Анализ данных. Лекции. Неделя 1
Обучение с учителем. Метрики. Генерация признаков.

Лектор: Степан Зимин
Конспект: Утешева Дарья

2022

Contents

1 Постановка задачи:

Пусть нам дано множество $\{x_1, x_2, ..., x_n\} \in X$ - объектов обучющей выборки. И множество $\{y_1, y_2, ..., yn\} = \{y(x_1), y(x_2), ..., y(x_n)\} \in Y$ - искомых ответов.
Задача: Найти $a : X \to Y$ - алгоритм a, который приближает y на всём множестве X.
Объект обучающей выборки: $x_i = (x^1, x^2, ..., x^m)$ - вектор признаков объекта x_i, где x^j - j-й признак объекта x_i.

Обучающая выборка: $\begin{pmatrix} x_1^1, & \cdots & x_1^m \\ \vdots & \ddots & \vdots \\ x_n^1 & \cdots & x_n^m \end{pmatrix}$

2 Типы признаков объектов:

В теории:

- Бинарный признак(0,1, мужчина или женщина)
- Номинальный признак(нельзя упорядочить, например, времена года)
- Порядковый признак(можем упорядочить, например, уровень образования)
- количественный признак(возраст, зарплата)

На практике:

- Категориальные(значит, конечное возможное число элементов):
  - Бинарный признак
  - Номинальный признак
  - Порядковый признак
- Количественные(значит, имеет смысл взятие среднего):
  - Числа
  - Бинарный признак
  - Порядковый признак

3 Виды ответов:

- $y_i \in \{0, 1\}$ - бинарный ответ()
- $y_i \in \{C_1, C_2, ..., C_k\}$- ответ принадлежит одному из k классов(оценка товара от 1 до 5)
- $y_i \in \mathbf{R}$ - ответ - это число

Первые два типа ответов - задача **классификации**(бинарной в случае бинарных ответов и многоклассовой в случае k классов). Если же ответ - вещественное число, то это уже задача **регресии**.

4 Этапы обучения и применения:

1. Построение алгоритма
По объектам $x_1, x_2, ..., x_n \in X$ и ответам $y_1, y_2, ..., y_n$ нужно построить алгоритм a

$$\begin{pmatrix} x_1^1, & \cdots & x_1^m \\ \vdots & \ddots & \vdots \\ x_n^1, & \cdots & x_n^m \end{pmatrix}; \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \to a \tag{1}$$

2. Применение алгоритма:
По объектам $x'_1, ..., x'_l \in X$ без ответа с помощью обученного алгоритма a нужно найти ответы $a(x'_1), ..., a(x'_l)$

$$\begin{pmatrix} {x'_1}^1 & \cdots & {x'_1}^m \\ \vdots & \ddots & \vdots \\ {x'_l}^1 & \cdots & {x'_l}^m \end{pmatrix}; a \to \begin{pmatrix} a(x'_1) \\ \vdots \\ a(x'_l) \end{pmatrix} \tag{2}$$

3. Объяснение работы алгоритма
4. Некоторые типы проблем:

1. Неоднородность train-test
   Пример: Модель предсказания банкротства клиентов банка
   Обучение: Данные о клиентах банка в Москве
   Предсказание: Данные о клиентах банка в Самаре
2. Переобучение Возникает из-за избыточной сложности алгоритма и конечности обучающей выборки. Можно минимизировать с помощью ограничения на параметры модели и проверки на кросс-валидацию.
3. Разные ответы у идентичных объектов(одинаковые заёмщики, один отдал кредит, второй нет)
4. Недостаток данных
5. Избыточность
6. Нет описания признаков
7. Пропуски в данных
8. Зашумлённые данные

5 Мера качетсва модели:

Вопрос: Каким образом a приближает y?
Введём понятие меры качества модели:

$$F = F(Y, a(X)) \tag{3}$$

Мера качества модели бинарной классификации:

$$F = \frac{1}{n} \sum I[a(x_i) = y_i] \tag{4}$$

Здесь I - идентификаторная функция, F смотрит на количество совпадающих ответов и делит их на общее число ответов. Минус данного подхода - несбалансированность выборки. Пример:
Обучающая выборка состоит из 1000000 транзакций, 100 из них фродовые. Построим **матрицу ошибок**(confusion matrix) модели:

truth/model	0	1
0	True negative(TN)	False positive(FP)
1	False negative(FN)	True positive(TP)

(5)

Здесь 0 и 1 - ответы: по горизонтали модели, по вертикали реальные. Тогда мы можем записать формулу (4) в виде: $F = \frac{TP+TN}{TP+TN+FP+FN}$
Введём новые обозначения:

$PR = \frac{TP}{TP+FP}$ - precision и $RC = \frac{TP}{TP+FN}$ - recall - две метрики ошибки модели. Как понять какую выбрать?

$$F_\beta = (1+\beta^2)\frac{PR * RC}{(\beta^2 * PR) + RC} \tag{6}$$

Здесь β- параметр, меняя который мы понимаем какая метрика важнее. Например:
$F_1 = 2\frac{PR*RC}{PR+RC}$
F_2- важнее RC
$F_{0,5}$ - важнее PR
F_β - среднее гармоническое. Минус данных метрик в том, что мы не можем их никак оптимизировать. Для этого лучше использовать следующую функцию (логистическая loss функция):

$$logloss = -\frac{1}{l}\sum_{i=1}^{l}(y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot\log(1-\hat{y}_i)) \tag{7}$$

Здесь $y_i \in \{0,1\}$ -реальные ответы, а $\hat{y}_i \in [0,1]$ - предсказанные скоры.
Самый главный плюс данного подхода - дифференцируемость.

6 Мера качества модели регрессии:

- $MSE = F_{MSE}(y, a(x)) = \frac{1}{n}\sum(y_i - a(x_i))^2$
- $MAE = F_{MAE}(y, a(x)) = \frac{1}{n}\sum|y_i - a(x_i)|$
- $RMSE = F_{RMSE}(y, a(x)) = \sqrt{\frac{1}{n}\sum(y_i - a(x_i))^2}$
- $RMSLE = F_{RMSLE}(y, a(x)) = \sqrt{\frac{1}{n}\sum(\log(y_i+1) - \log(a(x_i)+1)^2}$
- $MAPE = F_{MAPE}(y, a(x)) = \frac{100\%}{n}\sum|\frac{y_i - a(x_i)}{y_i}$
- $SMAPE = F_{SMAPE}(y, a(x)) = \frac{100\%}{n}\sum\frac{2|y_i - a(x_i)|}{|y_i|+|a(x_i)|}$

7 Основные типы алгоритмов машинного обучения:

- **Линейные:**
  - Linear regression
  - Logistic regression

  - Ridge, Lasso regression
  - SVM
- **Логические - алгоритмы на деревьях:**
  - Дерево решений
  - Случайный лес
  - XGBoost
- **Метрические:**
  - kNN
  - Метод потенциальных функций

8 Итог: Алгоритм решения задачи машинного обучения:

- Просмотр и анализ имеющихся данных
- Предобработка данных
- Изобретение признаков
- Выбор модели машинного обучения
- Построение модели машинного обучения
- Оптимизация модели машинного обучения(переобучение,...)
- Оценка качества модели
- Можно использовать на других данных(однородных)