Постановка задачи

Пусть нам дана модель, так что целевая переменная **y** немонотонно зависит от некоторого признака χ. Тем не менее, у нас есть подозрение, что зависимость можно представить в кусочно-линейном виде. Эта зависимость может быть предоставлена экспертами, однако нас такой случай не интересует. Мы будем решать более общую постановку, когда кроме смутных предчувствий в нашем распоряжении ничего нет.
Итак, есть предчувствие, есть признак, есть некоторая выборка, хотим вместо признака ввести несколько других, от которых целевая переменная зависела бы уже монотонно.

Пути к решению ¶

1-й путь

Если выборка достаточно велика, можно попытаться сделать следующее: "кластеризуем" нашу выборку так, чтобы внутри одного кластера расстояние между остальными признаками было $\leq \varepsilon$. Из предположения "достаточности" размера выборки следует, что внутри одного кластера будет много объектов обучающей выборки. Тогда для каждого кластера мы можем нарисовать график зависимости целевой переменной от нашего немонотонного признака. Таким образом мы сможем на глаз определить наше предположение.
Постараемся ввести некий формализм. Будем считать, что признак χ нормализован, т.е. его значения лежат на отрезке $[-1, 1]$. Разобьем отрезок $[-1, 1]$ на **k** равных частей (для конкретики можно взять **k** = 10, однако это число взято с потолка, из чисто интуитивных соображений). Итак, мы разбили отрезок на **k** равных частей. Теперь мы хотим подобрать такое $\varepsilon > 0$, что число ε-шаров, которые содержат хотя бы по одному объекту для каждой части, такому что значение χ для него попадает в эту часть разбиения, максимально. Проще говоря, мы хотим максимизировать число наших "кластеров", для которых χ представлено более или менее на всем отрезке $[-1, 1]$.
Тем не менее, мы не хотим, чтобы само ε было слишм велико. Для простоты рассуждений давайте считать вообще все признаки нормализованными. Скажем, что мы не хотим, чтобы ε было больше некоторого ε_0, которое можно взять, например, равным 0.2.
Всё, что нам теперь остается, это перебирать значения ε на отрезке $[0, \varepsilon_0]$ (например, по сетке с шагом $\frac{\varepsilon_0}{20}$), пытаясь найти наиболее оптимальное значение ε. Что значит оптимальное? Ну например под оптимальным значением можно понимать argmin функции $f(\varepsilon) = \alpha\varepsilon + \beta m^{-1}$, где m - кол-во ε-шаров ("кластеров"), удовлетворяющих нашему требованию (про принадлежность значений χ всем k частям отрезка), а α и β - некоторые константы, о подборе которых мы говорить не будем. Вообще говоря, мы ищем минимум не на всех ε, а только на тех, где m > 0 (в противном случае нам пришлось бы делить на ноль, что неблагоразумно), те же ε, где m = 0, мы исключаем из рассмотрения. Если для всех рассмотренных значений ε **m** = 0, то нужно либо уменьшать k, либо увеличивать ε_0, либо признать, что выборка недостаточно большая.
Предположим, что мы нарисовали графики и убедились, что зависимость кусочно-линейная (хотя бы приблизительно). Если всё настолько плохо, что даже не удается на глаз определить количество частей отрезка, на которых зависимость линейная, то, возможно, наше изначальное предположение о такой зависимости неверное. Есть еще вариант, что ε получилось всё же слишком большим. Чтобы его исключить, увеличим α и повторим эксперимент.
Итак, предположим самый лучший исход: мы на глаз смогли определить **s** - число участков линейности, а также примерные диапазоны. Самое простое, что теперь можно сделать: уточнять диапазоны с помощью кросс-валидации. В принципе, если очень хочется, можно и руками, а можно каким-нибудь поиском по сетке.

2-й путь

В случае с большой выборкой всё замечательно. Если данных с избытком, то первый метод, кажется, должен сработать довольно неплохо. Однако предположим, что выборка у нас не слишком велика, а уверенность насчет кусочной-линейности, напротив, огромна. В обозначениях, введенных выше, мы хотим определить **s** (число участков линейности), и чисел $0 < x_1 < x_2 < \ldots < x_{s-1} < 1$ (сами участки).
Давайте попробуем выкинуть наш "плохой" признак вовсе. Найдем коэффициенты логистической регрессии каким-нибудь из классических методов (хотя бы тем же градиентным спуском). Будем считать, что и без χ всё не так плохо (если χ - чуть ли не единственный информативный признак и без него точность на тестовой выборке низка, то такой метод не сработает). Иными словами, мы надеемся, что и без учета χ можно построить неплохую модель, а рассмотрение χ лишь улучшит результаты, но не слишком значительно.
Найдя коэффициенты при остальных признаках, попытаемся теперь повторить по сути предыдущие рассуждения (путь 1), однако теперь не будем заботиться об ε. Разобьем $[-1, 1]$ на **k** равных отрезков (**k** должно быть точно не меньше искомого **s**, т.е. нужно брать его заведомо большим). Для каждого посмотрим на все объекты, у которых значение χ попало в этот отрезок. Для каждого признака (кроме нашего немонотонного) найдем его медианное значение на этих объектах. Т.к. мы считаем, что мы знаем примерную зависимость целевой переменной от этих признаков, давайте приведем все ее значения к тем, как если бы признаки на объектах принимали медианные значения.
Пример: пусть медианные значения $M_1, M_2, \ldots, M_u$ при признаках $\chi_1, \ldots, \chi_n$. Найденные веса $w_1, \ldots, w_n$. Также свободный коэффициент w_{n+1}. Пусть на некотором объекте X_j ответ дан ответ y_j. Тогда новым ответом на нем будем считать
$y'_j = y_j - w_1(X_{j1} - M_1) - w_2(X_{j2} - M_2) - \ldots - w_n(X_{jn} - M_n)$.
Таким образом мы надеемся, что зависимости теперь линейны (т.к. **k**, в идеале, сильно больше чем **s**, то мы рассчитываем, что на каждом отрезке завимость действительно линейна).
Теперь опять же желательно построить графики. Если на них линейной зависмости не наблюдается, то метод оказался плох и продолжать в том же направлении смысла не имеет. Если же все более или менее хорошо, то можно либо действовать на глазок, как в **путь 1**, либо попыться программно найти кусочно-линейную функцию, максимально хорошо описывающую наши данные.
Т.к. в каждом отрезке (а их k штук, напомним), мы полагаем зависмость линейной, давайте методом наименьших квадратов попытаемся найти зависмость. Получим кусочно-линейную функцию, состоящую из k линейных участков. Далее двигаясь вдоль этих участков, попытаемся понять, в каких местах у нас перелом, т.е. коэффициент наклона отличается больше чем на некоторое ε_1. Тут нужно быть острожным с теми отрезками, которые содержат в себе точку излома - на них найденный коэффициент наклона будет, скорее всего, сильно отличаться и от предыдущего отрезка, и от последующего. Собственно, именно эти отрезки мы и возьмем на карандаш: можно либо просто по сетке искать на них точку излома, либо, если число **k** было взято достаточно большим, просто взять в качестве точки излома середину "подозрительного" отрезка.

Финал

После того, как найдены **s** и $x_1, \ldots, x_{s-1}$, мы подбираем на новых признаках коэффициенты (например, методом градиентного спуска). Если найдено только **s**, а для $x_1, \ldots, x_{s-1}$ лишь примерные интервалы, то мы можем искать значения $x_1, \ldots, x_{s-1}$ по сетке, много раз повторяя процесс обучения для разных значений $x_1, \ldots, x_{s-1}$ и проводя затем кросс-валидацию.