Постановка задачи обучения по прецедентам

1. В области машинного обучения объектом называют то, для чего необходимо решить поставленную задачу. Пример – пациенты в задаче медицинской диагностики или письма в задаче определения спама. Признаками называют численные или категориальные значения, которыми можно описать объект в задаче. Пример – симптомы пациентов или слова, встречающиеся в письмах. Ответом называют вывод об объекте, который можно дать после исследования его признаков, он же может быть и ответом на вопрос поставленной задачи. Пример – болен ли пациент или является ли письмо спамом.
2. В матрице объектов-признаков содержатся совокупности признаковых описаний объектов, расположенные построчно.
3. Моделью алгоритма называют параметрическое семейство отображений из пространства признаков и параметров в пространство ответов. Обучение – это процесс подбора оптимальных параметров модели по обучающей выборке. Конкретный метод обучения совершает сам подбор параметров.
4. Этапы обучения:
   1. Построение алгоритма .
   2. Применение алгоритма a для новых объектов и выдача ответов .
5. Переобучение – это явление, когда вместо того, чтобы находить закономерность в данных, алгоритм начинает приближать конкретные точки обучающей выборки, а для других данных давать худший результат. В результате переобучения качество работы алгоритма на объектах, не входящих в тестовую выборку, оказывается гораздо хуже, чем на обучающей выборке. Как бороться – разбивать выборку с известными ответами на обучающую и тестовую и выбирать оптимальные параметры по лучшему результату тестовой выборки.

Метод решающих деревьев

1. Область данных разбивается гиперповерхностями вида . Постепенно спускаясь от корня дерева к листьям и выбирая определённую ветвь в зависимости от ответа на поставленное неравенство, мы относим точку к определённому классу.
2. Для разделения узлов в самых информативных признаках применяются 3 разных критерия: ошибка классификации, критерий Джини и энтропийный критерий.
3. Достоинства метода:
   1. Формирует чёткие и понятные правила классификации.
   2. Быстро обучается и прогнозируется.
   3. Легко визуализируется.
   4. Не требует много параметров модели.

Недостатки метода:

1. Переобучение при отсутствии ограничения на дерево.
2. Деревья чувствительны к шуму в данных, и небольшие изменения обучающей выборки могут привести к глобальным изменениям в модели.
3. По сравнению с другими методами зачастую даёт не лучший ответ.
4. Сложный поиск оптимального решения.