11. Neparametarske metode

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v1.4

1 Zadatci za učenje

1. [***Svrha: Razumjeti sličnosti i različitosti algoritma k-NN i SVM.***] Napišite dualni model algoritma SVM te model algoritma k-NN. Što ova dva modela imaju zajedničko? Po čemu se algoritmi razlikuju?

2. [***Svrha: Isprobati klasifikator k-NN na konkretnom primjeru. Razumjeti kako hiperparametar k i broj primjera N utječu na složenost modela.***]

   (a) Klasifikator 4-NN s euklidskom udaljenošću učen je na sljedećim primjerima iz $\mathbb{R}^3 \times \{0,1\}$:

   $$\mathcal{D} = \left\{\left((\mathbf{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{6} = \left\{\big((4,4,0),1\big),\ \big((4,3,1),1\big),\ \big((6,0,2),1\big),\ \big((5,2,2),0\big),\ \big((5,1,1),0\big),\ \big((7,2,0),0\big)\right\}.$$

   Odredite klasifikaciju primjera $\mathbf{x}^{(1)} = (4,2,1)$ i $\mathbf{x}^{(2)} = (0,3,3)$.

   (b) Ponovite klasifikaciju s težinskim modelom 4-NN, primjenom inverzne kvadratne jezgre.

   (c) Skicirajte (za općenit slučaj) pogrešku učenja i ispitnu pogrešku kao funkcije od k.

   (d) Skicirajte (za općenit slučaj) pogrešku učenja i ispitnu pogrešku kao funkcije broja primjera N za $k = 1$ i $k = 3$ (nacrtajte dva zasebna grafikona).

3. [***Svrha: Shvatiti uzročne veze između naoko nevezanih veličina.***] Obrazložite u kakvim su odnosima sljedeći pojmovi: (a) složenost modela, (b) broj parametara modela, (c) dimenzija ulaznog prostora n i (d) broj primjera N. Analizirajte odnose između svih parova pojmova, posebno za parametarske, a posebno za neparametarske metode.

2 Zadatci s ispita

1. (T) Algoritam SVM može biti parametarski i neparametarski, ovisno o tome provodimo li optimizaciju u primarnoj ili dualnoj formulaciji. U oba slučaja preferiramo da je model rijedak, tj. da je nakon treniranja što više parametara postavljeno na nulu. **Kako rijetkost modela ovisi o hiperparametru C?**

   A Što je C veći, to je neparametarski model manje rijedak, dok je parametarski to rjeđi jer λ pada

   B Što je C veći, to je neparametarski model manje rijedak, dok parametarski model nije rijedak jer ima L_2-regularizaciju a ne L_1-regularizaciju

   C Što je C manji, to je neparametarski model rjeđi, ali to nema utjecaja na rijetkost parametarskog modela jer on nema potporne vektore

   D Što je C manji, to je neparametarski model rjeđi, a također je to rjeđi i parametarski model jer λ raste

2. (N) Bavimo se zadatkom određivanja etimologije riječi. Zanima nas je li neka nama nepoznata riječ latinskog ili slavenskog porijeka. Zadatak rješavamo kao binarnu klasifikaciju. Prikupili smo označeni skup primjera, koji se sastoji od latinskih riječi i riječi iz svih dvanaest živućih slavenskih

jezika. Npr., u našem skupu imamo (*stroj*, 1), (*strues*, 0), (*tracto*, 0) i (*trasa*, 1), gdje 1 označava da je to slavenska riječ, a 0 da je latinska. Na ovom skupu primjera treniramo algoritam k-NN (k najbližih susjeda). Kao funkciju udaljenosti koristimo Levenshteinovu udaljenost. Levenshteinova udaljenost L između dviju riječi najmanji je broj umetanja, brisanja i zamjena jednog znaka potrebnih da se jedna riječi pretvori u drugu. Npr., $L(\textit{stroj}, \textit{straja}) = 2$. Razmatramo dva modela. Model h_1 je 3-NN. Model h_2 je težinski k-NN s jezgrenom funkcijom definiranom kao $\kappa(\mathbf{x}, \mathbf{x}') = 1/(1+L(\mathbf{x}, \mathbf{x}'))$. **Koja je klasifikacija riječi $\mathbf{x} =$ *straja* prema modelima h_1 i h_2?**

A $h_1 = h_2 = 0$ B $h_1 = h_2 = 1$ C $h_1 = 1$, $h_2 = 0$ D $h_1 = 0$, $h_2 = 1$

3. (N) Algoritam k-NN koristimo za višeklasnu klasifikaciju riječi prema jeziku kojemu pripadaju. Skup za učenje sastoji se od sljedećih riječi i oznaka klasa:

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\} = \{(\text{“water”}, 0), (\text{“voda”}, 1), (\text{“zrak”}, 1), (\text{“luft”}, 2), (\text{“feuer”}, 2)\}$$

Kao mjeru sličnosti između primjera koristimo jezgrenu funkciju nad znakovnim nizovima, definiranu kao $\kappa(\mathbf{x}_1, \mathbf{x}_2) = |\mathbf{x}_1 \cap \mathbf{x}_2|/|\mathbf{x}_1 \cup \mathbf{x}_2|$, gdje je su operacije unije i presjeka definirane nad skupovima slova od kojih se riječi sastoje. Npr., $\kappa(\text{“water”}, \text{“voda”}) = 1/8 = 0.125$. Razmatramo dvije varijante algoritma: 3-NN i težinski k-NN. Kod potonjeg u obzir uzimamo sve primjere, tj. $k = N$. Odredite klasifikaciju primjera $\mathbf{x} =$ “zemlja” pomoću ova dva algoritma. U slučaju jednake sličnosti za dva primjera, kao susjed se uzima onaj koji je u skupu $\mathcal{D}$ naveden prvi. U slučaju izjednačenja glasova između klasa, prednost se daje klasi s numerički manjom oznakom y. **U koju će klasu biti klasificiran primjer x algoritmom 3-NN, a u koju algoritmom težinski k-NN?**

A $y = 0$ i $y = 0$ B $y = 0$ i $y = 1$ C $y = 0$ i $y = 2$ D $y = 1$ i $y = 1$