2-TOM, 6-SON

SODDA BAYES USULI YORDAMIDA MATNLARNI TASNIFLASH

Maxmudov Zaynidin Maxamadiyevich[1], Shamiyev Muxammadi Olimovich[2]
Toshkent axborot texnologiyalari universiteti Samarqand filiali
"Axborot texnologiyalari" kafedrasi [1]dotsenti va [2]o'qituvchisi.

Annotatsiya. Maqolada axborot-qidiruv tizimi tushunchasi hamda uni yaratish usuli muhokama qilinadi va axborot-qidiruv tizimlarining soda bays usuli tasniflari keltirilgan. Tizimida axborotni izlash, qayta ishlash va saqlashning usullari tahlil qilinib, axborot-qidiruv tizimlaridan foydalanishga misollar hamda axborot izlashda tasniflashga misollar keltiriladi.

Kalit so'zlar. Axborot-qidiruv tizimlari, Sodda Bayes usuli, izlash, klassifikator.

Foydalanuvchilar qidiruv tizimiga bir yoki bir nechta so'rovlar yuborish orqali o'zlarining vaqtinchalik axborot ehtiyojlarini qondiradilar. Biroq, ko'p foydalanuvchilarning doimiy ma'lumotga ehtiyoji bor. Masalan, ular ko'p yadroli kompyuter chiplarini ishlab chiqishga oid yangiliklarni kuzatishi mumkin. Ushbu ehtiyojni qondirish uchun har kuni ertalab "ko'p yadroli AND kompyuter AND chip" so'rovidan foydalanib, yangiliklarni qidirish mumkin. Endi "Bu takroriy vazifani qanday avtomatlashtirish kerak?" degan savolni ko'rib chiqish zarur. Bunga erishish uchun ko'plab tizimlar foydalanuvchining qidiruv so'roviga obuna bo'lishini qo'llab-quvvatlaydi (doimiy so'rovlar yordamida). Doimiy so'rov ham odatda boshqa so'rovlarga o'xshaydi, ammo u muntazam ravishda yangi hujjatlar bilan yangilanib turadigan to'plamda amalga oshiriladi.

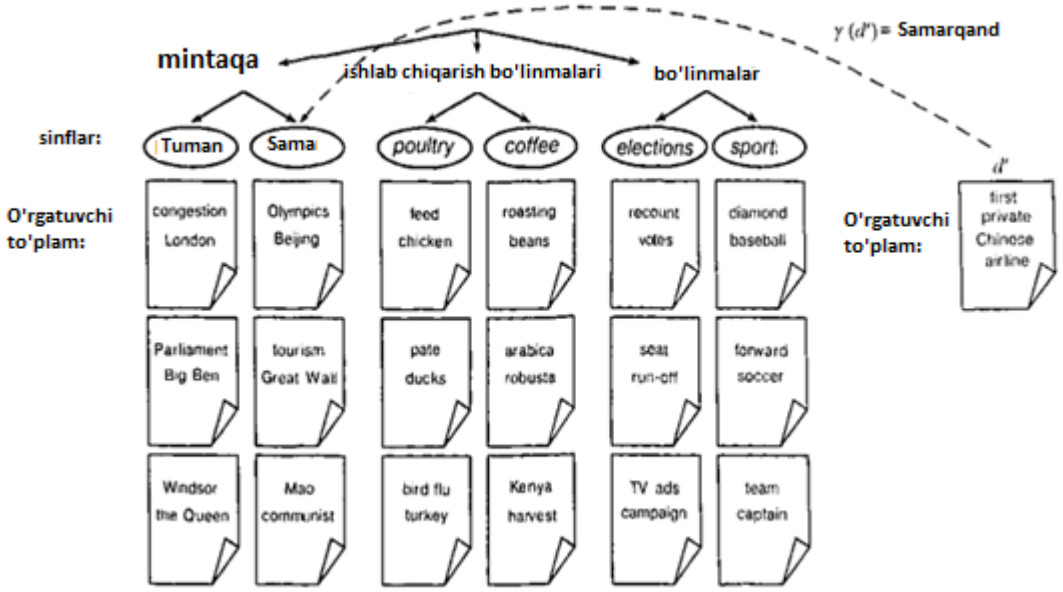
Agar doimiy so'rov "ko'p yadroli AND kompyuter AND chip" bo'lsa, foydalanuvchi "ko'p yadroli protsessorlar" kabi boshqa atamalar ishlatadigan ko'plab yangi relevant hujjatlarni o'tkazib yuborishi mumkin. Qidiruvning etarli darajada to'liqligini ta'minlash uchun doimiy so'rovlar vaqt o'tishi bilan takomillashtirilishi va shuning uchun asta-sekin murakkablashishi kerak bo'ladi. Ushbu misolda, mantiqiy qidiruv tizimida stemmingdan foydalanib natijaga erishish mumkin, yani: (ko'p yadroli OR ko'p – yadroli) AND (chip OR protsessor OR mikroprotsessor).

Umumiylik darajasini tavsiflash va doimiy so'rovlarni qayta ishlash bilan bog'liq bo'lgan mavzu sohasining chegaralarini belgilash uchun tasniflash muammosini aniqlaymiz, sinflar to'plami berilganberilgan bo'lsa, bizga berilgan ob'ekt u sinflardan qaysi biriga tegishli ekanligini aniqlash kerak bo'ladi. Ushbu misolda doimiy so'rovni qayta ishlash, tasniflash vazifasi ko'rib chiqiladi, unda zarur ma'lumotlarning ikki sinfi mavjud bo'lib,

“ko'p yadroli kompyuter chiplari” haqidagi hujjatlar va “ko'p yadroli kompyuter chiplari haqida” bo’lmagan hujjatlar. Bu masala binar tasnif deb ham ataladi. Doimiy so'rovlar asosida tasniflash, marshrutlash yoki filtrlash deb ham ataladi.

"Ko'p yadroli kompyuter chiplari" doimiy so'rovida bo'lgani kabi sinf tor ham bo'lishi shart emas. U ko'pincha "Samarqand" yoki "qahva" kabi juda keng mavzuni qamrab oladi. Bunday katta sinflar odatda mavzular deb ataladi va tasniflash masalasi matnni tasniflash, matnni turkumlashtirish, mavzuli tasniflash yoki mavzuni aniqlash ham deb ataladi. "Samarqand" sinfining namunasi 1-rasmda ko'rsatilgan.

Doimiy so'rovlar va mavzular spetsifikatsiya darajasida farqlanadi, lekin marshrutlash, filtrlash va matnni tasniflash usullari asosan bir xil bo’ladi. Shuning uchun, ushbu tasvirlangan umumiyroq matn, tasniflash masalasiga maxsus misollar sifatida marshrutlash va filtrlash masalalarini kiritdik.



1-rasm. Sinflar o’rgatuvchi to’plam va matnni tasniflash masalasida test to'plami.

Tasniflash tushunchasi juda umumiy bo'lib, axborotni qidirish sohasida ham, undan tashqarida ham ko'plab qo'llanmalarga ega. Masalan, kompyuterni ko'rish sohasida klassifikator yordamida tasvirlarni "landshaft, portret va hech biri" kabi sinflarga ajratish mumkin. Axborot izlashda tasniflash misollariga e'tibor qaratamiz.

• Indekslash uchun zarur bo'lgan ba'zi dastlabki ishlov berish bosqichlari: hujjatning kodlanishini aniqlash (ASCII, Unicode UTF-8 va boshqalar) so'z segmentatsiyasi (ikki harf orasidagi bo'shliq (probel) so'z chegarasi bo'ladimi yoki yo'qmi), so'zning haqiqiy holatini aniqlash va hujjat tili.

• Veb-spamni avtomatik aniqlash (bunday sahifalar qidiruv tizimi indeksiga kiritilmagan).

• Axloqiy bo'zuq mazmundagi kontentni avtomatik aniqlash (bunday materiallar faqat foydalanuvchi xavfsiz qidiruv opsiyasini o'chirib qo'ygan bo'lsa, qidiruv natijalariga kiritiladi).

• Film yoki mahsulot haqidagi fikr-mulohazalarni aniqlash yoki avtomatik tasniflash. Bunday ilovaga misol sifatida foydalanuvchi kamerani sotib olishdan oldin uning yashirin kamchiliklari yoki sifat muammolari yo'qligiga ishonch hosil qilish uchun salbiy sharhlarni qidiradigan qidiruvdir.

• Elektron pochtani saralash. Foydalanuvchi xabarlar, hisoblar, do'stlardan xatlar va boshqalar kabi bir nechta papkalarni yaratishi, kiruvchi elektron pochta xabarlarini tasniflashi va ularni avtomatik ravishda tegishli papkaga joylashtirishi mumkin. Uyushtirilgan xabarlar papkalarida xabarni topish juda katta kirish qutisiga qaraganda ancha oson amalga oshiriladi. Ko'pincha bu xususiyat spamni filtrlash uchun ishlatiladi.

• Tematik yoki vertikal qidiruv. Vertikal qidiruv tizimlari ma'lum bir mavzu bo'yicha qidiruvlarni cheklaydi. Masalan, "Informatika" so'roviga javoban, "Samarqand" mavzusi bo'yicha vertikal qidiruv tizimi "Informatika Samarqand" so'roviga javoban an'anaviy qidiruv tizimiga qaraganda "Samarqandda kompyuter fanlari" bo'limlarining aniqroq va to'liq ro'yxatini topa oladi.

Buning sababi shundaki, vertikal qidiruv tizimi boshqa ma'noda Samarqand atamasini o'z ichiga olgan veb-sahifalarni (masalan, Samarqand qog'ozi haqida) o'z ichiga olmaydi, lekin ularda "Samarqand" atamasi aniq ko'rsatilmagan bo'lsa ham, tegishli sahifalarni o'z ichiga oladi.

Maxsus ma'lumot qidirishda tartiblash funksiyasi tasniflagichga asoslanishi mumkin.

Ushbu ro'yxat ma'lumotni qidirish sohasida tasniflashning asosiy qo'llanilishini tavsiflaydi. Ko'pgina zamonaviy qidiruv tizimlarida u yoki bu shaklda tasniflagichlardan foydalanadigan bir nechta komponentlar mavjud.

Tasniflash kompyuterlarsiz amalga oshirilishi mumkin. Ko'pgina tasniflash muammolari an'anaviy ravishda qo'lda hal qilingan. Masalan, kutubxonachilar kitoblarga Kongress kutubxonasi toifalarini belgilaydilar. Biroq, qo'lda tasniflashni o'lchash qiyin. Misol so'rovi "Ko'p yadroli kompyuter chiplari" muqobil yondashuvni ko'rsatadi: qoida sifatida talqin qilinishi mumkin bo'lgan doimiy so'rovlar yordamida tasniflash, ko'pincha qo'lda yoziladi. Bizning misolimizda qoidalar mantiqiy ifodalarga (ko'p yadroli OR ko'p – yadroli) AND (chip OR protsessor OR mikroprotsessor) ekvivalentdir.

Qoida sinfga xos bo'lgan bir nechta kalit so'zlardan iborat. Qo'lda qoidalar yaxshi miqyosga ega, ammo vaqt o'tishi bilan ularni yaratish va saqlash tobora qiyinlashadi. Tajribali odamlar (masalan, muntazam iboralar bilan tanish bo'lgan domen mutaxassislari) avtomatik ravishda yaratilgan klassifikatorlar bilan raqobatlashadigan yoki hatto ularni yengadigan qoidalar to'plamini yaratishi mumkin, ammo bunday odamlarni topish odatda qiyin.

Qo'lda tasniflash va qo'lda yozilgan qoidalardan tashqari, matnlarni tasniflashning uchinchi yondashuvi, ya'ni mashinali o'rganishga asoslangan matn tasnifi mavjud. Mashinali o'rganishda qoidalar to'plami yoki umuman olganda qaror mezoni o'rgatuvchi to'plamdan avtomatik ravishda olinadi. Agar o'rganish usuli statistik bo'lsa, bu yondashuv statistik matn tasnifi deb ataladi. Statistik matnlarni tasniflash uchun har bir sinf uchun bir qator misollar (yoki o'quv hujjatlari) keltirish mumkin. Qo'lda tasniflash zarurati ham istisno qilinmaydi, chunki o'rgatuvchi hujjatlari ularni belgilagan shaxsdan bog'liq bo'ladi. Belgilash - bu har bir hujjatning sinfini ajratish jarayoni bo'ladi. Hujjatlarni tematik belgilash tasniflash qoidalarini tuzishdan ko'ra ancha sodda tartib bo'ladi.

Hujjatga qaragan deyarli har bir kishi bu Samarqand haqidami yoki yo'qmi, ayta oladi. Ba'zan bunday belgilar mavjud texnologik jarayonning bir qismi bo'ladi. Misol uchun, har kuni ertalab takroriy so'rovga javoban olingan yangiliklarni ko'rish va tegishli maqolalarni "ko'p yadroli protsessor" kabi maxsus papkaga joylashtirish orqali tegishli fikr-mulohazalarni bildirishingiz mumkin.

Shu jumladan, ushbu ma'ruza matnni tasniflashga umumiy holatda kirish, uning rasmiy ta'rifi bilan boshlanadi. Keyinroq biz sodda Bayes yondashuvini tasvirlaymiz, bu juda oddiy va samarali tasniflash usuli bo'ladi. Biz o'rganayotgan barcha tasniflash algoritmlarida hujjatlar yuqori o'lchamli bo'shliqlar (probellar) elementlari sifatida ko'rib chiqiladi. Ushbu algoritmlarning samaradorligini oshirish uchun odatda makonning o'lchamini kamaytirish maqsadga muvofiq bo'ladi va xususiyatni tanlash usuli qo'llaniladi.

Matinlarni tasniflash

Matnni tasniflash masalasida $d \in X$ xujjat tavsifi berilgan bo'lsin, bunda X - hujjatlar maydoni va $C=\{s_1, s_2,\ldots,s_j\}$ sinflarning belgilangan to'plami. Sinflar toifalar va teglar deb ham ataladi. Qoidaga ko'ra, X hujjat maydoni katta o'lchamga ega va sinflar "Samarqand" misolida yoki «ko'p yadroli kompyuter chiplari» haqida "Tailand" hujjatlarida bo'lgani kabi, dasturga qarab mutaxassislar tomonidan belgilanadi. Bundan tashqari, $<d,c>$ etiketli hujjatlarning D o'rgatuvchi to'plami berilgan, bu erda $<d,c> \in X*S$ Masalan, $<d,c>$ = <Samarqand jahon savdo tashkilotiga qo'shiladi, O'zbekiston> juftligi o'rgatuvchi

to'plamidir, bu erda *d* - bir jumladan iborat hujjat " Samarqand jahon savdo tashkilotiga qo'shiladi" va *c* - sinf "O'zbekiston".

O'rgatish usuli yoki o'rgatish algoritmidan foydalanib, biz hujjatlarni sinflarga ajratadigan tasniflagich yoki tasniflash funktsiyasini olamiz.

$$g:X \rightarrow C \qquad (1)$$

Bu usul nazorat ostida o'qitish deb ataladi, chunki nazoratchi (sinflarni belgilaydigan va orgatuvchi hujjatlarini tayyorlaydigan shaxs) orgatish jarayonini boshqaradigan o'qituvchi rolini o'ynaydi. O`qituvchi bilan o'rgatish usulini *G* harfi bilan belgilaymiz va *G(D)=g* deb yozamiz. *G* o'rganish usuli *D* o'rgatuvchi to'plamini kirish sifatida qabul qiladi va ***y*** tasniflash funktsiyasini hosil qiladi.

G tasniflagichi va o'rgatish usuli *G* ko'pincha ajratilmaydi. Ular "Sodda Bayes klassifikatori mukammal" deganda, sodda Bayes usulidan ko'plab o'rgatish muammolarini hal qilishda foydalanish mumkinligini anglatadi va og'ir darajada murakkab klassifikator yaratilishi mumkin emasligini bildiradi. Biroq, "Sodda Bayes usulining xatosi 20%" deganda, ma'lum bir klassifikator ***y*** (Sodda Bayes o'rgatuvchisi yordamida yaratilgan) 20% xatolikka ega bo'lgan tajribani tasvirlaymiz.

1-rasmda Reuters-RCVl to'plamidan matn tasniflash misoli ko'rsatilgan. Ushbu misolda oltita sinf (SAG, Samarqand sporti) mavjud, har birida uchta o'rgatuvchi hujjati mavjud. Rasmda har bir hujjat mazmunidan bir nechta xarakterli so'zlar ko'rsatilgan. O'rgatuvchi to'plami har bir sinf uchun odatiy misollarni o'z ichiga oladi, shuning uchun tasniflash funktsiyasini va ***y*** ni olishingiz mumkin. Natijadagi ***y*** funktsiyasini sinov to'plamiga (yoki test ma'lumotlariga), masalan, sinfi noma'lum bo'lgan birinchi xususiy Samarqand kompaniyasining yangi hujjatiga qo'llashimiz mumkin. 1-rasmda tasniflash funktsiyasi yangi hujjatni *g(d')* = Samarqand sinfiga belgilaydi, ya'ni u to'g'ri javob beradi.

Matnni tasniflash muammolaridagi sinflar ko'pincha qiziqarli tuzilishga ega bo'ladi, masalan, 1-rasmda ko'rsatilgan ierarxik tuzilishga ega bo'ladi. U sanoat mintaqasi va tematik soha uchun ikkita toifaga ega bo'ladi. Ierarxiya tasniflash muammosini hal qilishda muhim rol o'ynashi mumkin. Biz shunchaki sinflar to'plamni tashkil qiladi, deb faraz qilamiz, unda kichik to'plamlar hech qanday tarzda bir-biriga qisqartirilmaydi.

"Sodda Bays usuli" (Simple Bayesian method) tasniflash uchun foydalaniladigan oddiy va samarali usullardan biridir. Bayes teoremasiga asoslangan bu usul, statistik klassifikatsiya uchun keng qo'llaniladi, ayniqsa Naive Bayes klassifikatori orqali. Naive Bayes klassifikatori quyidagi bosqichlarda ishlaydi:

1. Ma'lumotlarni tayyorlash:

Avval ma'lumotlar to'plamini to'plang va ularni tayyorlang. Bu bosqichda ma'lumotlar tozalanishi, tegishli xususiyatlar ajratib olinishi kerak.

2. Xususiyatlar ekstraktsiyasi:

Ma'lumotlardan xususiyatlar ajratib olinadi. Xususiyatlar (features) - bu tasniflash uchun asos bo'ladigan belgilardir.

3. Mashg'ulot (Training) jarayoni:

- Trening ma'lumotlar to'plamidan foydalangan holda modelni o'rgatish.
- Har bir toifadagi (klassdagi) xususiyatlarning ehtimolini hisoblash. Naive Bayes usuli shuni nazarda tutadiki, xususiyatlar mustaqil ravishda taqsimlangan, ya'ni har bir xususiyatning qiymati boshqa xususiyatlarning qiymatlariga bog'liq emas.

1. **Tasniflash (Classification)**:

Sinov (test) ma'lumotlari uchun toifadagi ehtimollarni hisoblash va eng yuqori ehtimolli toifani tanlash.

Naive Bayes algoritmini Python orqali qanday amalga oshirish mumkinligini misol ko’rib chiqamiz:

```
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score
# Misol ma'lumot to'plami
X = np.array([[1, 20], [2, 21], [3, 22], [4, 23], [5, 24], [6, 25]])
y = np.array([0, 0, 0, 1, 1, 1])
# Ma'lumotlarni bo'lish
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=42)
# Naive Bayes modelini yaratish
model = GaussianNB()
# Modelni o'rgatish
model.fit(X_train, y_train)
# Bashorat qilish
y_pred = model.predict(X_test)
# Natijalarni baholash
accuracy = accuracy_score(y_test, y_pred)
print(f"Modelning aniqligi: {accuracy * 100:.2f}%")
# Misol ma'lumot to'plami
```

```
X = np.array([[1, 20], [2, 21], [3, 22], [4, 23], [5, 24], [6, 25]])
y = np.array([0, 0, 0, 1, 1, 1])
```

Ushbu kodda:

- Naive Bayes modelini yaratilgan va uni ma'lumotlar bo'yicha o'rgatish ko'rsatilgan.
- Keyin test to'plamida bashoratlar qilamiz va aniqlikni hisoblaymiz.

Naive Bayes algoritmi ko'pchilik tasniflash masalalarida, ayniqsa matn tasniflashda juda yaxshi natijalar beradi. Matn tasniflashda, masalan, elektron pochta spamini aniqlash yoki yangiliklar va maqolalarini turli toifalarga ajratish kabi masalalarda keng qo'llaniladi.

FOYDALANILGAN ADABIYOTLAR

1. Маннинг Кристофер Д. Введение в информационный поиск / Маннинг Кристофер Д., Рагхаван Прабхакар, Шютце Хайнрих: Пер. с англ. – М.: ООО «И.Д. Вильямс», 2011. 528 с.
2. Кнут Дональд Э. Искусство программирования. Т. 3. Сортировка и поиск / Кнут Дональд Э. – 2-е изд.: Пер. с англ. – М.: Вильямс, 2005. – 824 с.
3. Вдовин В.М., Суркова Л.Е., Валентинов В.А. Теория систем и системный анализ. – М.: Издательско-торговая корпорация «Дашков и К», 2010.
4. Бесекерский В.А, Попов Е.П., Теория систем автоматического управления, С.П., «Профессия», 2004.
5. Пугачев В.С., Синицын И.Н., Теория стохастических систем. Москва, «Логос», 2014.
6. Трояновский В.М., Информационно-управляющие системы и прикладная теория случайных процессов, Москва «Гелиос АРВ», 2014.