Бақыт Жансая Втипо -35
8-билет

1. Тихонов регуляризациясы
2. K-жақын көршілер әдісі
3. 9- есеп

1. Тихонов регуляризациясы

Тихонов регуляризациясы, сондай-ақ Ridge Regression немесе L2 регуляризациясы деп аталады, коэффициенттердің квадраттарына штраф қосу арқылы сызықтық регрессиядағы модельдің күрделілігін төмендету әдісі.
Бұл әдіс модельдің тұрақтылығы мен дәлдігін жақсартады, әсіресе белгілер корреляцияланған кезде немесе бақылаулар белгілерден аз болған кезде.

Ридж регрессиясы – тәуелсіз айнымалылар жоғары корреляцияланған сценарийлердегі көп регрессия модельдерінің коэффициенттерін бағалау әдісі. Ол эконометрика, химия және инженерия сияқты көптеген салаларда қолданылған. Сондай-ақ, Андрей Тихонов атындағы Тихонов регуляризациясы ретінде белгілі, бұл нашар қойылған мәселелерді реттеу әдісі.

Теорияны алғаш рет 1970 жылы Хоэрл мен Кеннард «Ридж регрессиялары: ортогональды емес есептерді объективті бағалау» және «Ридж регрессиялары: ортогональды емес есептердегі қолданбалар» атты Technometrics мақалаларында енгізді.

Ридж регрессиясы сызықтық регрессия модельдерінде кейбір мультиколлинеарлық (жоғары корреляциялық) тәуелсіз айнымалылар болған кезде, жотаның регрессия бағалаушысын (RR) жасау арқылы ең аз квадратты бағалаушылардың дәл еместігіне ықтимал шешім ретінде әзірленді.

Артықшылықтары:
- Регуляризация модельді қайта оқытудан қорғайды, әсіресе бақылаулар санымен салыстырғанда белгілер саны көп болған жағдайда.
- Реттелетін Модель әдетте жаңа деректерде жақсы жалпылау қабілетіне ие.
- Көп өлшемді регрессияда қолданылады, стандартты әдістер мультиколлинеарлықтан зардап шегуі мүмкін.

Ол жұмыс істейтін Принцип келесідей: мүмкін болатын шешімдердің ішінен жоғары нормалары бар шешімдерді алып тастау арқылы оңтайландыру мәселесіндегі шешімдер аймағын шектеу. Мысалы, сызықтық регрессиялық модельде бұл көбінесе Шарттың|\beta|^2 шарттарының қосымша мәні ретінде пайда болады. (бета коэффициенттерінің квадраттық нормасы), бұл жалпы шешімді тұрақтандырады және асыра сілтеуді болдырмауға көмектеседі.
Осы формулалар

$$S = (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}).$$

$$(X^TX)\mathbf{b} = X^T\mathbf{y}.$$

2. K-жақын көршілер әдісі

Nearest Neighbors - жіктеу үшін де, регрессия үшін де қолданылатын параметрлік емес алгоритм.

kNN объектінің жіктелуі оның көршілерінің көпшілік дауысымен анықталады, объектісі оның ең жақын көршілерінің арасында жиі кездесетін сыныпқа тағайындалады.
Объект k арасында ең көп таралған класс беріледі.
Регрессия жағдайында болжамды мән жақын көршілердің k мәндерінің орташа немесе медианалық мәні болып табылады. Нысандар арасындағы қашықтық әдетте евклидтік қашықтықты пайдаланып өлшенеді, дегенмен Манхэттен қашықтығы сияқты басқа көрсеткіштерді де қолдануға болады.
Артықшылығы қарапайымдылығы k-NN іске асыру және түсіну оңай, кеңінен қолданылады.

KNN-болжаудың қарапайым әдістерінің бірі: сәйкес келетін модель жоқ. Бұл KNN пайдалану Автоматты процедура дегенді білдірмейді. Болжау нәтижелері белгілердің қалай талданғанына, ұқсастықтың қалай өлшенгеніне және K шамасы қандай болатынына байланысты. Сонымен қатар, барлық болжаушылар сандық түрде болуы керек. Біз осы әдістің жұмысын жіктеу мысалымен түсіндіруге болады.

Жақын көршілердің K әдісі (KNN, ағылш. k-nearest neighbors) өте қарапайым идея жіктелетін немесе болжанатын әрбір жазба үшін:
1. K-ға ұқсас белгілері бар жазбаларды табу .
2. Жіктеу үшін: осы ұқсас жазбалардың арасынан мажоритарлық сыныпты анықтап, осы сыныпты жаңа жазбаға тағайындау.
3. Болжау үшін (KNN регрессиясы деп те аталады): осы ұқсас жазбалардың арасынан орташа мәнді тауып, жаңа жазбаның орташа мәнін болжау.
Көрші -болжалды мәндері басқа жазбаға ұқсас жазба.
- Қашықтықтың метрикалық көрсеткіштері - бір жазбаның екіншісінен қаншалықты алыс екенін бір санмен қорытындылайтын метрикалық көрсеткіштер.
- Стандарттау - орташа мәнді алып тастап, стандартты ауытқуға бөлу.
- K - Жақын көршілердің алгоритмін есептеу кезінде ескерілетін көршілердің саны.

K жақын көршісі кластеризация түрі болып табылады . Және есептеп анық дәл шығару үшін метрика қолданады .
Nearest Neighbors жақын көршілерде осы ең маңызды танымал метрикаларды қолданады.

Евклид $\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$

Манхэттон $+ |p_n - q_n|$

Чебышева $= \max(|p_1 - q_1|$

3. 9- есеп

9. Предположим, у вас есть набор данных, содержащий информацию о пассажирах титаника, которые выжили или не выжили в катастрофе. Вам нужно разработать модель классификации, которая бы предсказывала, выживет ли пассажир в

катастрофе, основываясь на различных признаках, таких как пол, возраст, класс каюты, наличие родственников на борту и т.д. Задача может состоять из следующих шагов: 1. Предобработка данных: перед использованием данных в модели классификации, необходимо выполнить предобработку данных, такую как заполнение пропущенных значений, кодирование категориальных признаков и т.д. 2. Разбиение данных на обучающую и тестовую выборки: данные должны быть разделены на обучающую и тестовую выборки для оценки производительности модели. 3. Обучение модели классификации: вы можете использовать любую модель классификации, такую как логистическую регрессию, решающее дерево, случайный лес и т.д. 4. Оценка производительности модели: после обучения модели на обучающих данных, вы должны оценить ее производительность на тестовых данных, используя различные метрики, такие как метрика Чебышева, Манхэттен, Евклида. Задача может заключаться в том, чтобы реализовать модель классификации и применить ее к набору данных, чтобы предсказать, выживет ли пассажир в катастрофе на основе его характеристик. После этого студентам нужно будет проанализировать результаты и оценить производительность модели, используя различные метрики, а также обсудить возможные улучшения модели, такие как использование других методов классификации или улучшение предобработки данных.

https://colab.research.google.com/drive/1tZfpLdq4ZZFA8oFYCcRjQ3FzpB60rOrX?usp=sharing