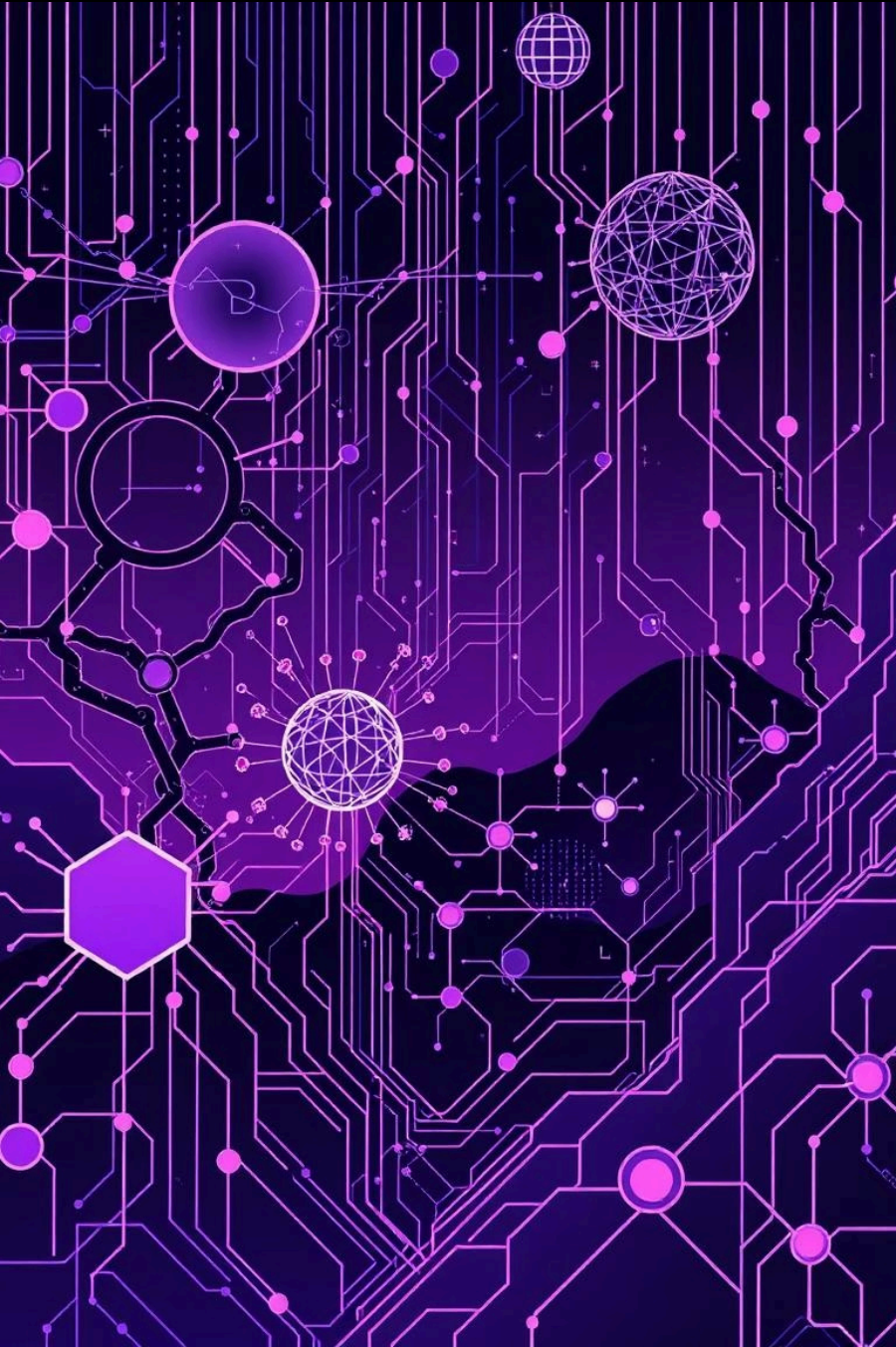




Дәріс 9. Машиналық оқыту алгоритмдері (1 бөлім)

Оқытушы: Сарсембаева Т.С.

Дәрістің мақсаты

Машиналық оқытудың (ML) негізгі ұғымдарымен, оның жасанды интеллект (AI) саласындағы орнымен және деректер ғылымындағы маңыздылығымен таныстыру.

Дәріс машиналық оқыту алгоритмдерінің модель құрудағы "жүрегі" екенін түсіндіреді. Осы негізгі ұғымдардан кейін, үш негізгі, іргелі алгоритммен таныстыру көзделеді: К-жақын көршілер (KNN), Аңғал Байес (Naïve Bayes) және Қарапайым сызықтық регрессия.

Соңында, осы әр алгоритмнің негізгі идеясы, математикалық негізі, артықшылықтары мен кемшіліктері талданатын болады.

Негізгі сұрақтар

1. Машиналық оқыту дегеніміз не және оның негізгі мақсаты қандай?
2. К-жақын көршілер (KNN) алгоритмі қалай жұмыс істейді және "k" не үшін қажет?
3. "Аңғал Байес" (Naïve Bayes) классификаторы неліктен "аңғал" деп аталады?
4. Қарапайым сызықтық регрессияның негізгі теңдеуі қандай және оның мақсаты неде?
5. Осы үш алгоритмнің негізгі айырмашылықтары мен қолдану салалары қандай?

Қысқаша мазмұн (тезистер)

Машиналық оқыту (ML)

Бұл жасанды интеллекттің (AI) маңызды саласы. Ол компьютерлерге ережелер арқылы емес, деректер үлгілеріндегі үлгілерді табу арқылы өздігінен білім алуға мүмкіндік береді. ML деректер ғылымында болжау (prediction), классификациялау (classification) және кластерлеу (clustering) үшін негізгі құрал болып табылады.

Алгоритмдер модельдің "жүрегі"

Олар мәліметтерден заңдылықтарды іздейді. Бұл дәріс үш негізгі алгоритмге арналған: KNN, Аңғал Байес және Сызықтық регрессия.

К-жақын көршілер (KNN) - бұл бақыланатын оқыту (supervised learning) алгоритмі. Негізгі идеясы – жаңа объектінің класы оның ең жақын "k" көршілерінің класына (көпшілігіне) негізделіп анықталады. Ол объектілер арасындағы қашықтықты (мысалы, Евклидтік) есептейді.

Аңғал Байес (Naïve Bayes) - бұл ықтималдыққа негізделген жіктеу алгоритмі. Ол Байес теоремасын қолдана отырып, объектінің белгілі бір класқа жату ықтималдығын есептейді. "Аңғал" деп аталуының себебі – ол әрбір белгінің (feature) басқа белгілерден тәуелсіз екенін қабылдайды.

Қарапайым сызықтық регрессия - бұл тәуелді айнымалы (y) мен тәуелсіз айнымалы (x) арасындағы сызықтық қатынасты модельдейтін болжамды талдау әдісі. Модель теңдеуі: $y = w \cdot x + b$. Мақсаты – қателік функциясын (мысалы, MSE) азайтатын ең жақсы w (салмақ) және b (ығысу) мәндерін табу.

Қорытындылайтын болсақ, KNN ұқсастыққа, Naïve Bayes ықтималдық заңына, ал Сызықтық регрессия сызықтық қатынасты модельдеуге негізделген.

Кіріспе

Машиналық оқыту (Machine learning) - бұл компьютерлік жүйелерге нақты нұсқаулар бермей, тәжірибе мен мәліметтер негізінде өздігінен білім алуға және жетілдіруге мүмкіндік беретін жасанды интеллекттің (Artificial Intelligence) маңызды саласы.

Яғни, машина белгілі бір тапсырманы ережелер арқылы емес, деректер үлгілеріндегі үлгілерді табу арқылы орындауды үйренеді.

Машиналық оқыту қазіргі уақытта деректер ғылымы, робототехника, медицина, қаржы, маркетинг, киберқауіпсіздік және басқа да көптеген салаларда шешім қабылдаудың негізгі құралына айналды. Ол үлкен деректерді (Big Data) талдау, болжау (prediction), классификациялау (classification), кластерлеу (clustering) және үлгіні тану (pattern recognition) сияқты тапсырмаларды орындайды.

Машиналық оқытудың негізгі мақсаты - модель құру, яғни нақты әлемдегі құбылыстарды математикалық сипаттай алатын функцияны табу. Бұл модель жаңа деректер келген сайын болжамын жақсарту арқылы тәжірибеден сабақ алады.

Машиналық оқытуда алгоритмдер модельдің "жүрегі" болып табылады. Олар мәліметтерден заңдылықтарды іздейді, функцияларды реттейді (түзу, қисық, ықтималдық шекарасы және т.б.), және жаңа мысалдар үшін дұрыс шешім қабылдау үшін модельді бейімдеңіз.

Әрбір алгоритмнің өзіндік ерекшеліктері бар: олардың кейбіреулері категориялық белгілерді болжауға бағытталған (мысалы, KNN немесе Naïve Bayes), кейбіреулер сандық мәндерді есептейді (мысалы, сызықтық регрессия), басқалары күрделі көпөлшемді модельдерді түсіндіруге арналған (мысалы, нейрондық желілер, шешім ағаштары).

Бұл дәрісте біз машиналық оқытудың үш негізгі алгоритмін талдаймыз. Бұл әдістер күрделі модельдердің (нейрондық желілер, SVM, кездейсоқ орман) теориялық негізін құрайды және деректер ғылымы бағытындағы алғашқы қадам ретінде маңызды.

Зерттелетін Алгоритмдер

К-жақын көршілер

K-Nearest Neighbors, KNN

Аңғал Байес

Naïve Bayes

**Қарапайым сызықтық
регрессия**

Simple Linear Regression

К жақын көршілер

K-near neighbors алгоритмі бақыланатын оқыту(supervised learning) түріне жатады және жіктеу және регрессия мәселелерін шешу үшін қолданылады.

Негізгі идеясы - жаңа объектінің класы оның ең жақын көршілерінің класы негізінде анықталады.

Математикалық негіз

KNN объектілер арасындағы қашықтықты есептейді. Ең жиі қолданылатын метрика-Евклидтік қашықтық:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

x, y — объектілердің белгі векторлары; n - белгілердің саны.

Алгоритм қадамдары

- Көршілер саны k таңдалады.
- Берілген нүкте мен барлық деректер арасындағы қашықтық есептеледі.
- Ең жақын k көршілері анықталды.
- Көршілердің көпшілігі қай сыныпқа жатады-жаңа нысан бір сыныпқа жатады.

Python мысалы

```
from sklearn.neighbors import KNeighborsClassifier

model = KNeighborsClassifier(n_neighbors=5)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```


Аңғал Байес Классификаторы

Naïve Bayes-ықтималдыққа негізделген жіктеу алгоритмі. Ол байес теоремасын қолдана отырып, объектінің белгілі бір класқа жату ықтималдығын есептейді.

Байес формуласы

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

$P(C|X)$ - x белгілері бойынша сыныптың ықтималдығы, $P(X|C)$ - класс белгілі болған кезде белгілердің ықтималдығы, $P(C)$ — Класс ықтималдығы, $P(X)$ - барлық белгілердің ықтималдығы.

Оның "аңғал"деп аталуының себебі: бұл модель әрбір белгінің тәуелсіз екенін аңғал түрде (naively) қабылдайды. Мысалы, "электрондық пошта спамы немесе жоқ" дегенде, ол әр сөздің басқа сөздерден тәуелсіз болу ықтималдығын есептейді.

Түрлері

- Gaussian Naïve Bayes — үздіксіз (сандық) деректерге
- Multinomial Naïve Bayes — мәтін және санау типіндегі белгілерге
- Bernoulli Naïve Bayes — бинарлы белгілерге (иә/жоқ, бар/жоқ)

Python мысалы

```
from sklearn.naive_bayes import GaussianNB

model = GaussianNB()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```


Қарапайым сызықтық регрессия

Сызықтық регрессия - тәуелді айнымалыны қолданатын болжамды талдау әдісі ($\hat{y}$) тәуелсіз айнымалы (x) арасындағы сызықтық қатынасты модельдейді.

Модель теңдеуі

$$\hat{y} = w \cdot x + b$$

$\hat{y}$ – болжамды мән,

w -салмақ (сызық қатынасы, градиент),

b - ығысу (intercept).

Мақсат-қателік функциясын азайту

$$L = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Модельді оқыту процесінде градиенттің түсуі қолданылады.

Python мысалы

```
from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

Салыстырмалы кестесі

Алгоритм	Түрі	Негізгі идея	Артықшылығы	Кемшілігі
KNN	Классификация, регрессия	Ең жақын көршілердің көпшілігі бойынша шешім	Түсінікті, қарапайым	Үлкен деректерде баяу
Naïve Bayes	Классификация	Байес ықтималдығы және тәуелсіздік болжамы	Жылдам, аз дерекке тиімді	Белгілер тәуелсіз болмауы мүмкін
Linear Regression	Регрессия	Сызықтық байланыс	Нақты түсіндіру	Тек сызықтық байланысқа жарайды

Бұл үш алгоритм Деректер ғылымы саласындағы ең негізгі, іргелі әдістердің бірі болып табылады. Олар күрделі модельдерге негіз болатын негізгі ұғымдарды үйретеді (Шешім Ағашы, Кездейсоқ Орман, SVM, Нейрондық Желілер).

- **KNN**
Көршілер ұқсастыққа сүйенеді
- **Naïve Bayes**
Ықтималдық заңына негізделген
- **Сызықтық Регрессия**
Сызықтық қатынасты модельдейді

Материалды меңгеруді бақылау сұрақтары

1. Машиналық оқыту дегеніміз не және оның дәстүрлі бағдарламалаудан айырмашылығы неде?
2. KNN алгоритмінде "k" параметрін таңдау неліктен маңызды?
3. "Аңғал Байес" алгоритмі неліктен "аңғал" деп аталады? Бұл "аңғалдық" оның дәлдігіне қалай әсер етеді?
4. Сызықтық регрессияның негізгі теңдеуін түсіндіріңіз. ω (салмақ) және b (ығысу) нені білдіреді?
5. KNN, Аңғал Байес және Сызықтық регрессияның қайсысы классификацияға, қайсысы регрессияға жатады?

Әдебиеттер тізімі

1. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. <https://www.deeplearningbook.org>
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer. <https://hastie.su.domains/ElemStatLearn/>
3. Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing and Decision Making. SN Computer Science, Springer.
4. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>