Тақырыбы: Классификация әдістері: Жалпы түсінік, Алгоритмдер және Модельдерді Бағалау

Классификация әдістері: Жалпы түсінік және алгоритмдер

Классификация дегеніміз не?

Классификация (Classification) — бұл бақылауларды белгілі бір категорияларға немесе сыныптарға жатқызу үшін қолданылатын бақыланатын машиналық оқытудың (Supervised Machine Learning) негізгі міндеті. Классификациялау моделі деректердің ерекшеліктері (features) мен олардың сәйкес сыныптық жапсырмалары (class labels) арасындағы байланысты үйренеді.

- Мысалдар:

– Денсаулық сақтау: Пациенттің медициналық сынақ нәтижелеріне сүйене отырып, оның ауруға шалдығу ықтималдылығын ("иә" немесе "жоқ") болжау.

– Қаржы: Несие алушыны оның қаржылық тарихына сүйене отырып, "төлем қабілетті" (тәуекел төмен) немесе "төлем қабілетсіз" (тәуекел жоғары) деп жіктеу.

– Маркетинг: Тұтынушының мінез-құлқына байланысты оның белгілі бір өнімді сатып алуы ("сатып алады" немесе "сатып алмайды") туралы болжам жасау.

Классификацияның негізгі мақсаты — оқыту деректері негізінде, жаңа, бұрын көрілмеген мәліметтерді алдын ала дайындалған модель негізінде дұрыс сыныпқа жатқызу үшін жалпыланған ережелерді құру.

Классификациялық модельдер шығарылатын сыныптар санына байланысты екіге бөлінеді:

1. Екілік классификация (Binary Classification): Екі ғана мүмкін сынып бар (мысалы, 0 немесе 1, Иә немесе Жоқ).
2. Көп сыныпты классификация (Multi-class Classification): Екі немесе одан да көп сынып бар (мысалы, "кіші", "орташа", "үлкен").

Логистикалық регрессия (Logistic Regression)

Логистикалық регрессия — сызықтық регрессияның кеңейтілген түрі. Атына қарамастан, бұл — классификация алгоритмі. Ол екілік классификация мәселелерін шешуге арналған және нәтижені 0 мен 1 аралығындағы ықтималдық мәні ретінде ұсынады.

Логистикалық Функция (Сигмоид)

Логистикалық регрессияның негізгі принципі деректердің сызықтық комбинациясын z ықтималдыққа түрлендіретін логистикалық (сигмоидты) функцияны қолдану болып табылады. Сигмоид функциясы кез келген нақты санды қабылдайды және нәтижені [0, 1] диапазонына салады , бұл оны ықтималдық ретінде интерпретациялауға мүмкіндік береді.

Логистикалық регрессияның формуласы:

$$p(y=1|X) = \frac{1}{1+e^{-z}}$$ мұндағы:

- p(y=1 | X) — ықтималдық, яғни жаңа бақылаудың y = 1 сыныбына жату ықтималдылығы.
- X— деректердің ерекшеліктері (факторлар) жиынтығы.
- e — натуралды логарифмнің негізі (≈ 2.718).
- z — сызықтық комбинация (сызықтық регрессияның нәтижесі):

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

мұндағы beta_i — модельдің салмақтары (коэффициенттері).

Шешім Қабылдау

Сигмоид функциясы арқылы алынған ықтималдық p шешім қабылдау үшін қолданылады. Бұл үшін арнайы шешім табалдырығы (threshold) белгіленеді.

- Егер p g $\geq$ 0.5$ болса, класс y=1деп анықталады.
- Егер p < 0.5 болса, класс y=0 деп анықталады.

Логистикалық регрессияның қолданылуы

- Бинарлық классификация: Ең негізгі қолданылу аясы. (Спамды анықтау, ауруды диагностикалау).
- Көп сыныпты классификация (Multinomial Logistic Regression): Екі сыныптан көп болған жағдайда қолданылады. Бұл үшін көбінесе "Бірге қарсы қалғандары" (One-vs-Rest) стратегиясы немесе Softmax функциясы қолданылады. Softmax бір мезгілде барлық сыныптардың ықтималдығын есептейді, олардың қосындысы 1-ге тең болады.

k-NN (k-жақын көрші - k-Nearest Neighbors)

k-NN — ленив (lazy) оқытуға негізделген, қарапайым, бірақ өте тиімді классификация алгоритмі. Ол параметрикалық емес (non-parametric) модель болып табылады, яғни деректердің таралуы туралы ешқандай болжам жасамайды.

k-NN әдісі жаңа нысанның классын оның оқыту деректер жиынындағы ең жақын k көршісіне қарап болжайды.

1. Қашықтықты есептеу: Жаңа бақылау x_new және оқыту жиынтығындағы әрбір нүкте x_i арасындағы қашықтықты есептеу. Ең жиі қолданылатын қашықтық өлшемі — Евклидтік қашықтық (Euclidean Distance):

$$d(x_a, x_b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

2. Көршілерді таңдау: Есептелген қашықтықтар негізінде ең жақын k нүктені (көршіні) анықтау.
3. Классификация: Таңдалған k көршілердің ішіндегі ең көпшілігінің сыныбы жаңа бақылауға тағайындалады (дауыс беру принципі).

k-NN алгоритмінің формуласы (Дауыс беру):

$$y_{\text{new}} = \text{mode}(\{y_{k1}, y_{k2}, \ldots, y_{kk}\})$$

мұндағы:

- y_new — жаңа бақылаудың болжанған сыныбы.
- mode— ең жиі кездесетін мән (яғни, көпшілік дауыс).
- y_k1, y_kk — ең жақын k көршілердің сыныптары.

4.3.2. k мәнін таңдау

k параметрін дұрыс таңдау модельдің тиімділігі үшін өте маңызды:

- Кіші k (мысалы, k=1): Модель өте күрделі болады және шуға сезімтал, нәтижесінде артық оқыту (overfitting) қаупі жоғары.
- Үлкен k: Модель өте тегіс болады және қарапайым қателіктерге әкелуі мүмкін (underfitting), себебі алыс нүктелер де ескеріледі.

Әдетте, k мәнін оңтайландыру үшін кросс-валидация қолданылады.

Артықшылықтары мен Кемшіліктері

Артықшылықтар	Кемшіліктер
Оңай түсініледі: Қарапайым логикаға негізделген.	Есептеу шығыны жоғары: Әрбір жаңа нүкте үшін барлық оқыту деректерімен қашықтықты есептеу керек.
Жалқау оқыту (Lazy Learning): Оқыту кезеңі іс жүзінде жоқ.	Деректер көлеміне тәуелді: Үлкен деректер жиынында жұмысы баяулайды.
Параметрикалық емес: Деректердің таралуы туралы болжамдар жоқ.	Өлшемділіктің қарғысы (Curse of Dimensionality): Атрибуттар саны артқан сайын, модельдің дәлдігі төмендейді.

Шешім ағашы (Decision Trees)

Шешімдер ағашы — мәліметтерді әртүрлі белгілер (факторлар) бойынша бөлшектеу арқылы классификация жасау әдісі. Бұл құрылымдық алгоритм, адамның шешім қабылдау логикасына ұқсас.

Құрылымы және Принципі

Шешімдер ағашының құрылымы:

- Түбір түйін (Root Node): Бастапқы деректер жиынтығын көрсетеді.
- Ішкі түйіндер (Internal Nodes): Белгілі бір қасиеттер бойынша деректерді бөлу шартын (сұрақ) қамтиды.
- Жапырақ түйіндер (Leaf Nodes): Деректердің соңғы сыныптарын көрсетеді.

Ағашты Құру Алгоритмі (CART)

Шешімдер ағашын құрудың ең кең таралған алгоритмі — CART (Classification and Regression Trees). Оның негізгі мақсаты — әрбір түйінде деректерді таза сыныптарға (бір сынып басым болатын) бөлетін ең жақсы ерекшелікті табу.

- Түпкі шарт: Алдымен, барлық деректерді бір түбірге жинайды.
- Ең жақсы бөлу: Әрбір ішкі түйінде деректерді ең жақсы бөлетін белгіні таңдайды. Бұл таңдау үшін әдетте Джини индексі (Gini Index) немесе Ақпараттық өсім (Information Gain/Entropy) сияқты өлшемдер қолданылады.
    - Джини индексі: Топтағы сыныптардың араласу дәрежесін өлшейді. Ең кіші Джини индексі бар бөлу ең жақсы болып саналады.
- Қайталау: Әрбір бөлуден кейін деректер топтары кішігірім топтарға бөлінеді, бұл процесс тоқтау критерийіне (мысалы, түйіндегі деректер саны тым аз болса немесе топ толығымен таза болса) жеткенше қайталанады.

Артықшылықтары мен Кемшіліктері

Артықшылықтар	Кемшіліктер
Интерпретациясы жеңіл: Ағаштың құрылымы адамға түсінікті.	Overfitting қаупі: Ағаштың тереңдігі тым көп болса, оқыту деректеріне өте жақсы бейімделіп, жаңа деректерде нашар жұмыс істейді.
Масштабтылық: Үлкен деректер жиындарымен жұмыс істей алады.	Тұрақсыздық (Instability): Оқыту деректеріндегі кішігірім өзгерістер ағаштың құрылымын түбегейлі өзгертіп жіберуі мүмкін.

Артықшылықтар	Кемшіліктер
Бірнеше сыныпты классификациялау: Табиғатынан көп сыныпты мәселелерді шешуге қабілетті.	Оңтайлы ағашты табу: Есептеу тұрғысынан өте күрделі.

Классификациядағы қателерді бағалау (Model Evaluation)

Классификациялық модельдердің тиімділігін бағалау үшін әртүрлі метрикалар қолданылады. Бұл метрикалар Қателік Матрицасы (Confusion Matrix) негізінде есептеледі.

Қателік Матрицасының Элементтері:

- TP (True Positive): Модель y=1 деп болжады, шын мәнінде де y=1. (Дұрыс оң)
- TN (True Negative): Модель y=0 деп болжады, шын мәнінде де y=0. (Дұрыс теріс)
- FP (False Positive): Модель y=1 деп болжады, шын мәнінде y=0. (Type I қате - Жалған оң)
- FN (False Negative): Модель y=0 деп болжады, шын мәнінде y=1. (Type II қате - Жалған теріс)

Дәлдік (Accuracy)

Дәлдік — модельдің барлық болжамдарының ішіндегі қаншасының дұрыс болғанын көрсететін ең қарапайым және жиі қолданылатын метрика.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{Дұрыс болжамдар саны}}{\text{Барлық болжамдар саны}}$$

Нақтылық (Precision)

Нақтылық (немесе Дәлдік) — модель оң деп болжағандардың ішіндегі шын мәнінде қаншасының оң болғанын көрсетеді. Жалған оң (FP) санын азайту маңызды болғанда қолданылады (мысалы, спам емес поштаны спам деп белгілеу қаупін азайту).

$$\text{Precision} = \frac{TP}{TP + FP}$$

Толықтық (Recall/Sensitivity)

Толықтық (немесе Сезімталдық) — барлық шын оң сыныптардың ішінде модельдің қаншасын дұрыс анықтағанын көрсетеді. Жалған теріс (FN) санын азайту маңызды болғанда қолданылады (мысалы, ауруды анықтауда ауру адамды сау деп жібермеу).

$$\text{Recall} = \frac{TP}{TP + FN}$$

F-мера (F-Measure немесе F1-Score)

F-мера — Нақтылық (Precision) пен Қамтудың (Recall) арасындағы теңгерімді көрсететін гармоникалық орта. Егер сыныптар тең емес (Imbalanced Classes) болса, бұл метрика өте пайдалы.

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Қорытынды

Классификация әдістері деректерді талдау және модельдеуде маңызды орын алады. Логистикалық регрессия, k-NN және шешімдер ағашы сияқты негізгі алгоритмдердің әрқайсысының өзіндік артықшылықтары мен кемшіліктері бар, және оларды таңдау деректердің түріне, көлеміне және проблеманың сипатына байланысты болады.

Сонымен қатар, модельдердің дәлдігін бағалау үшін қолданылатын Дәлдік, Нақтылық, Толықтық және F-мера сияқты метрикалар ғылыми және практикалық жұмыстарда дұрыс және негізделген шешімдер қабылдауға мүмкіндік береді, әсіресе сыныптар тең емес жағдайда.

Қолданылған әдебиеттер тізімі

1. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques (3rd ed.)*. Elsevier. (Классификация алгоритмдерінің негіздері)
2. Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer. (Бағалау метрикаларының егжей-тегжейлі талдауы)
3. Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill. (Машиналық оқыту және шешімдер ағашының теориялық негіздері)
4. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. (Логистикалық регрессияның математикалық сипаттамасы)
5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)*. Springer. (Классификацияның статистикалық негізі)