Синтез и отбор признаков
Синтез и отбор объектов

Лекция 10

Отбор признаков

Почему количество признаков иногда нужно уменьшать?

1. Экономия времени и памяти
2. Когда много признаков труднее найти закономерность
3. Между нецелевыми признаками могут существовать зависимости – многие модели предсказания в этой ситуации работают плохо.

Какие признаки – кандидаты на удаление?

Допустим, у нас есть таблица с нецелевыми признаками X1, X2,… Xn и целевым признаком Y.

Какие из признаков X1, X2,… Xn можно удалить?

Ваши предложения?

Какие признаки – кандидаты на удаление?

Следующие признаки рекомендуется удалять

1. Признаки с большим числом пропусков и косяков данных.
2. Числовые признаки с ОЧЕНЬ малым отклонением (в частности все константные признаки).
3. Если между признаками X1, X2 очень высокая корреляция, то один из них можно удалить.
4. Можно удалить признак X, если его корреляция с Y близка к 0 (тут надо аккуратнее).
5. Для признаков Xi вычислить их информативность (энтропию, неопределенность Джини…) и удалить признаки с наихудшими показателями.

Какие признаки – кандидаты на удаление?

6. Запустить модель предсказания, которая (помимо своей основной работы) умеет определять значимость каждого из признаков. К таким моделям относится, например, линейные модели (в том числе линейная регрессия, логист. регрессия, их регуляризации и лассо).
Значимость каждого признака у линейной модели – это…

Какие признаки – кандидаты на удаление?

6. Запустить модель предсказания, которая (помимо своей основной работы) умеет определять значимость каждого из признаков. К таким моделям относится, например, линейные модели (в том числе линейная регрессия, логист. регрессия, их регуляризации и лассо).
Значимость каждого признака у линейной модели – это коэффициент при этом признаке.

$$Y=1.5X_1+0.01X_2-2X_3+10$$

Отбор признаков в несколько итераций

Можно перебрать все подмножества признаков, для каждого подмножества построить модель предсказания. Выбрать подмножество с наилучшим качеством предсказания.

Но это очень трудоемко.

Отбор признаков в несколько итераций

Можно так: Фиксируем небольшое число N, перебираем все комбинации по N признаков, выбираем лучшую комбинацию, потом перебираем комбинации из N+1 признаков так, что предыдущая лучшая комбинация признаков зафиксирована, а перебирается только новый признак. Таким образом можно перебирать, пока не упремся в максимально допустимое число признаков или пока качество модели не перестанет значимо расти.

Отбор признаков в несколько итераций

Последний алгоритм можно развернуть: начинать с полного пространства признаков и выкидывать признаки по одному, пока это не портит качество модели или пока не достигнуто желаемое число признаков.

Синтез новых признаков

Зачем это делать?

1. Из нескольких плохих признаков можно состряпать один хороший.
2. Улучшение работы моделей МО.

Методы получения новых признаков

1. Нормализация (приведение признаков к одному масштабу). Без этого метрические методы МО работают плохо.
2. Логарифмирование. Для борьбы с большими числами и получения нормального распределения значений признака.
3. Житейская логика. Например, если мы предсказываем анорексию у девушек из Playboy, то значимым тут признаком является «индекс массы тела», а не «рост» и «вес» по отдельности.

Методы получения новых признаков

Выделение признаков для картинок, текстов, видео – это отдельная тема.

Про преобразование категориальных признаков см. [2]

Синтез новых объектов

Зачем это делать?

Если тренировочная выборка объектов несбалансирована (то есть доля объектов одного класса гораздо больше доли объектов второго класса), то могут возникнуть проблемы.

Например, такая: алгоритм предсказания просто забудет про меньший класс и все объекты будет относить к большему классу.

С этим нужно что-то делать!!!

Методы балансировки выборки

1. Удаление выбросов – они тоже мешают работе алгоритмов предсказания.
2. Undersampling – удаление объектов большего класса. Объекты большего класса можно кластеризовать, а потом из каждого кластера оставить лишь эталонные объекты (объекты из середины кластера).
3. Oversampling – размножение объектов меньшего класса.
4. Создание синтетических объектов (см. след. слайд)

Синтетические объекты (SMOTE-алгоритм)

По паре объектов A,B можно построить синтетический объект как их линейную комбинацию aA+(1-a)B, где a – случайное число из отрезка [0,1].
Например, при a=0.1 объекты

Объект	Рост	Вес	Пол (Y)
A	200	100	1
B	150	50	0

Дают синтетический объект

Объект	Рост	Вес	Пол (Y)
C	155	55	0.1

Проблемка: категориальные признаки

Категориальные признаки при вычислении лин. комбинации могут потерять смысл. Например, пол=0.1
Возможные пути решения:
1.

Объект	Рост	Вес	Пол (Y)
C	155	55	0.1

Проблемка: категориальные признаки

Категориальные признаки при вычислении лин. комбинации могут потерять смысл. Например, пол=0.1

Возможные пути решения:

1. Округлить до ближайшего допустимого значения.
2. Провести случайное испытание в соответствии с полученной вероятностью.
3. Сделать признак числовым. Если это делается для целевого признака, то задача классификации превращается в задачу регрессии.

Объект	Рост	Вес	Пол (Y)
C	155	55	0.1

Выводы

Синтез (отбор) признаков и объектов – не панацея, так как...

На новом множестве признаков (тренировочных объектов) качество предсказания не обязательно возрастет. Оно даже может ухудшиться ((((

Короче, нужно пробовать.

Использованная литература

1. https://habrahabr.ru/company/ods/blog/325422/
2. https://alexanderdyakonov.wordpress.com/2016/08/03/python-категориальные-признаки/
3. https://ru.wikipedia.org/wiki/Логнормальное_распределение
4. https://habrahabr.ru/post/264915/ (про энтропию при отборе фич)
5. https://habrahabr.ru/post/270367/ (зачем распределения признаков делать нормальными)