Тема 6: Методы классификации

6.1. Задача классификации

Классификация – системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

Классификация – упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранных для определения сходства или различия между этими объектами.

Под *классификацией* будем понимать отнесение объектов (наблюдений, событий) к одному из заранее известных классов.

Классификация – это выявленная *закономерность*, позволяющая делать *вывод* относительно определения характеристик конкретной группы. Таким образом, для проведения *классификации* должны присутствовать признаки, характеризующие группу, к которой принадлежит то или иное событие или *объект* (обычно при этом на основании анализа уже классифицированных событий формулируются некие правила).

Классификация относится к стратегии *обучения с учителем* (*supervised learning*), которое также именуют контролируемым или управляемым обучением.

Задачей *классификации* часто называют предсказание категориальной зависимой переменной (т.е. зависимой переменной, являющейся категорией) на основе выборки непрерывных и/или категориальных переменных.

Цель процесса *классификации* состоит в том, чтобы построить модель, которая использует прогнозирующие атрибуты в качестве входных параметров и получает *значение* зависимого *атрибута. Процесс классификации* заключается в разбиении *множества* объектов на классы по определенному критерию.

Классификатором называется некая сущность, определяющая, какому из предопределенных классов принадлежит *объект* по вектору признаков.

Для проведения *классификации* с помощью математических методов необходимо иметь формальное описание объекта, которым можно оперировать, используя математический аппарат *классификации*. Таким описанием может выступать *база данных*. Каждый *объект* (*запись базы данных*) несет информацию о некотором свойстве объекта.

Набор исходных данных (или выборку данных) разбивают на два *множества*: обучающее и тестовое.

Обучающее множество (*training set*) – множество, которое включает данные, использующиеся для обучения (конструирования) модели.

Такое множество содержит входные и выходные (целевые) значения примеров. Выходные значения предназначены для обучения модели.

Тестовое множество (*test set*) также содержит входные и выходные значения примеров. Здесь выходные значения используются для проверки работоспособности модели.

Процесс *классификации* состоит из двух этапов: конструирования модели и ее использования.

1. Конструирование модели: описание множества предопределенных классов.
   - Каждый пример набора данных относится к одному предопределенному классу.
   - На этом этапе используется обучающее множество, на нем происходит конструирование модели.
   - Полученная модель представлена классификационными правилами, деревом решений или математической формулой.
2. Использование модели: *классификация* новых или неизвестных значений.
   - Оценка правильности (точности) модели.
     - Известные значения из тестового примера сравниваются с результатами использования полученной модели.

- Уровень точности – процент правильно классифицированных примеров в тестовом множестве.

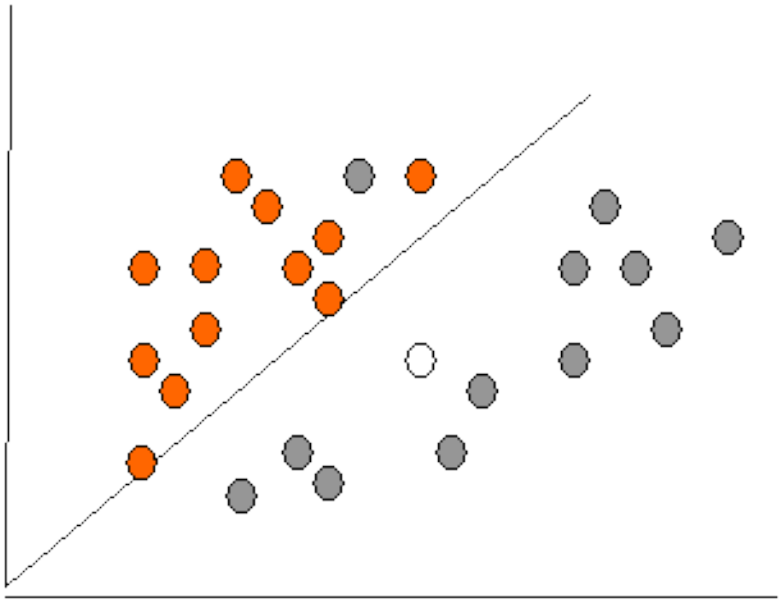
- Тестовое множество, т.е. множество, на котором тестируется построенная модель, не должно зависеть от обучающего множества.

- Если точность модели допустима, возможно использование модели для *классификации* новых примеров, класс которых неизвестен.

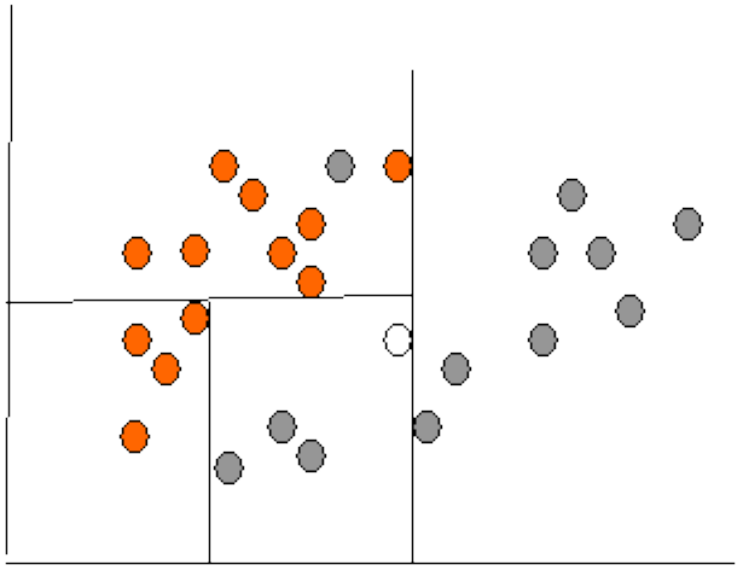
6.2. Основные методы, применяемые для решения задач классификации:

- *классификация* с помощью деревьев решений;
- байесовская (наивная) *классификация*;
- *классификация* при помощи *искусственных нейронных сетей*;
- *классификация* методом *опорных векторов*;
- статистические методы, в частности, линейная регрессия;
- *классификация* при помощи метода ближайшего соседа;
- *классификация CBR*-методом;
- *классификация* при помощи *генетических алгоритмов*.

- Схематическое решение задачи *классификации* некоторыми методами (при помощи линейной регрессии, деревьев решений и нейронных сетей) приведены на рис. 6.1 – 6.3.



- Рис. 6.1. Решение задачи классификации методом линейной регрессии



- Рис. 6.2. Решение задачи классификации методом деревьев решений

- Рис. 6.3. Решение задачи классификации методом нейронных сетей

6.3. Точность классификации: оценка уровня ошибок

Оценка точности *классификации* может проводиться при помощи *кросс-проверки*. *Кросс-проверка* (Cross-validation) – это процедура оценки точности *классификации* на данных из тестового *множества*, которое также называют кросс-проверочным множеством. *Точность классификации* тестового *множества* сравнивается с точностью *классификации* обучающего *множества*. Если *классификация* тестового *множества* дает приблизительно такие же результаты по точности, как и *классификация* обучающего *множества*, считается, что данная модель прошла кросс-проверку.

Разделение на обучающее и тестовое *множества* осуществляется путем деления выборки в определенной пропорции, например обучающее множество – две трети данных и тестовое – одна треть данных. Этот способ следует использовать для выборок с большим количеством примеров. Если же *выборка* имеет малые объемы, рекомендуется применять специальные методы, при использовании которых обучающая и тестовая выборки могут частично пересекаться.

6.4. Оценивание классификационных методов

Оценивание методов следует проводить, исходя из следующих характеристик [21]: скорость, *робастность*, интерпретируемость, *надежность*.

Скорость характеризует время, которое требуется на создание модели и ее использование.

Робастность, т.е. *устойчивость* к каким-либо нарушениям исходных предпосылок, означает возможность работы с зашумленными данными и пропущенными значениями в данных.

Интерпретируемость обеспечивает возможность понимания модели аналитиком.

Свойства классификационных правил:

- размер дерева решений;
- компактность классификационных правил.

Надежность методов *классификации* предусматривает возможность работы этих методов при наличии в наборе данных шумов и выбросов.

Только что мы изучили задачу *классификации*, относящуюся к стратегии "*обучение с учителем*".