

УДК 004.588
МРНТИ 20.00.00

МАШИНАМЕН ОҚЫТУ

ЖУНУСОВА А.

ал-Фараби атындағы Қазақ Ұлттық университеті

Аңдатпа: Машинамен оқыту күн сайын оның кеңінен қолданылуының арқасында қазіргі технологиялық әлемде маңызды орын алады. Жол қозғалысын талдаудан бастап, көптеген тапсырмалар өз бетімен жүретін машиналардан, өздігінен үйренетін автомобильдерге ауысуда. Қазір машинамен оқыту екінші және үшінші деңгейлердің түйіскен жерінде, әлемде осы технологиямен өзгеру қарқыны күн сайын артып келеді.

Түйінді сөздер: машинамен оқыту, нейрондық желілер, терең оқыту, машиналық аударма, соңғы өңдеу

МАШИННОЕ ОБУЧЕНИЕ

Аннотация: Машинное обучение растет с каждым днем благодаря широкому спектру приложений. От анализа трафика до автомобилей с автоматическим управлением многие задачи переходят на автомобили с самообучением. Теперь на пересечении второго и третьего уровней машинного обучения, темп изменений в мире с помощью этой технологии растет с каждым днем.

Ключевые слова: машинное обучение, нейронные сети, углубленное обучение, машинный перевод, постредактирование

MACHINE LEARNING

Abstract: Machine learning is growing every day thanks to a wide range of applications. From traffic analysis to self-driving cars, many tasks are shifting to self-learning cars. Now, at the intersection of the second and third levels of machine learning, the pace of change in the world with the help of this technology is growing every day.

Key words: machine learning, neural networks, advanced learning, machine translation, post-editing

Қазіргі уақытта жасанды интеллект, машинамен оқыту және терең білім беру – бұл қазіргі заманғы технологиялардың, сонымен қатар болашақтағы технологиялардың негізі. Олардың әрқайсысының негізгі идеясы – адамдар жасай алатын тапсырманы орындау немесе адамның миы сияқты жұмыс істеу.

Жасанды интеллект, машинамен оқыту және терең білім бірдей ме? Жұмыста осы мәселелер толығымен қарастырылып, зерттелді.

Олардың барлығы бірдей болуы мүмкін, бірақ іс жүзінде, мүлдем бірдей емес. Жасанды интеллект – машинамен оқыту мен терең білім алудың тамыры. Жасанды интеллект машинамен оқытуға және терең оқуға бөлінеді.

Машинамен оқыту кезінде біз машинаны белгілі бір мәліметтер типінен және оның нәтижесінен жағдайдың белгілі бір түрін үйренуге мәжбүр етеміз, ал оның оқу тәжіри-

бесінен жағдайдың ұқсас типі үшін белгісіз мәліметтерді болжай аламыз.

Машинамен оқыту мен терең оқытудың айырмашылығы - олардың жұмыс істеу тәсілі. Машинамен оқыту кезінде модель кеңістіктегі деректерді даналарға немесе мәліметтер нүктелеріне бөледі. Оларды ағаштың көмегімен немесе екі немесе одан да көп даналарды жіктеуге болатын шешім жазықтығын құру немесе бастапқы нәтижемен салыстыру нәтижесін болжау арқылы шешуге тырысады. Жоғалтуды есептеу арқылы, содан кейін әрбір итерация соңында болжанғаннан кейін, шығынды барынша азайту үшін қайтадан итерация қолданылады. Терең білім алу тұжырымдаманы деректер қабатына бөлу болып табылады. Қабат неғұрлым терең болса, оны болжау дәлірек болады, бірақ тереңдігі тым терең болмауы керек, әйтпесе нашар нәтиже береді.

Машинамен оқыту қалай басталды? Машинамен оқыту (Engl. Machine learning, ML) – бұл жасанды интеллект әдістерінің классы. Оның сипаттамасы мәселені тікелей шешу емес, көптеген ұқсас мәселелерге шешім қолдану процесінде жаттығу болып саналады. Мұндай әдістерді құру үшін математикалық статистика, сандық әдістер, оңтайландыру әдістері, ықтималдық теориясы, графтар теориясы, сандық түрде мәліметтермен жұмыс істеудің әртүрлі әдістері қолданылады. Машинамен оқыту күнделікті біздің өмірімізде көптеген қолданыстарға ие болуда. Оның қолданылуының кеңдігіне байланысты IT технологияда маңызды орын алады. Қазіргі кезде машинамен оқыту әдістеріне негізделген кейбір қосымшалар тиімді жұмыс істейді.

Машинамен оқыту жасанды интеллекттің бір саласы болып есептеледі. Оның негізгі идеясы – компьютер алдын ала жазылған алгоритмді қолданумен ғана шектеліп қоймай, мәселені өздігімен шешуді үйрену. Кез келген жұмыс машинамен оқыту технологиясын шартты түрде қолжетімділікке байланысты үш деңгейдің біреуіне тағайындалуы мүмкін. Бірінші деңгей – бұл Google немесе IBM деңгейіндегі әртүрлі технологиялық алыптар

үшін қолжетімді болған кезде. Екінші деңгей – белгілі бір білімі бар студент оны қолдана алатын кезде. Үшінші деңгей – бұл тіпті ата-әжелер оны басқара алатын кез. Қазір машинамен оқыту екінші және үшінші деңгейлердің түйіскен жерінде, осы технологияның көмегімен әлемнің өзгеру қарқыны күн сайын артып келеді.

Машинамен оқыту корпустарды үйренуден басталды. Интуитивті түрде «үйрену» дегеніміз – белгілі бір модель қандай да бір жолмен «үйреніп», содан кейін нәтиже бере бастайды, яғни бір нәрсені болжай алады. Томас Митчеллдің «Машинамен оқыту» [1] кітабында келтірілген «үйренуге» деген жалпы анықтама беруге болады: «Компьютерлік бағдарлама D проблемалары мен объективті функцияларының белгілі бір класына қатысты тәжірибе жинақталған сайын үйренеді, егер оларды шешудің сапасы болса, тапсырмалар (P-ге қатысты) жаңа тәжірибемен жақсарады. Бұл анықтама өте жалпылама және дерексіз болып көрінсе де, іс жүзінде кейбір маңызды тұстарды нақтылауға мүмкіндік береді. Мысалы, ондағы орталық орынды деректер (объективті функция) алады. Кез келген практикалық мәселені шеше бастағанда, объективті функцияны анықтау және нәтижелерді бағалайтындығының мәні зор. Объективті функцияны тандау барлық келесі жұмысты толық анықтайды, тіпті ұқсас міндеттерде де әртүрлі объективті функциялар мүлдем басқа модельдерге әкелуі мүмкін. Мысалы, «компьютерді оқуды үйрету» жақсы болар еді, бірақ алдымен оның нені білдіретінін анықтау керек. «Оқу» мәтіні бойынша сұрақтарға дұрыс жауап бере алу үшін, сөйлемді талдауда, осы мәтінге қатысты ең қажетті Wikipedia-ны көрсету қолданылады. Сан түрлі жауаптар әрқелкі модельдерге және зерттеу бағыттарына әкеледі. Бірақ Митчеллдің анықтамасы бізге жеткіліксіз болады. Машиналарды оқытудың қандай қиындықтары бар, ол неден тұрады? Осы сұрақ машинамен оқытудың классификациясымен толық байланысты. Енді машинамен оқыту тапсырмаларының негізгі классификациясына тоқталсақ: машинамен оқытудың негізгі

екі сыныбы мұғаліммен бірге оқу (бақыланатын оқыту) және мұғалімсіз оқу (бақыланбайтын оқыту) болады. «Бақыланатын» түсінігінде адамның мәліметтерді өңдеуге араласуы деп ұғынады. Адам машинамен оқытуға араласқан кезде бізде белгілі бір болжамалы ақпараттар бар. Ал адамсыз оқу кезінде бізде тек анықталатын мәліметтер болады.

Бақыланатын оқыту [2]

Мысалы, бізде Алматы қаласындағы 10 000 пәтерлер туралы мәліметтер бар. Сондай-ақ әр пәтердің ауданы, бөлмелер саны, орналасқан қабаты, автотұрақтың болуы, метро станциясына дейінгі қашықтық, сонымен қатар әр пәтердің құны және басқалары белгілі. Біздің міндетіміз – осы белгілер негізінде, пәтердің құнын болжай алатын модель құру. Бұл бақыланатын оқытудың классикалық мысалы. Мұндай тапсырма регрессиялық есеп деп аталады. Басқа мысалдар: әрқилы медициналық көрсеткіштерге негізделген ақпараттар бойынша науқаста қатерлі ісіктің бар-жоғын болжау немесе электрондық поштаның мәтініне сүйене отырып, осы спамның ықтималдығын болжау және т.б.

Бақыланбайтын оқыту. Оқытудың бұл түрі өте қызықты болып келеді, себебі біз нақты дұрыс жауабын білмейміз. Атап айтсақ, белгілі бір адамдардың бойы мен салмағы туралы деректер берілсін. Деректерді 3 санатқа (немесе топ) топтастыру керек. Осыған орай адамдардың әр санаты үшін қолайлы өлшемдегі көйлек жасау керек. Бұл тапсырма кластерлік тапсырмасы деп аталады.

Мұғаліммен жаттығу кезінде жаттығу мысалдарының жиынтығы енгізіледі, оны әдетте жаттығулар немесе жаттығулар туралы мәліметтер жиынтығы (жаттығу жиынтығы немесе жаттығу үлгісі) деп атайды, ал міндет – жаңа тәжірибеге бұрыннан белгілі жауаптарды, не болмаса, тест түрінде көрсетілетін жалғастыру, мәліметтер жиынтығы (тест жиынтығы, сынама үлгісі). Мұндағы басты болжам – оқыту үшін қолжетімді мәліметтер дайындалған модельдің мәліметіне ұқсас келеді, әйтпесе жалпылау мүмкін болмайды. Мәтінді «оқу» үшін мұғалім-

мен сабақ жүргізудің мысалы: адамдар нақты сөйлемдер құрған ағаштар жиынтығынан талдаушы ағаштарын (сөздердің қай сөйлемдерге байланысты) құрайтын модельді үйрету. Мұндағы болжам бойынша, талдаушы ағаштар бірдей заңдарға сәйкес жасалады, ал талданатын ағаштардың белгілі бір жиынтығында дайындалған үлгіні оқу жиынына кірмеген жаңа сөйлемдерге қолдануға болады. Егер бұл болжам бұзылса, модель жұмыс істемейді. Мысалы, егер лингвистер сөйлемдерді ағылшын тілінде жазып, содан кейін үйренген модельді неміс тілінде қолдана, онда әріптер бірдей, бірақ синтаксис мүлде өзгеше, модельден мағыналы сөйлемдерді күту мүмкін емес. Мұғаліммен жүргізілетін оқу міндеттері кейде жіктеу және регрессиялық тапсырмалар болып бөлінеді. Дәлдік машинамен оқыту моделінде басым рөл атқарады және ол модельдің жұмысына тікелей пропорционал, ал терең оқытуда ол маңызды рөл атқарады, бірақ ол модельдің жұмысына тікелей пропорционалды емес. Машинамен оқыту жануарларды ажыратуда қолданылады. Егер жіктеуді тереңдетіп оқыту моделі оқытылса және оның моделінің дәлдігі 1 болса, аталған модель бетінде тырнақтары бар итке қарсы сынақтан өткізілсе, онда ол оны ит ретінде танымайды, өйткені дайындалған мәліметтерге қатысты шамалы өзгеріс дұрыс емес нәтиже беруі мүмкін. 80% немесе 90% дәлдік терең модельдеу кезінде 100% дәлдігі бар модельге қарағанда, жақсы үлгіге айналуы мүмкін. Модель өте сезімтал болады. Жіктеу мәселесінде, мысалы, жануарлардың суреттерін қай жануар екенін анықтау үшін, алдымен жануарларды барлық жануарларға: мысықтарға, иттерге, жылқыларға бөліп, қасиеттерін анықтау керек. Дәл сондай адамдардың желідегі суреттерін анықтау үшін де адамдар анықтай алатындай желінің бағдарламасын оқыту қажет. Егер сіз мысалды тілмен жалғастыратын болсаңыз, онда типтік жіктеу тапсырмасы сөйлеу бөліктеріндегі сөздерді белгілеу болып табылады. Ал регрессия мәселесінде белгілі бір функцияның мәнін болжау керек, ол негізінде шексіз көптеген әртүрлі мәндерге ие болады. Мысалы,

адамның бойына байланысты оның салмағын болжау, ертеңгі ауа-райын болжау, акцияның бағасын болжау немесе суретте адамның беті орналасқан тіктөртбұрышты бөліп көрсету – бұл тіктөртбұрыштар жоғарыда аталған классификаторға берілуі үшін қажет. Регрессия мен классификацияға бөлу, әрине, өте еркін, «аралық» мысалдарды оңай табуға болады (сөйлемді бірдей талдау – «ағаш салу» міндетіне жатады). Дегенмен, қандай мәселені шешіп жатқанымыз анық және бұл бөлінудің мағыналы мәні бар: объективті функциялар дұрыс анықтау, нәтижесінде оқу процесін тура бағытқа өзгертеді. Бұл қарапайым жіктеуге сәйкес келмейтін тапсырмалардың әрқелкі түрлері бар. Мысалы, іздеу және ұсыну жүйелерінде рейтингті үйрену міндеті жиі кездеседі. Ол былай қойылады: қолда бар мәліметтерге сәйкес (іздеу жүйесінде бұл құжаттардың мәтіндері және бұрынғы пайдаланушылардың әрекеттері болады), қолжетімді нысандарды объективті функцияның төмендеу реті бойынша реттелуі қажет. Мұндағы реттеу (іздеу жүйесі мен оның өзектілігін көрсетеді: құжат оны жауап ретінде көрсету үшін қаншалықты орынды болатынын көрсетеді) сұралады. Ендігі міндет регрессия мәселесіне біршама ұқсас – үнемі объективті функцияны, оның өзектілігін болжауымыз керек. Бірақ функцияның мәндері деректерге мүлде кірмейді және олар өте маңызды емес. Бұл функцияның сан түрлі нысандарда салыстырудың нәтижелері ғана маңызды (іздеу нәтижелері бойынша қандай нәтижелер шығатыны мәнді). Бұл бірқатар қызықты және нақты оқыту әдістеріне әкеледі. Егер белгілі бір тапсырмаға сәйкес белгіленген мәліметтер жиынтығы болмаса, бірақ сізде «белгілі бір мағынаны табу» керек мәліметтер болса, онда мұғалімсіз оқу қиындықтары болады. Мұғалімсіз оқу проблемасының типтік мысалы – кластерлеу: бір кластерге берілген ұпайлар бір-біріне мүмкіндігінше жақын, тиісінше жақын болу үшін, белгілі бір ұқсастықпен бұрын белгісіз сыныптарға мәліметтерді бөлу керек және әртүрлі кластерлердің нүктелері мүмкіндігінше, бір-бірінен алыстап кетпей, керісінше ұқсас болуын

қамтамасыз ету қажет. Мысалы, кластерлік тапсырманы шеше отырып, сіз гендік отбасыларды нуклеотидтер тізбегінен немесе веб-сайтыңыздың кластерлік пайдаланушыларынан бөліп, әр кластер үшін жекелендіре аласыз немесе ісіктің қай жерде екенін түсіну үшін медициналық суретті сегментациялай аласыз. Мұғалімнің көмегінсіз оқудың тағы бір жиі кездесетін міндеті – кірістірілген мәліметтер үлкен өлшемге ие болған кезде деректер өлшемдерінің азаюы (мысалы, егер сіз мәтінге ондаған, мыңдаған өлшемдер сөзбен жазылған болсаңыз, ал фотосуреттер миллиондаған өлшемнен тұрған болса), онда тапсырма бастапқы деректерді толығымен көрсететін кішірек өлшемдегі деректердің тұсаукесерін құру болып табылады.

Машинамен оқудағы шектеулер [3]:

Дәстүрлі түрде машинамен оқуда бірнеше жалпы және маңызды шектеулер бар. Асып кету – бұл машинаны оқыту алгоритміне, әсер етуі мүмкін статистикалық мәселе. Машиналарды оқыту алгоритмінде деректерді талдау мен болжау кезінде белгілі бір «қателіктер» бар. Алгоритм сәйкес айнымалылар арасындағы байланысты көрсетеді деп болжанады, бірақ шамадан тыс бөлшектеу кезінде ол қателіктерді түсіре бастайды, бұл «шулы» немесе дұрыс емес модельге әкеледі. Машинамен жұмыс жасау модельдері, сонымен бірге олар оқыған мәліметтердің түсінбестіктеріне бейім болуы мүмкін, проблема алгоритмді тексеру үшін зерттеушілер деректердің бір бөлігін сақтаудың орнына, барлық қолжетімді мәліметтер базасында алгоритмдерді оқытқанда айқын көрінеді.

Машинамен оқыту дәстүрлі түрде алдын ала білуге және болжауға болатын шағын деректер жиынтығын пайдаланады. Деректердің аз мөлшерімен зерттеушілер машинаны оқу бағдарламасын деректерді түсінуге және үйренуге көмектесетін нақты мүмкіндіктерді анықтай алады. Алайда бағдарлама бұрыннан бар алгоритмдер негізінде жіктей алмайтын ақпаратқа еңсе, зерттеушілерге, әдетте проблемалық деректерді қолмен талдап, жаңа мүмкіндік жасау керек болады. Міне, осыған

байланысты, машинаны классикалық түрде үйрену, көбінесе үлкен көлемдегі мәліметтермен дұрыс өлшенбейді, бірақ ол кішігірім деректер жиынтығындағы қателіктерді азайтуға мүмкіндік береді.

Машинамен оқыту компьютердің жұмысына өзгермелі талаптар қояды. Орташа дербес компьютерде жұмыс істеуге болатын көптеген модельдер бар. Статистикалық және математикалық әдістер неғұрлым жетілдірілген болса, компьютерге деректерді тез өңдеудің қиынға соғатыны мәлім.

Машинамен оқытудың класстарына тоқталсақ: регрессиялық мәселелер. Әрқилы белгілерге сүйене отырып, материалдық реакцияны болжау. Басқаша айтқанда, жауап 1, 5, 23.575 немесе кез келген нақты сан болуы мүмкін [4].

Жіктеу мәселесі. Әркелкі белгілер негізінде категориялық жауапты болжау. Мысалдар: мәтінді қолжазбамен тану, фотосуреттегі адамның немесе мысықтың бар-жоғын анықтау.

Класстерлік мәселелер: мәліметтерді ұқсас категорияларға бөлу. Мысалдар: ұялы байланыс операторы клиенттерінің төлем қабілеттілігі бойынша бөлу, ғарыш объектілерінің ұқсас заттарға бөлу және т.б. (галактикалар, планеталар, жұлдыздар және басқалар).

Өлшемді азайту мәселесі: біздің мәліметтерімізді N белгілерімен емес, кішірек санмен

сипаттауға үйрену, (мейлінше кейінгі визуализация үшін 2-3). Визуализация қажеттілігінен басқа деректерді сығуды, мысал ретінде келтіруге болады.

Аномалияларды анықтауға байланысты мәселелер: белгілер негізінде аномалияларды «аномалиялардан» айыра білуді үйрену. Бұл міндеттің жіктеу мәселесінен еш айырмашылығы жоқ сияқты. Бірақ аномалияны анықтаудың ерекшелігі – бізде модельді үйрету үшін аномалиялардың мысалдары өте аз немесе мүлдем жоқ. Сондықтан біз классификация мәселесін шеше алмаймыз [5].

Қорытынды

Осылайша біздің зерттеуіміздің нәтижесінде мынадай тұжырымдар жасалады:

1) қазіргі заманғы машинамен оқыту жүйелері, ақпараттың үлкен көлемін аудару кезінде таптырмас көмекшілерге айналды, өйткені олар уақыт пен күш-жігердің шығындарын үнемдеуге мүмкіндік береді;

2) аударма сапасын арттыру мақсатында, жүйе үнемі жаңартылып, тексеріліп отырады, бұл аудармашылар жұмысын едәуір жеңілдетеді және жеделдетеді. Заманауи машиналық аударма технологияларының жұмысы талданады. Онлайн аудармашылардың жұмысы бұрын қазақ тіліне және керісінше аударылатын. Машинамен оқыту әдістері көмегімен машиналық аударманың сапасын жақсартуға болады.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР

1. Abdi tf, Williams L.J. Principal Component Analysis//Wiley Interdisciplinary Reviews: Computational Statistics, 2010, vol. 2, no. 4. – P. 433-459.
2. Ackley D. H., Hinton G. E., Sejnowski T. J. A Learning Algorithm for Boltzmann Machines // Connectionist Models and Their Implications: Readings from Cognitive Science / Norwood, NJ, USA: Ablex Publishing Corp., 1988. - P. 285-307.
3. Addressing the Rare Word Problem in Neural Machine Translation/T. Luong et al. //Proc. 53rd ACL and the 7th IJCNLP, Vol. 1: Long Papers, Beijing, China: ACL, 2015. – P. 11-19.
4. Adversarial Autoencoders / A. Makhzani et al. // arXiv, 2015. <http://arxiv.org/abs/1511.05644>.
5. Aggarwal C. C., Reddy C. K. Data Clustering: Algorithms and Applications, Chapman and Hall/CRC, 2013.
6. Antipov G.y Baccouche M., DugelayJ. Face Aging With Conditional Generative Adversarial Networks //arXiv, 2017. <http://arxiv.org/abs/1702.01983>.