**ДЕРЕКТЕРДІ ТАЛДАУ. ДЕРЕКТЕРДІ БАСҚАРУ**

**Деректерді талдау базасы. Жинау, жіктеу және болжамдау әдістері. Шешімдер дарағы. Деректердің үлкен көлемдерін өңдеу. Деректердің зияткерлік талдау әдістері мен кезеңдері. Деректердің зияткерлік талдау міндеттері. Деректердің визуалдылығы**

1. Деректерге кіріспе
2. Өңдеу және деректерді талдау процесі
3. Деректерді жинау

3.1 Деректерді талдау әдістері

1. Деректерді өңдеу
2. Деректердің визуалдылығы

5.1 Кестелерді құрастыруға арналған Python кітапханалары

1. Деректерді талдау

6.1 Жіктеме

6.2 Регрессия

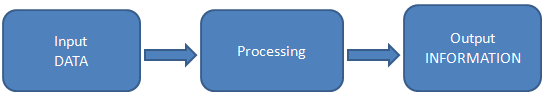
6.3 Шешімдер дарағы

1. Деректерді басқару

**Деректерге кіріспе**

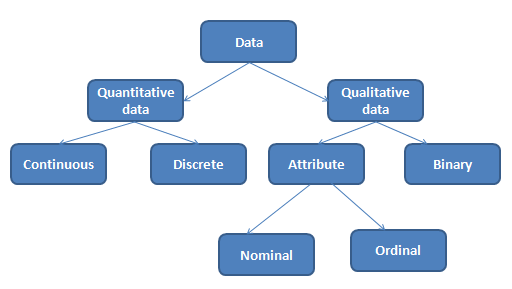
Деректер — ақпаратты алу немесе мәліметтерді жинау үшін түрлендіру қажетті, өңделмеген фактілер мен сандардың жиынтығы. Деректер компьютерлік жүйенің «ресурсы» ретінде пайдаланылады және де ақпарат содан құрастырылған базалық блок болып табылады.

Жалпы алғанда ақпарат – бұл тек өңделген деректер, нақты ақпарат деп жиі өңделген, ұйымдастырылған және құрылымдалған деректер нысаны түсіндіріледі.



|  |  |  |
| --- | --- | --- |
| Input DATA | Входные ДАННЫЕ | Кіріс ДЕРЕКТЕРІ |
| Processing | Обработка | Өңдеу |
| Output INFORMATION | Выходная ИНФОРМАЦИЯ | Шығу АҚПАРАТЫ |

Деректердің екі типі болады: сандық және сапалы. Сапалы деректер – бұл өзгеріске ұшырамайтын деректер. Оларға сезім, жеке тәжірибе, қарым-қатынастар және ниеттер кіреді. Бұл деректер фокус-топтар, сауалнамалар, хабарлы мәтіндер және есептер сияқты көздерден алынады. Сандық деректер – бұл жағдайларды өзара салыстыру үшін, сонымен қатар уақыт өтумен бір жағдайдың шеңберінде жағдайдың өзгеруін қадағалау үшін, жағдайларды объективті бағадауға мүмкіндік беретін, өзгеруге жататын сандық деректер.



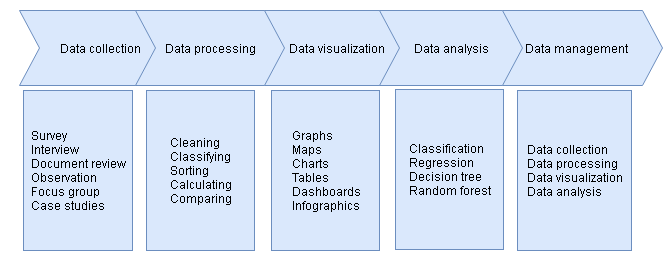
|  |  |  |
| --- | --- | --- |
| Data | Данные | Деректер |
| Quantitative Data | Количественные данные | Сандық деректер |
| Continuous | Непрерывные | Үздіксіз |
| Discrete | Дискретные | Дискретті |
| Attribute | Атрибутивные | Атрибутивті |
| Binary | Двоичные | Екілік |
| Nominal | Номинальные | Номиналды |
| Ordinal | Порядковые | Реттік |

Деректер типтерінің мысалдары

|  |  |  |
| --- | --- | --- |
| Тип | Анықтама | Мысал |
| Үздіксіз | Өзгертуге болатын деректер | Адамның жас мөлшері, бойы, салмағы  Ақша  Уақыт |
| Дискретті | Есептеу негізіндегі деректер | Студенттердің саны  Зообақтағы жануарлардың саны |
| Номиналды | Бірнеше қиылыспайтын топтарға бөлуге болатын Размеченные және аталған деректер | Қазақ, орыс, ағылшын  Ерлер, әйелдер  Ақ, қара,қызыл |
| Реттік | Деректер бір-біріне қатысты реттелген түрде белгілі бір тәртіпте болады | Өте жақсы, жақсы, жаман  Қанағаттанған, қанағаттанбаған  Бақытты, бақытсыз, қалыпты |
| Екілік | Тек қана екі ықтималды мәнді қабылдай алатын деректер | Ия. жоқ  Сәттілік, сәтсіздік  Шындық, өтірік |

**Өңдеу және деректерді талдау процесі**

Өңдеу және деректерді талдау процесіне шартты түрде бес кезең кіреді: өңдеу, визуалдылық, талдау және басқару. Деректермен жұмыс істеуді бастау үшін, біріншіден, біз оларды жинау керекпіз. Бұған барлық қаралатын деректерге сәйкестендіру және қолжетімділік кіреді. Сонан соң бізге деректерді зерттеу және оларды өңдеу қажет. Деректерді дайындауда бірінші кезең деректерді шығуын, мәнін, сапасын және форматын түсіну мақсатында, оларды нақты қарау болып табылады. Өңдеу процесіне деректерді тазарту, деректерді сүзгілеу немесе оларды қосалқы жиынтықтарға бөлу, бағдарламалармен оқылып, айырып танылуы мүмкін деректерді жасау, мысалы, өңделмеген деректерден деректердің бұдан да белгілі моделін модельдеу немесе деректерді белгілі форматқа қаптау кіреді. Келесі кезең – нәтижелердің визуалдылығы. Сонан соң дайындалған деректер ішіне талдамалық әдістерді таңдау, деректердің моделін құрастыру және нәтижелерді талдау кіретін, талду кезеңіне табысталады. Талдаудан кейін сіз өзіңіздің қалауыңыз бойынша деректерді басқара аласыз.



|  |  |  |
| --- | --- | --- |
| Data collection | Сбор данных | Деректерді жинау |
| Data processing | Обработка данных | Деректерді өңдеу |
| Data visualization | Визуализация данных | Деректердің визуалдылығы |
| Data analysis | Анализ данных | Деректерді талдау |
| Data management | Управление данными | Деректерді басқару |
| Survey Interview  Document review  Observation  Focus group  Case studies | Опрос Интервью  Пересмотр документов  Наблюдение  Фокус-группа  Анализ конкретных примеров | Сауалнама Сұхбаттасу  Құжаттарды қайта қарау  Бақылау  Фокус-топ  Нақты мысалдарды талдау |
| Cleaning  Classifying  Sorting  Calculating  Comparing | Очистка  Классификация  Сортировка  Расчет  Сравнение | Тазарту  Жіктеме  Сұрыптау  Есептеу  Салыстыру |
| Graphs  Maps  Charts  Tables  Dashboards  Infographics | Диаграммы  Карты  Графики  Таблицы  Информационные панели  Инфографика | Диаграммалар  Карталар  Графиктер  Кестелер  Ақпараттық панельдер  Инфографика |
| Classification  Regression  Decision tree  Random forest | Классификация  Регрессия  Дерево решений  «Случайный лес» | Жіктеме  Регрессия  Шешімдер дарағы  «Кездейсоқ орман» |
| Data collection  Data processing  Data visualization  Data analysis | Сбор данных  Обработка данных  Визуализация данных  Анализ данных | Деректерді жинау  Деректерді өңдеу  Деректердің визуалдылығы  Деректерді талдау |

**Деректерді жинау**

Деректерді жинау қаралатын саланың толық және нақты суреттемесін алу үшін әртүрлі көздерден ақпаратты жинауға және өзгертуге жүйелі тәсілдеме болып табылады. Ақпаратты жинаудың көптеген әдістері, сонымен қатар ақпарат көздерінің көп әртүрлілігі бар. Төменде деректерді жинаудың ең көп таралған әдістерінің кейбіреулері атап өтілген:

1. *Сауал жүргізу, сауалнама*. Сауалнама — бұл ең қолайлы ақпаратты немесе мәліметтерді автордың сұрақтарына жауап ретінде таңдау.

2. *Сұхбаттасу*. Сұхбаттасу — бір немесе бірнеше адамдардан егжей-тегжейлі ақпаратты жинаудың тамаша тәсілі. Бұл егер де қаралатын сұрақ бойынша сараптамалық пікірді алу немесе осы тақырыпта жақсы хабардар болған біреумен әңгімелесу қажет болса, пайдалы болады.

3. *Құжатарды қайта қарау*. Автор кейбір маңызды деректерді әртүрлі деректер базаларынан, жинақтарының, есептердің хаттамаларынан, қатысу журналдарынан, қаржылық құжаттардан, ақпараттық бюллетеньдерден жинай алады.

4. *Бақылау*. Автор әртүрлі позициялардан, орындарда тікелей бақылауларды жүргізе алады, бұл орындарда нақты жағдайда жүйелердің жүріс-тұрысын түсінуге көмектеседі.

5. *Фокус-топ*. Өзара бір нәрсе ортақ адамдармен оңайлатылған топтық сұхаттасу.

6. *Нақты мысалдарды талдау*. Бұл әдіте спецификалық және қзықты оқиғалар қаралады.

**Деректерді өңдеу**

Өңдеу процесінде деректерді жинап, пайдалы ақпаратқа түрлендіріледі. Бұл кезеңде өңделмеген деректер әртүрлі мақсаттар үшін түсіндіруге, талдауға және пайдалануға болатын, оқылатын форматқа түрлендіріледі. Бұл тазарту, жіктеме, сұрыптау, есептеу, сомалау, салыстыру және деректерді ақпаратқа айналдыратын тағы басқа сияқты деректерді өңдеудің нақты әдістері деп ұйғарылады.

Мамандар деректерді тазарту кезеңінде көп уақыт кетіреді. Оған мыналар кіреді:

1. Керексіз бақылауларды жою

2. Құрылымдық қателіктерді түзету

3. Экстремалды мәндерді сүзгілеу

4. Жетіспейтін деректер проблемаларының шешімі

**Деректердің визуалдылығы**

Үздік деректер – әрқайсысы бұл көріп, түсінуге болатын деректер. Сондықтан деректерді өңдеу және талдау жөніндегі мамандар ең толық және түсінікті визуалдылықтарды жасап, ұсынуды жөн көреді. Бұл әрдайым оңай емес: деректерді табу, оларды қарау, тазарту, сүзгілеу, ал сонан соң оның визуалдылығы үшін дұрыс құралды пайдалану талап етіледі.

Деректердің визуалдылығы — ақпаратты графикалық ұсыну. Ол деректердегі динамиканың, экстремалды мәндердің және заңдылықтардың қарауды және түсінуді жеңілдететін диаграммалар, графиктер және карталар сияқты визуалды элементтердің көмегімен орындалады.

Python деректерді талдау және өңдеу кезінде пайдаланылатын бағдарламалаудың ең танымал тілдерінің бірі болып табылады, сондықтан оның ашық БЖ-ның әзірлеушілерінің қоғамдастығымен әзірленген пайдалы кітапханалардың көптеген саны бар. Python ішіне әртүрлі функциялардың көп саны кіретін деректердің визуалдылығы үшін графиктерді құрастырудың бірнеше кітапханасын ұсынады.Бұл matplotlib, seaborn, ggplot, plotly, pandas visualization.

**Деректерді талдау**

Деректерді талдау – бұл деректерді ұсынудың барлық компонеттерін зерттеуге арналған талдамалық және логикалық әдістерді пайдаланумен деректерді бағалау процесі. Деректерді талдаудың бұдан да ілгері әдістемелеріне динамиканы, заңдылықтарды және өзара байланыстарды анықтау мақсатында деректердің үлкен ауқымдарын мақсатқа бағытты сұрыптауын қарастыратын, деректердің зияткерлік талдауы кіреді.

Деректерді зияткерлік талдау — бұл машиналық оқыту, статистика және деректер жоқ жүйелер қиылысында әдістерді қосатын талдау кезеңі. Оған деректерді алу, деректерді тазарту, деректердің қосылуы, деректердің көлемін қысқарту және белгілерді қалыптастыру сияқты алдын ала өңдеу міндеттері, сонымен қатар, шаблондар мен модельдерді түсіндіру, гипотезаларды растау және генерациялау сияқты алдын ала өңдеу кезеңдері кіреді.

Машиналық оқыту – бұл талдамалық модельдердің құрастыруын автоматтандыратын деректерді талдау әдісі. Ол жүйелерге айқын түрде бағдарламалаусыз, тәжірибемен жетілдіру және автоматты түрде оқу мүмкіндігін ұсынады. Машиналық оқыту алгоритмдерін әзірлеу кезінде пайдаланылатын бес негізгі қадам келесіге келтіреді:

1. Белгілерді таңдау және оқыту іріктемелерін жинау
2. Өнімділік метрикасын таңдау
3. Жіктеуішті және оңтайландыру алгоритмін таңдау

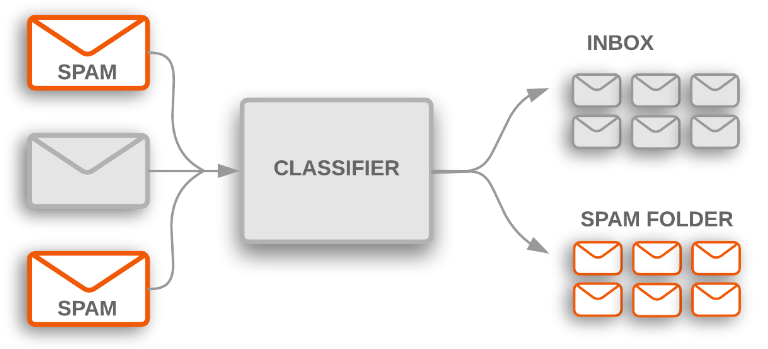
4. Модельдің өнімділігін бағалау

5. Алгоритмді баптау

Машиналық оқытудың екі типі бар: бақыланатын және бақыланбайтын оқыту.

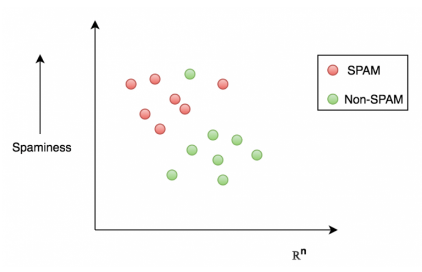
Бақыланатын оқытудың негізгі мақсаты – көзге көрінбейтін және кейінгі деректер туралы болжамдарды жасауға мүмкіндік беретін жоспарланған модельді оқыту деректері бойынша зерттеу.

Жағымсыз электрондық поштаның (спамның) сүзгілеу мысалын қарап, біз ол жаңа хатты «Кіріс» папкасына немесе «Спам» папкасына орналастыру қажет екенін дұрыс болжай алатындай, жоспарланған хаттардың жиналысында модельді машиналық оқытудың бақыланбалы алгоритмін оқытамыз. Кластардың дискреттік белгілерін пайдаланумен оқытудың бақыланбалы міндеті, мысалға, жағымсыз поштаны сүзгілеудің өткен мысалында сияқты, жіктеме аталады.



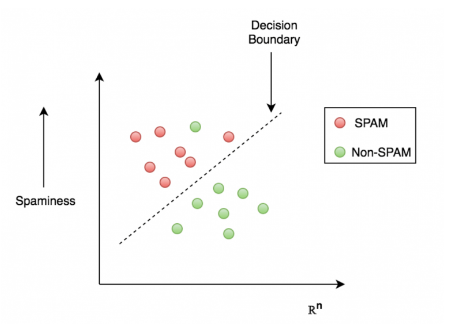
|  |  |  |
| --- | --- | --- |
| SPAM | СПАМ | СПАМ |
| CLASSIFIER | КЛАССИФИКАТОР | ЖІКТЕМЕ |
| INBOX | ПАПКА «ВХОДЯЩИЕ» | «КІРІС» ПАПКАСЫ |
| SPAM FOLDER | ПАПКА «СПАМ» | «СПАМ» ПАПКАСЫ |

Суретте көрсетілгендей, пошталық бағдарлама кіріс хаттың «Кіріс» папкасына орналастырылуын анықтау немесе «Спам» папкасына сүзгілеу үшін, мәтіндік жіктеуішті пайдаланады. Бұл мысал машиналық оқыту алгоритмдері екі ықтимал класты: «спам» және «спам-емес» хаттарды айыруға арналған қағидалар жиынына оқытатын, екілік жіктеменің үлгілік міндеті болып саналады.



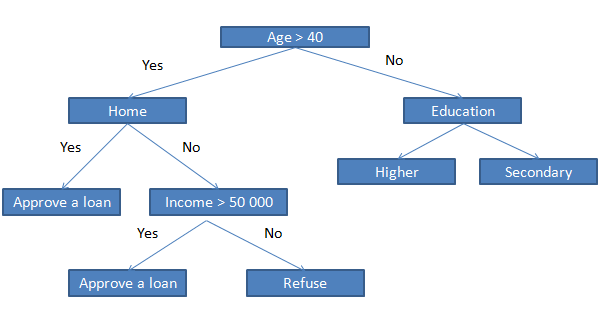
|  |  |  |
| --- | --- | --- |
| Spaminess | Спамность | Спамдық |
| SPAM | СПАМ | СПАМ |
| Non-SPAM | Не-СПАМ | СПАМ-емес |

Бақыланатын оқытудың басқа санаты – сандық ауыспалылар үшін нысанайлы мәнді бағалуға немесе болжауға тырысатын регрессия. Регрессияда прогностикалық (түсіндіретін) ауыспалылар қатары және үздіксіз (нысаналы) ауыспалы беріледі, және де бізге нәтижені болжауға мүмкіндік беретін, осы ауыспалылар арасындағы байланысты табу қажет.



|  |  |  |
| --- | --- | --- |
| Spaminess | Спамность | Спамдық |
| Decision Boundary | Граница принятия решений | Шешімдерді қабылдау шекарасы |
| SPAM | СПАМ | СПАМ |
| Non-SPAM | Не-СПАМ | СПАМ-емес |

Машиналық оқытудағы жіктеме алгоритмінің бірі шешімдер құралы болып табылады. Шешімдер құралы – бұл бірқатар мәселелерді қоюға негізделген шешімді қабылдай отырып, біздің деректерімізді бөлетін модель.



|  |  |  |
| --- | --- | --- |
| Age >40 | Возраст > 40 лет | Жас мөлшері > 40 жас |
| Yes | Да | Ия |
| No | Нет | Жоқ |
| Home | Дом | Үй |
| Education | Образование | Білім |
| Approve a loan | Одобрить кредит | Кредитті мақұлдау |
| Income > 50000 | Доход > 50000 | Кіріс > 50000 |
| Higher | Высшее | Жоғарғы |
| Secondary | Среднее | Орта |
| Refuse | Отказать | Бас тарту |

Біздің оқытатын іріктемеміздегі белгілерге сүйене отырып, шешімдер құралы негізіндегі модель олар бойынша ол үлгі үшін класс белгісі туралы қорытынды жасайтын бірқатар мәселелерге оқытады. Өткен суретте категориялды ауыспалыларға және нақты сандарға негізделген шешімдер құралы түріндегі банктік кредитті ұсыну идеясы көрсетілген. Шешімдерді қабылдау алгоритмін пайдалану кезінде, біз құрал түбірінен бастаймыз және де деректерді белгілер бойынша бөлеміз, бұдан кейін біз түпкілікті жапырақтарға дейін жеткенше, бұл бөлу рәсімін әрбір еншілес торапта қайталай аламыз. Бұл әрбір тораптағы үлгілердің бір класқа тиесілі екенін білдіреді.

Кластерлеу бақыланбайтын оқытуды жеке кластерлерге немесе сегменттерге деректерді топтастыру үшін пайдаланады. Басқаша айтқанда, кластерлеу деректерде табиғи топтастыруды табуға тырысады.

**Деректерді басқару**

Деректерді басқару – бұл ішіне деректердің өмірлік циклының ағымдағы қажеттіліктерін қанағаттандыру үшін пайдаланылатын деректермен жұмыс істеудің барлық процестері кіреді: деректерді жинау, деректерді өңдеу, деректердің визуалдылығы, деректеді талдау және т. б.