny a součiny tenzorů
- báze radiálních funkcí
- jednovrstvé dopředné neuronové sítě
- a další.