

Sissejuhatus tugivektor-masinatele

Hando Tint
Arvutiteaduse instituut, Tartu Ülikool
htint@ut.ee

Andmekaevandamise uurimisseminar MTAT.03.169.
Arvutiteaduse instituut, Tartu Ülikool
Detsember 2003, lk. 136–147

Kokkuvõte

Tugivektor-masinatele (SVM-support vector machines) on suhteliselt uus ja väga tõhus meetod, mida kasutatakse klassifitseerimisel ja regressioonis ning mis põhineb klasside lineaarsel eraldamisel. Lineaarselt mitteeralduvad klassid kujutatakse kõrgema dimensiooniga ruumi ja lineaarne eraldamine teostatakse seal. SVM tugevuseks on see, et püütakse leida optimaalne eraldav hüperatasand, mistõttu nende tulemus on tuntud meetodite omadest üks parimaid. Optimaalse tasandi leidmisel minnakse statistilise õppimise valdkonda. Selles artiklis toome välja SVM tööpõhimõtted ja kirjeldame ka vajalikku osa statistilisest õppimisteooriast.

1 Sissejuhatus

Üks andmekaevanduse tähtsaid ülesandeid on klassifitseerimine. Andmekaevanduse ühe alamosa, neurovõrkude vallas, kasutatakse selleks näiteks mitmekihilisi pertseptroone ja RBF võrke. Sinna valda kuuluvad näiteks ka otsustuspuud. Klassifitseerimise all mõtleme protsessi, kus me otsustame elemendi tunnuste järgi, millisesse klassi ta kuulub. Tavaliselt kasutavad klassifitseerimismeetodid õigete otsustusreeglite leidmiseks näidisandmeid, kus elementidele on juba mõne muu otsustaja poolt klass määratud. Sellist näidisandmete järgi otsustusreeglite genereerimist kutsutakse süsteemi treenimiseks. Üks viimase aja lisandusi selliste meetodite

vallas on tugivektor-masinad (edaspidi SVM - support vector machine). Vaatame juhtu, kus meil on tegemist n -dimensionaalses ruumis kahe klassi lineaarse eraldamisega. Kahe klassi lineaarseks eraldamiseks tekitab SVM ruumi hüper-tasandi, millest ühele poole jäävad ühe ja teisele poole teise klassi isendid. Erinevalt paljudest teistest meetoditest ei piirdata sellise tasandi leidmisega, mis kõiki näidiselemente õigesti klassifi tseerib, vaid püütakse leida *optimaalne* tasand.

Artikli teises peatükis läheme statistilise õppimise ja optimeerimismeetodite valda, kus toome välja Lagrange kordajate meetodi ja näitame, kuidas esialgsel probleemist on võimalik tuletada duaalne probleem, mille lahendamine teatud juhtudel palju efektiivsem on.

Kolmandas peatükis seletame, milles otsustustasandite optimaalsus seisneb ja kuidas selleni jõuda. Edasi näitame, et SVM-ga saab töötada ka lineaarselt mitte eralduvate klassidega, tuues esmalt sisse veatolerantsi ja hiljem andmeruumi kõrgdimensionaalsesse ruumi kujutamise, kus eraldamine lihtsam on.

2 Lagrange'i duaalsus

Enne SVM-de juurde jõudmist toome välja ühe vajaliku meetodi statistilise õppimise vallast. Nimelt piiratud optimeerimisülesannete lahendamise.

Olgu meil järgmine probleem. Meil on vaja minimeerida mingi funktsiooni $f : \mathbb{R}^M \rightarrow \mathbb{R}$, arvestades piirangutega $\{h_i = 0\}$, kus $h_i : \mathbb{R}^M \rightarrow \mathbb{R}, i = 1, 2, \dots, N$. Formaalset võime siis selle probleemi kirja panna järgmiselt:

$$\begin{aligned} \min_{\mathbf{w}} \quad & f(\mathbf{w}) \\ \text{nii et} \quad & h_i(\mathbf{w}) = 0, \quad i = 1, 2, \dots, N, \end{aligned}$$

kus $\mathbf{w} \in \mathbb{R}$. Ehk siis meil oleks vaja leida selline \mathbf{w} , mis kuulub elementide hulka, mis etteantud piiranguid rahldavad, ning samas ei leidu selles hulgas elemente, millega f saaks väiksema väärtuse.

Sellist probleemi saab teatud puhkudel lahendada *Lagrange'i kordajate meetodi* abil. Selles meetodis defi neeritakse *Lagrange'i funktsioon* järgmiselt:

$$L(\mathbf{w}, \beta) = f(\mathbf{w}) + \sum_{i=1}^N \beta_i h_i(\mathbf{w}),$$

kus β_i kutsutakse *Lagrange'i muutujateks*. Lahenduse saaksime, kui võtaksime tuletised nii \mathbf{w} ja β järgi, seaksime tulemused võrdseks 0-ga (ekstreemum) ja la-

hendaksime vastavad võrrandid

$$\frac{\partial L}{\partial \mathbf{w}} = 0; \quad \frac{\partial L}{\partial \beta_i} = 0.$$

Viime nüüd probleemi natuke üldisemale pinnale.

$$\begin{aligned} \min_{\mathbf{w}} \quad & f(\mathbf{w}) \\ \text{nii et} \quad & h_i(\mathbf{w}) = 0, \quad i = 1, 2, \dots, N \\ & g_i(\mathbf{w}) \leq 0, \quad i = 1, 2, \dots, N. \end{aligned}$$

Üldistatud Lagrange'i funktsioon oleks siis

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^N \alpha_i g_i(w) + \sum_{i=1}^N \beta_i h_i(w),$$

kus $\alpha_i \geq 0$ on samuti Lagrange'i kordajad.

Vaatame suurust

$$\theta_P(w) = \max_{\alpha, \beta: \alpha_i \geq 0} L(w, \alpha, \beta),$$

kus P tähistab, et tegu on *peamise* probleemiga.

Meil on vaja, et

$$\theta_P(w) = \begin{cases} f(w) & \text{kui } w \text{ rahuldab kõiki piiranguid} \\ \infty & \text{muidu} \end{cases}$$

Siit näeme, et minimiseerimisprobleem

$$\min_w \theta_P = \min_w \max_{\alpha, \beta: \alpha_i \geq 0} L(w, \alpha, \beta)$$

omab sama lahendust kui meie algne probleem.

Tekitame nüüd teise ehk *duaalse* probleemi:

$$\begin{aligned} \max_{\alpha, \beta: \alpha_i \geq 0} \quad & \theta_D(w), \\ \text{kus} \quad & \theta_D(w) = \min_w L(w, \alpha, \beta). \end{aligned}$$

Kui me nüüd tähistame peamise probleemi lahendi P^* ja duaalse probleemi lahendi D^* , siis kasutades teadmist, et $\max_a \min_b$ millestki on alati väiksem või võrdne $\min_b \max_a$ -ga, näeme, et:

$$D^* = \max_{\alpha, \beta: \alpha_i \geq 0} \min_w L(w, \alpha, \beta) \leq \min_w \max_{\alpha, \beta: \alpha_i \geq 0} L(w, \alpha, \beta) = P^*.$$

Meil oleks vaja leida olukord, kus $D^* = P^*$, sest siis võime me peamise probleemi lahendamiseks lahendada ka duaalse probleemi.

Lemma 1 (Ülo Kaasik & Kivistik 1982)(Nocedal & Wright 1999) Olgu f ja g_i kumerad ja h_i afinne. Kehtigu ka tingimus, et eksisteerib selline w , nii et $g_i(w) < 0 \forall i$. Kui leiduvad sellised α^*, β^*, w^* , mis rahuldavadjärgmisi Karush-Kuhn-Tucker tingimusi:

$$\begin{aligned} \frac{\partial}{\partial w_i} L(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \dots, n \\ \frac{\partial}{\partial \beta_i} L(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \dots, l \\ \alpha_i^* &\geq 0, \quad i = 1, \dots, k \\ g_i(w^*) &\leq 0, \quad i = 1, \dots, k \\ \alpha_i^* g_i(w^*) &= 0, \quad i = 1, \dots, k \end{aligned}$$

siis on α^*, β^*, w^* on nii primaalse kui ka duaalse probleemi lahenditeks.

Viimast tingimust nimetatakse Lagrange'i duaali täiendavaks tingimuseks. Ta näitab, et kui $\alpha_i^* > 0$, siis $g_i(w^*) = 0$. Põhjus, miks on kasulik algne probleem vahevahel duaalseks probleemiks ümber formuleerida, on selles, et uues probleemis esinevad vabade muutujatena vaid Lagrange kordajad. On mitmeid algoritme sellisel kujul probleemide kiireks lahendamiseks. Näiteks (Vishwanathan, Smola, & Murty 2003). Järgmistes peatükkides näeme, et meid SVM puhul on ülesande duaalne püstitus veel eriti kasulik.

Statistilise õppimise kohta sügavama seletuse saamiseks on soovitatav pöörduda (Nocedal & Wright 1999) ja (Ülo Kaasik & Kivistik 1982) poole.

3 Lineaarselt eralduvad klassid

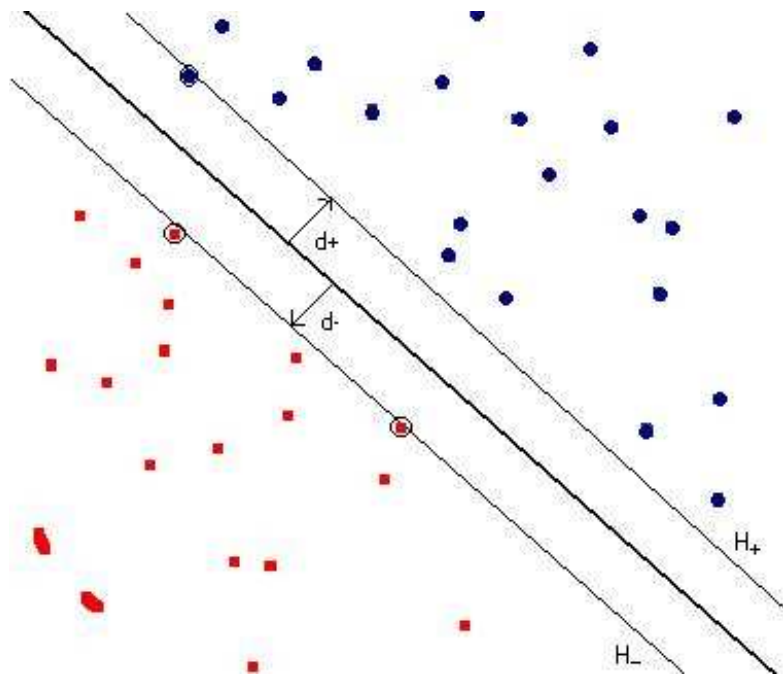
Eelmises peatükis olevat ära kasutades võime nüüd minna SVM-de ühe peamise komponendi, klasse eraldava otsustustasandi genereerimise juurde.

Olgu meil hulk treeningnäiteid kahest lineaarselt eraldatavast klassist $\{(\mathbf{x}_i, d_i)\}_{i=1}^N$, kus $\mathbf{x}_i \in \mathbb{R}^n$ on i . sisendnäide ruumist \mathbb{R}^n ja $d_i \in \{-1, +1\}$ talle vastava klassi identifi kaator. Ruumis \mathbb{R}^n asuv kahte klassi eraldav tasand, mida edaspidi nimetame *otsustustasandiks* on väljendatav võrrandiga

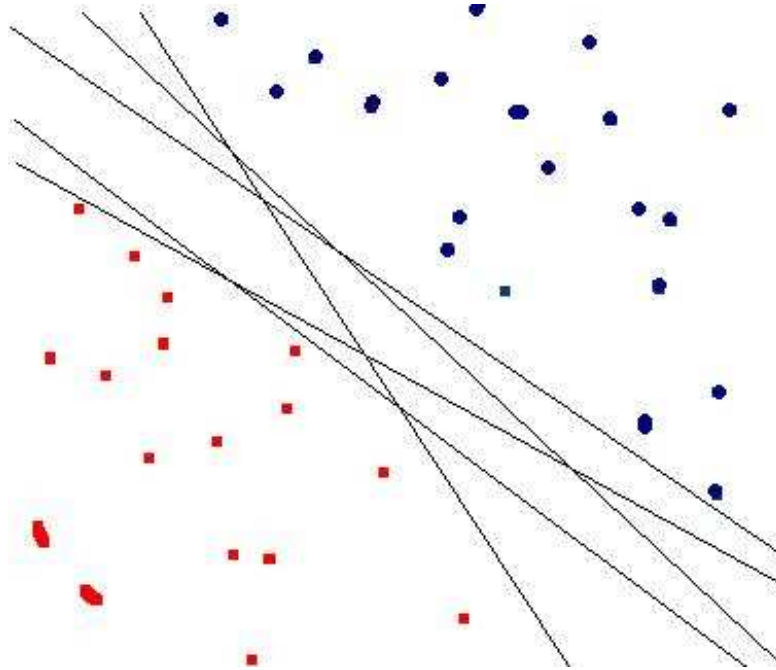
$$\mathbf{w}^T \mathbf{x} + b = 0$$

kus $\mathbf{w} \in \mathbb{R}^n$ ja b on konstant. Me võime seega kirjutada, et

$$\begin{aligned} \mathbf{w}^T \mathbf{x} + b &\geq 0 \quad \text{iga } d_i = +1 \\ \mathbf{w}^T \mathbf{x} + b &< 0 \quad \text{iga } d_i = -1 \end{aligned}$$



Joonis 1: Kahe klassi eraldamine 2-dimensionaalses ruumis. Ringidega on ära näidatud tugivektorid



Joonis 2: Võimalikud klasside eraldamised

On ilmne, et kahte lineaarselt eraldatavat klassi võib eraldada lõpmatu arvu tasanditega (vt joonis 2). Kui näiteks mitmekihiliste pertseptronite treenimise puhul on kõik õigesti klassifi tseerivad tasandid võrdväärsed, siis SVM puhul üritatakse leida *optimaalne tasand*. Optimaalseks tasandiks peetakse seda otsustustasandit kõigist võimalikest, mille puhul on *kaugus tasandist lähima näidisvektorini maksimiseeritud*. Edaspidi nimetame neid näidisvektoreid *tugivektoriteks*. Intuitiivselt on eesmärk põhjendatud. Mida lähemal otsustustasand tugivektoritele on, seda suurem on oht, et klassis leidub elemente, mis on klasside tegelikule eralduspiirile lähemal kui treeningnäidetest leitud tugivektorid ja seega võivad sattuda leitud otsustustasandi valele poolele. Eelpoolmainitud tasand minimiseerib seda ohtu.

Mõlema klassi tugivektorid asuvad otsustuspinnaga $\mathbf{w}_0^T \mathbf{x} + b_0 = 0$ paralleelselt asuvatel tasanditel

$$\begin{aligned} H^+ &: \mathbf{w}_0^T \mathbf{x} + b_0 = +1 \\ H^- &: \mathbf{w}_0^T \mathbf{x} + b_0 = -1 \end{aligned} \tag{1}$$

Paneme tähele, et konstantide $+1$ ja -1 kasutamine on täiesti lubatud, kuna me võime alati w_0 ja b_0 ümber skaleerida, et sinna just need konstandid saada (Ng 2003).

Optimaalse tasandi leidmisel tuleb meile appi järgmine lemma.

Lemma 2 *Optimaalse tasandi leidmine on ekvivalentne $\|\mathbf{w}\|$ minimiseerimisega. Tõestus.*

Tähistame valemis 1 toodud tasandeid vastavalt H_+ ja H_- . Meil on vaja, et otsustustasand oleks neist mõlemast maksimaalselt kaugel. Seega on meie ülesandeks maksimeerida H_+ ja H_- vahelist kaugust. Vastavad tasandite kaugused koordinaatide aluspunktitst on siis $d_+ = \frac{|b-1|}{\|\mathbf{w}\|}$ ja $d_- = \frac{|b+1|}{\|\mathbf{w}\|}$. Kuna tasandid on üksteisega paralleelsed, siis saab leida nendevahelise kauguse

$$d = |d_+ - d_-| = \frac{||b-1| - |b+1||}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}. \quad (2)$$

Näeme, et tasandite vaheline kaugus Viimane võrdus näitab meile seda, et otsustuspinnaga ja tugivektori vahelise kauguse maksimiseerimine on ekvivalentne kaaluvektori \mathbf{w}_0 minimiseerimisega. ■

Me võime seega sõnastada oma ülesande järgmiselt:

Kasutades treeningnäiteid $\{(\mathbf{x}_i, d_i)\}_{i=1}^N$, leida vektorile \mathbf{w} ja nihkele b sellised optimaalsed väärtused, mis rahuldavad võrratust

$$d_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \quad \forall i = 1, 2, \dots, N \quad (3)$$

ja vektor \mathbf{w} minimiseerib hinnangufunktsiooni

$$\Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w}. \quad (4)$$

Me võime lahendada selle optimiseerimisprobleemi kasutades *Lagrange'i kor-dajate meetodit*. Kõigepealt konstrueerime *Lagrange'i funktsiooni*

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^N \alpha_i [d_i(\mathbf{w}^T \mathbf{x}_i + b) - 1],$$

Ülesandeks oleks seda funktsiooni minimiseerida \mathbf{w} ja b suhtes. Diferentseerimine \mathbf{w} ja b järgi annab meile:

$$\frac{\partial J(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = \mathbf{0}$$

$$\mathbf{w} = \sum_{i=1}^N \alpha_i d_i \mathbf{x}_i \quad (5)$$

$$\frac{\partial J(\mathbf{w}, b, \alpha)}{\partial b} = 0$$

$$\sum_{i=1}^N \alpha_i d_i = 0 \quad (6)$$

Kasutades saadud tingimusi esialgse Lagrange'i funktsiooni lihtsustamisel saame, et

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^N \alpha_i d_i \mathbf{w}^T \mathbf{w} - b \sum_{i=1}^N \alpha_i d_i + \sum_{i=1}^N \alpha_i$$

Kolmas liige võrduses on valemist 6 tulenevalt 0. Kasutades valemit 5 saame, et

$$\mathbf{w}^T \mathbf{w} = \sum_{i=1}^N \alpha_i d_i \mathbf{w}^T \mathbf{x}_i = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

Saame

$$J(\mathbf{w}, b, \alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

Nagu näha, sõltub uus hinnangufunktsioon ainult α -st. Veelgi enam, uues funktsioonis esinevad sisendvektorid skalaarkorrutisena. Selle olulisust näeme viimases peatükis. Me saime selle valemi minimiseerides algset funktsiooni \mathbf{w} ja b suhtes. Arvestades tingimust $\alpha_i \geq 0$, mis meil on kogu aeg olnud ja tingimust 6 saame duaalse probleemi funktsiooni, mida peame nüüd maksimiseerida α suhtes:

$$W(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

Lugeja võib kontrollida, et kõik tingimused $P^* = D^*$ tarvis ja ka KKT tingimused

$$\begin{aligned} \sum_{i=1}^N \alpha_i d_i x_i &= \mathbf{w} \\ \sum_{i=1}^N \alpha_i d_i &= 0 \\ d_i(\mathbf{w}^T \mathbf{x}_i + b) &\geq 1 \quad i = 1, \dots, l \\ \alpha_i &\geq 0 \quad \forall i \\ \alpha_i(d_i(\mathbf{w}^T \mathbf{x}_i + b) - 1) &= 0 \quad \forall i \end{aligned}$$

on täidetud. Nüüd võime siis lõplikult sõnastada oma optimeerimisülesande:

Leida treeningnäidete $\{(\mathbf{x}_i, d_i)\}_{i=1}^N$ jaoks Lagrange'i kordajad $\{\alpha_i\}_{i=1}^N$, mis maksimeerivad funktsiooni

$$W(a) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

ja rahuldavad tingimusi

1. $\sum_{i=1}^N \alpha_i d_i = 0$
2. $\alpha_i \geq 0 \quad \forall i = 1, 2, \dots, N$

Leidnud optimaalsed Lagrange'i kordajad α_0 , võime meid huvitava vektori \mathbf{w}_0 leida valemist

$$\mathbf{w}_0 = \sum_{i=1}^N \alpha_{0,i} d_i \mathbf{x}_i \tag{7}$$

ja nihke b_0 võttes arvesse, et KKT tingimuste täiendava tingimuse põhjal tugivektori korral $\alpha_i = 0$,

$$b_0 = 1 - \mathbf{w}_0^T \mathbf{x}_s,$$

kus \mathbf{x}_s on suvaline tugivektor.

4 Otsustustasand lineaarselt mitteeralduvatele klassidele

Liigume edasi raskema ülesande poole, kus klassid pole enam lineaarselt eraldatavad. Toome sisse uued mittenegatiivsed muutujad $\{\xi\}_{i=1}^N$ ja kirjutame valemi 3

üumber:

$$d_i(\mathbf{w}^T \mathbf{x} + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, N$$

ξ_i nimetatakse *nihkemuutujateks* ja nad mõõdavad punkti kaugust ideaalsest klassi eralduskohast. Paneme tähele, et $\xi_i \leq 1$ puhul jääb punkt küll kahe klasse eraldava tasandi vahele, kuid klassifi tserub siiski õigesti. Kirjutame ka valemi 4 ümber linearselt mitte eralduvate klasside jaoks:

$$\Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i$$

C on kasutaja poolt valitud konstant, mis määrab ära, kui tähtsaks me valesti klassifi tserimise probleemi peame.

Järgides eelmise peatüki tegevusviisi, jõuame jällegi duaalse probleemini:

Leida treeningnäidete $\{(\mathbf{x}_i, d_i)\}_{i=1}^N$ jaoks Lagrange'i kordajad $\{\alpha_i\}_{i=1}^N$, mis maksimiseerivad funktsiooni

$$W(a) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

ja rahuldavad tingimusi

1. $\sum_{i=1}^N \alpha_i d_i$
2. $0 \leq \alpha_i \leq C \quad \forall i = 1, 2, \dots, N,$

kus C on kasutaja poolt valitud positiivne parameeter. Lahendus on jällegi toodud valemiga

$$\mathbf{w}_0 = \sum_{i=1}^{N_s} \alpha_i d_i \mathbf{x}_i,$$

kus N_s on tugivektorite arv. KKT täiendtingimused on antud juhul

$$\alpha_i [d_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi] = 0, \quad i = 1, 2, \dots, N \quad (8)$$

$$\mu \xi = 0$$

Juhul, kui $\alpha_i < C$, siis $\xi = 0$. Seega võttes suvalise sellise punkti ja kasutades teda valemis 8, saame me leida b_0 .

5 Tugivektor-masinad

Nüüd on meil kõik vajalik tugivektor-masinate (edaspidi SVM) tööpõhimõtete seletamiseks. SVMi töö käib kahes etapis:

1. Mittelineaarne sisendvektori kujutamine kõrgdimensionaalsesse *varjatud ruumi*.
2. Varjatud ruumis optimaalse tasandi leidmine klasside eraldamiseks.

Esimene punkt täidetakse vastavalt Coveri teoreemile (Cover 1965). See teoreem väidab, et multidimensionaalset ruumi, milles asub kaks lineaarselt mitteeraldatavat klassi, on võimalik teisendada kõrgema dimensiooniga ruumiks, kus klassid on suure tõenäosusega lineaarselt eraldatavad.

Teises punktis kasutatakse meetodikat, mida me eelmistes peatükkides kirjeldasime.

Olgu meil definiieritud kujutis $\{\varphi_j(\mathbf{x})\}_{j=1}^m$, mis sisendruumi m -dimensionaalsesse ruumi kujutab. Otsustustasandi võime siis kirjutada järgmiselt:

$$\sum_{j=1}^m w_j \varphi_j(\mathbf{x}) + b = 0$$

Kui me võtame $\varphi_0(\mathbf{x}) = 1$ ja $w_0 = b$, saame kirjutada

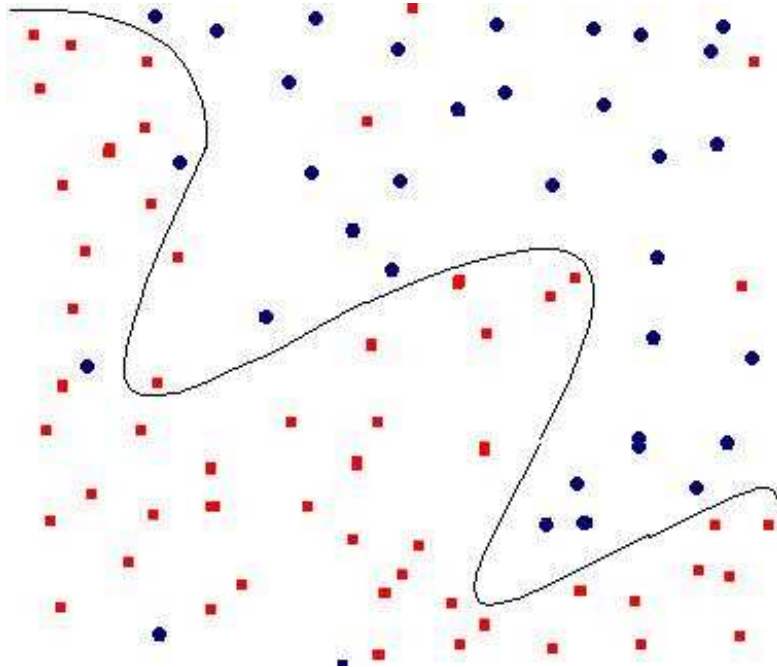
$$\sum_{j=0}^m w_j \varphi_j(\mathbf{x}) = 0$$

või veel lihtsamini

$$\mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = 0$$

Paneme tähele, et sisendruumist vaadates, ei ole tegu enam tasandiga, vaid keerukama mittelineaarse otsustuspinnaga, mis võib hakkama saada ka lineaarselt mitte eralduvate klassidega. (vt joonis 3). Näeme ka kohe, et selle edasimineku eest tuleb meil ka lõivu maksta. Meil on vaja otsida tasantit kõrgema dimensiooniga ruumis ja pealegi sisendid sinna kõigepealt kujutada, mis kõik lisavad algoritmi töömahukust.

Kui me nüüd vaatame eelmises peatükis saadud tulemus, siis võime selle ümber kirjutada järgmiselt:



Joonis 3: Kõrgemas dimensioonis eraldav tasand võib olla alges ruumis mittelineaarne funktsioon

Leida treeningnäidete $\{(\mathbf{x}_i, d_i)\}_{i=1}^N$ jaoks Lagrange'i kordajad $\{\alpha_i\}_{i=1}^N$, mis maksimiseerivad funktsiooni

$$W(a) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j K(\mathbf{x}_i, \mathbf{x}_j)$$

ja rahuldavad võrratusi

1. $\sum_{i=1}^N \alpha_i d_i$
2. $0 \geq \alpha_i \geq C \quad \forall i = 1, 2, \dots, N,$

kus C on kasutaja poolt valitud positiivne parameeter ja

$$K(\mathbf{x}_i, \mathbf{x}_j) = \varphi^T(\mathbf{x}_i) \varphi(\mathbf{x}_j) = \sum_{k=0}^m \varphi_k(\mathbf{x}_i) \varphi_k(\mathbf{x}_j).$$

$K(\mathbf{x}_i, \mathbf{x}_j)$ nimetatakse *tuumaks* (kernel või inner product kernel). Tuumal on tugivektor-masinate juures täita oluline osa just töömahukuse vähendamises. Kui meil õnnestub leida funktsioon $K(\mathbf{x}_1, \mathbf{x}_2)$, mis on mingi φ vastavaks tuumaks, siis saame me optimeerimisel kasutada otse K -d ilma otseselt φ -d ja seega ka kõrgema dimensiooniga ruumi arvesse võtmata. Sellist meetodit nimetatakse *tuumatrikiks* (*kernel trick*). Ühesõnaga saame me opereerida kõrgdimensionaalses ruumis ilma arvutusi sinna ruumi otseselt viimata.

Võtame näiteks laialt levinud tuuma $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + c)^p$. Võtame $p = 2$. Sellisel juhul

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{y}^T \mathbf{x} + c)^2 = \sum_{i,j=1}^N (x_i, x_j)(y_i, y_j) + \sum_{i=1}^N (\sqrt{2cx_i})(\sqrt{2cy_i}) + c^2$$

$n = 3$ puhul oleks kujutis järgmine

$$\varphi(\mathbf{x}) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \\ \sqrt{2cx_1} \\ \sqrt{2cx_2} \\ \sqrt{2cx_3} \\ c \end{bmatrix}.$$

Paneme tähele, et samal ajal kui kujutise arvutamise ajaline keerukus on $O(N^2)$, siis tuuma K leidmine on vaid $O(N)$. Veel enam, p kasvades tõuseb ka kujutise loomise ajaline keerukus, samas kui tuuma leidmine jääb alati lineaarseks.

Kuidas siis kindlaks teha, kas mingi funktsioon sobib tuumaks? Selles osas tuleb appi *Merceri teoreem* (Mercer 1909). Enne teoreemi enda juurde jõudmist defini neerime *tuuma maatriksi*. Olgu meil mingi kindel lõplik punktide hulk $\{x^{(1)}, \dots, x^{(m)}\}$. Tuuma maatriks \mathbf{K} on $m \times m$ maatriks, mille (i, j) element on $K(x^{(i)}, x^{(j)})$. Lihtsustatult ütleb Merceri teoreem:

Olgu meil funktsioon $K : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$. Selleks, et K oleks sobiv tuum, on tarvilik ja piisav, et iga punktihulga $\{x^{(1)}, \dots, x^{(m)}\}$ puhul on talle vastav tuuma maatriks sümmeetriline ja pool-positiivselt määratud.

Teine tihedat kasutust leidev tuumfunktsioon on RBF närvivõrkudes kasutatav

$$K(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{y}\|^2\right).$$

Hetkeprobleemiks on see, et kasutuskõlblikke ja erikujulisi tuumasid, mis seega Merceri teoreemile vastaks, pole palju. On arusaadav, et erinevad teisendused kõrgdimensionaalsesse ruumi võivad tuua erinevaid tulemusi. Seega eeldab tuuma valimine teatavat informatsiooni valdamist sisendandmete kohta, mida meil alati ei tarvitse käepärast olla. Universaalse lahenduse leidmine, mis ise valiks parima teisenduse, ja üldse selle olemasolu tõestamine või ümberlukkamine on hetkel lahtised küsimused.

6 Kokkuvõte

Oleme toonud lühiülevaate populaarse andmekaevanduse algoritmi SVM kohta vaadates teda klassifi tseerimisprobleemide vaatenurgast. SVM-de kaks peamist trumpi on:

- Suvalise sobiva klasside eraldustasandi asemel optimaalse leidmine, mis pärisandmete korral vägagi suurt täpsuse kasvu võib põhjustada.
- Võimalus kõrgdimensionaalsesse ruumi üleminekul hoiduda arvutusraskuse liiga järsust kasvust.

Tööst jäeti välja duaalse probleemi lahendamise algoritm. Sellele probleemile on välja pakutud palju lahendusi. Lugeja võib näiteks tutvuda tööga (Vishwanathan,

Smola, & Murty 2003). Samal põhjusel ei toodud ka välja konkreetset näidet, kuna see oleks nõudnud samuti mingi optimeerimisalgoritmi kasutamist ja selle tutvustamist. Vastavaid näiteid võib leida materjalidest (Haykin 1999) ja (Hastie, Tibshirani, & Friedman 2001). Kasutades mitmeid SVM-e järjestikku on võimalik ka üle kahe klassilisi probleeme lahendada. SVM-d leiavad kasutust ka mitte-lineaarse regressiooni ülesannetes (Hastie, Tibshirani, & Friedman 2001).

Huvitava ja lahendamata probleemina võiks ära märkida selle, et puudub kriteerium otsustamiseks, milline tuum kõige paremini sisendruumi teisendama sobiks. Eriti siis, kui puudub informatsioon sisendruumi struktuuri kohta. Probleemiks on ka töötavate tuumade vähesus ja fakt, et iga tuum ja seega teisendus kõrgema dimensiooniga ruumi on mõeldud töötama suhteliselt kindla sisendruumi struktuuri jaoks ja seega ei pruugi teisendus alati oodatud efekti and või läheb vaja liiga suurt dimensiooni tõstmist.

Viited

- Cover, T. 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*.
- Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer.
- Haykin, S. 1999. *Neural Networks, A Comprehensive foundation*. Prentice Hall, Inc.
- Ülo Kaasik, and Kivistik, L. 1982. *Operatsioonianaliüs*. Tallinn Valgus.
- Mercer, J. 1909. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Transactions of the London Philosophical Society* 209.
- Ng, A. 2003. Tcs229 lecture notes, part v, support vector machines.
- Nocedal, J., and Wright, S. J. 1999. *Numerical Optimization*. Springer Series in Operation Search. Springer.
- Vishwanathan, S.; Smola, A.; and Murty, M. 2003. SimpleSVM. In *Proceedings of the Twentieth International Conference on Machine Learning*.