Dzitsoev OpenQ

Dan Immortal

December 2024

1

Основная идея в том, что если train и test выборки происходят из одного и того же распределения и нет значимых сдвигов в данных, то никакая модель не сможет надежно отличить, к какому набору данных относится отдельный образец, иначе между распределениями есть существенная разница и разбиение не репрезентативно.

Метод:

Объединим train и test в один пул и пометим каждую запись меткой 0 train и 1 test. Далее перемешаем записи, чтобы исключить влияние порядка и разделим перемешанную выборку на новую train и test но уже для бинарной классификации {train, test}. После обучим классификатор, пытаясь предсказать метку для каждого образца.

Интерпретировать будем так:

1)Если точность на валидации близка к 50% те случайному угадыванию, значит классификатор не смог найти паттерны, по которым можно отличить train от test. Знаичит разбиение на train/test вероятно репрезентативно.

2)Если точность заметно выше 50%, это говорит о том, что есть систематические различия в распределениях признаков между выборками (нужно пересмотреть разбиение данных стратифицировать, убедиться, что нет временного сдвига, проверить корректность случайной выборки и тд). Если модель(достаточно мощзная) не справляется с этой задачей, значит разбиение можно считать репрезентативным.

2

Предполагаемая модель кластеризации

Пусть $\mathbf{c}_2$ — это центр кластера №2. Каждый клиент i имеет исходный вектор признаков $\mathbf{x}_i \in \mathbb{R}^{10}$. Надо найти вектор $\mathbf{x}'_i$ такой, что он будет ближе к центру кластера $\mathbf{c}_2$, чем к центрам других кластеров и так, чтобы изменения от исходного вектора $\mathbf{x}_i$ были минимальными.

Сформулируем в виде минимизации комбинированной функции потерь:

$\min_{\mathbf{x}'_i} \; D(\mathbf{x}'_i, \mathbf{c}_2) + \lambda \cdot \|\mathbf{x}'_i - \mathbf{x}_i\|^2$

Где

D — мера расстояния между новым вектором клиента и центром кластера 2 (например евклидово или его квадрат).

$\|\mathbf{x}'_i - \mathbf{x}_i\|^2$ — штраф за отклонение от исходных характеристик клиента (чтоб не получить слишком нереалистичное решение)

–λ — кэф регулировании "цены"за изменение признаков

Получим:

$\min_{\mathbf{x}'_i} \; \|\mathbf{x}'_i - \mathbf{c}_2\|^2 + \lambda\|\mathbf{x}'_i - \mathbf{x}_i\|^2$

Хорошей идеей ввести доп ограничения

- Признаки менять нельзя или разрешить только небольшое смещение возраста (реалистично: возраст — это не просто меняется и сделать так чтобы он не выходил за разумные пределы)

- Основные изменения должны касаться транзакционных признаков: количество покупок, средний чек, периодичность транзакций и т.д. Эти параметры клиент потенциально может изменить своим поведением (совершая покупки чаще, увеличивая средний чек и тд).

3 3

Random forest снижает разброс за счёт увеличения числа деревьев. Чем их больше, тем надёжнее итоговый прогноз. 1 лес из 1000 деревьев даёт более стабильный результат, чем два леса по 500, при прочих равных условиях. 1 большой ансамбль проще в применении и поддержке: хранится 1 модель вместо 2ух. Отсутствует дополнительный шаг объединения результатов двух отдельных лесов. С точки зрения статистики общая точность чаще немного выше, а эксплуатация проще. Итого, 1 лес с 1000 деревьев обычно предпочтительнее, но в редких случаях при обучении 2ух лесов по 500 деревьев на слегка разных наборах данных или с отличными параметрами, теоретически можно добиться более устойчивого предсказания при их объединии, но в условии вопроса это не оговорено.

4 4

Очистим и нормализуем признаки, применим k-средних на историческом датасете с параметром $n_{clusters} = K$, где K — число кластеров. После обучения алгоритма на исторических данных у каждого клиента будет метка

кластера.По каждому кластеру вычислим процент клиентов в дефолте и получим вектор вероятностей дефолта для кластеров.

После к новому клиенту у которого признаки известны мы применяем обученный k-средних. Находим к какому кластеру относится новый клиент,предполагая что новый клиент схож с клиентами этого кластера, получим вероятность его дефолта как среднюю долю дефолтников в соответствующем кластере. Признаки надо подготовить для нового клиента так же как и для исторических данных. Количество кластеров и качество кластеризации сильно влияют на итого точность оценки.

5

Кажется, что можно использовать банальную дискретизацию (разбить диапазон на интервалы для классификации используя квантильное разбиение, опираяст на распределение доходов в выборке, чтоб интервалы были относительно равномерно наполнены объектами). Можно логарифмировать перед разбиением на интервалы (случай выбросов), брать медиану вместо среднего (если у интевалов разная ширина) и потениально можно побробовать доработать результат с помощью допольнителной регресси внутри каждого интервала.

Возможно, не понял "подвох"вопроса.