Лекция 5. Регрессия и логистическая регрессия

Линейная регрессия

Теперь предлагаю посмотреть на чуть более сложные модели классификации - линейные классификаторы.

> ❗ **Линейный классификатор** - алгоритм классификации, основанный на построении линейной разделяющей поверхности. В случае двух классов разделяющей поверхностью является гиперплоскость, которая делит пространство признаков на два полупространства. В случае большего числа классов разделяющая поверхность кусочно-линейна.

Самым простым примером является линейная регрессия. Несмотря на то, что алгоритм называется “линейная регрессия”, он применим и для задачи классификации. В этом случае при обучении модели будет строиться разделяющая гиперплоскость (в двумерном случае просто прямая) так, чтобы по одну сторону от неё находились объекты одного класса, а под другую - второго.

Начнем с исходной постановки задачи. Пусть объекты описываются *n* числовыми признаками $f_i : X \to \mathbb{R}, j = 1, ..., n$. Тогда пространство признаковых описаний объектов есть $X = \mathbb{R}^n$. Пусть Y- конечное множество номеров (имён, меток) классов.

Случай 2 классов

Для простоты пусть есть 2 класса и для упрощения работы модели положим метки -1 и 1, т.е. $Y = \{-1, +1\}$.

Линейным классификатором называется отображение из признакового пространства в пространство целевой переменной $f : X \to Y$ вида

$$f(x, \omega) = sign(\sum_{i=1}^{n} \omega_i f_i(x) - \omega_0 + \xi) = sign(X\vec{w} + \xi)$$

где ω_i - вес i-го признака, ω_0 - порог принятия решения, $\vec{\omega}$ - вектор весов, ξ- шумовая составляющая.

Обучение линейного классификатора

Обучение линейного классификатора обычно происходит методом минимизации эмпирического риска, который заключается в том, чтобы по имеющейся обучающей выборке построить алгоритм, минимизирующий функционал эмпирического риска или проще говоря, величины ошибки алгоритма на обучающей выборке

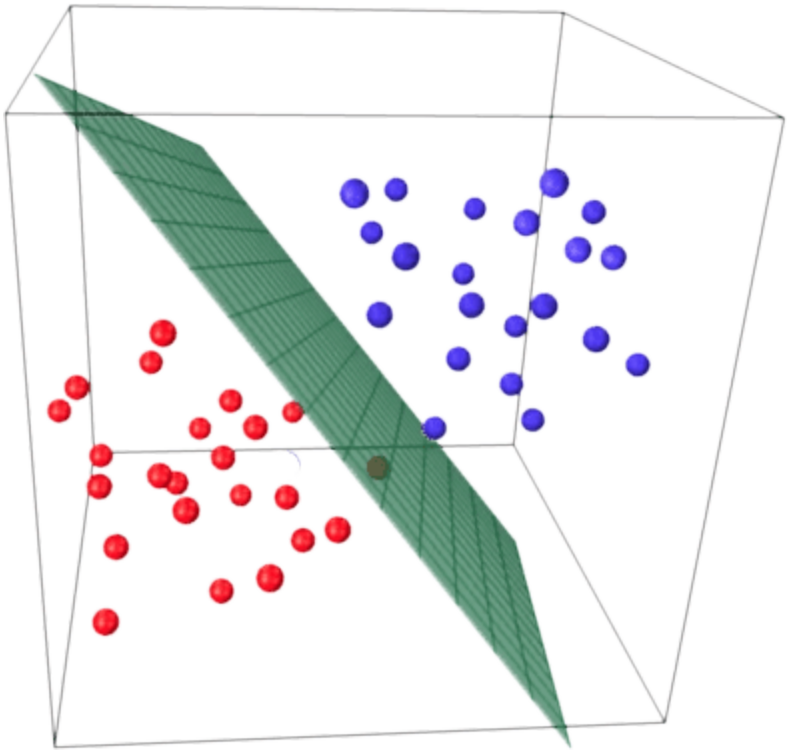
$$Q(\omega) = \sum_{i=1}^{n} [f(x_i, \omega) \neq y_i] \rightarrow min_{\omega}$$

Т.е. мы проверяем, на каком количестве прецедентов модель ошибается и далее при помощи метода наименьших квадратов минимизируем эту ошибку, меняя значения весов.

Для оптимизации метрик можно использовать различные численные методы, градиентный бустинг и так далее.

Логистическая регрессия

В отличие от обычной регрессии, в методе логистической регрессии не производится предсказание значения числовой переменной исходя из выборки исходных значений. Вместо этого, значением функции является вероятность того, что данное исходное значение принадлежит к определенному классу. Для простоты, давайте предположим, что у нас есть только два класса и вероятность, которую мы будем определять, вероятности того, что некоторый объект принадлежит положительному классу.



Основная идея логистической регрессии заключается в том, что пространство исходных значений может быть разделено гиперплоскостью на две области, соответствующих исходным классам. В случае двух измерений — это просто прямая линия без изгибов. В случае трех — плоскость, и так далее. Эта граница задается в зависимости от имеющихся исходных данных и обучающего алгоритма.

Эта плоскость называется линейным дискриминантом, так как она является линейной с точки зрения своей функции, и позволяет модели производить разделение точек на 2 класса.

Определение принадлежности классу

Но каким образом используется линейная граница в методе логистической регрессии для количественной оценки вероятности принадлежности точек данных к определенному классу?

Во-первых, давайте попробуем понять геометрический подтекст «разделения» исходного пространства на две области. Возьмем для простоты двумерный случай, когда признаковое пространство состоит из 2 значений x_1 и x_2:

$$\omega_0 + \omega_1 x_1 + \omega_2 x_2$$

Подставляя значения из обучающего датасета, мы получаем взвешенную сумму и в зависимости от положения исходного объекта, возможны следующие ситуации:

$\omega_0 + \omega_1 x_1 + \omega_2 x_2 > 0$. Объект принадлежит положительному (с меткой 1) классе и для этого объекта классификатор отработал верно. Чем больше это значение, тем более удаленной от границы является наша точка и тем выше вероятность принадлежности верному классу, т.е. $P_+ \in (0.5, 1]$.

$\omega_0 + \omega_1 x_1 + \omega_2 x_2 = 0$. Граничная ситуация, когда объект находится на самой границе. Модель не может однозначно определить какому-то классу и $P_+ = 0.5$

$\omega_0 + \omega_1 x_1 + \omega_2 x_2 < 0$. Тогда объект принадлежит негативному классу (метка 0) и тут аналогично с положительным классом, чем больше по модулю значение, тем выше уверенность модели, $P_+ \in [0, 0.5)$

Итак, мы имеем функцию, с помощью которой возможно получить значение в пределах $(-\infty; +\infty)$ имея точку исходных данных. Но каким образом преобразовать полученное значение в вероятность P_+, пределы которой [0, 1]? Ответ — с помощью функции отношения шансов (Odds Ratio - OR).

Пусть $P(X)$ - вероятность происходящего события X. Тогда, отношение шансов $OR(X)$ определяется из $\frac{P(X)}{1-P(X)}$, а это — отношение вероятностей того, произойдет ли событие или не произойдет. Очевидно, что вероятность и отношение шансов содержат одинаковую информацию. Но, в то время как $P(X)$ находится в пределах от 0 до 1, $OR(X)$ находится в пределах от 0 до ∞.

Давайте вычислим логарифм $OR(X)$, теперь границы значений для нового выражения будут находиться в диапазоне $(-\infty; +\infty)$.

Таким образом, мы получили способ интерпретации результатов, подставленных в граничную функцию исходных значений. В используемой нами модели граничная функция определяет логарифм отношения шансов класса "+". В двумерном случае алгоритм работы логистической регрессии выглядит так:

Шаг 1. Вычислить значение $\omega_0 + \omega_1 x_1 + \omega_2 x_2$ граничной функции, обозначим эту величину v.
Шаг 2. Вычислить отношение шансов: $OR_+ = e^v$
Шаг 3. Имея значение OR_+, вычислить P_+ с помощью простой зависимости $P_+ = \frac{OR_+}{1+OR_+}$
Получив значение v на шаге 1, можно объединить шаги 2 и 3:

$P_+ = \frac{e^v}{1+e^v}$

Правая часть уравнения, указанного выше, называется логистической функцией. Отсюда и название, данное этой модели обучения.

Обучение модели

Общая идея заключается в следующем:
Рассмотрим функцию $f(x)$, где x - точка данных обучающей выборки. Если x принадлежит положительному классу, то $f(x) = P_+$, иначе $f(x) = 1 - P_+$

Функция $f(x)$ проводит количественную оценку вероятности того, что точка обучающей выборки классифицируется моделью правильным образом. Поэтому, среднее значение для всей обучающей выборки показывает вероятность того, что случайная точка данных будет корректно классифицирована системой, независимо от возможного класса.

Скажем проще — механизм обучения логистической регрессии старается максимизировать среднее значение $f(x)$. А название этого метода — метод максимального правдоподобия. Основу этого метода составляет плотность вероятности совместного появления результатов выборки

$$L(y_1, y_2, ..., y_m; \omega) = p(y_1; \omega) * ... * p(y_m; \omega)$$

Согласно методу максимального правдоподобия в качестве оценки неизвестного параметра принимается такое значение $\omega = \omega(y_1, y_2, ..., y_m)$, максимизирующее функцию L.

Для упрощения предлагается работать с логарифмом этого функционала:

$$ln(L(y; \omega)) \to max$$

Обозначим через P_i вероятность появления единицы: $P_i = Prob(Y_i = 1)$. Эта вероятность будет зависеть от $x_i\theta$

$$P_i = f(x_i, \omega), f(z) = \frac{1}{1 + e^{-z}}$$

Далее, для вычисления коэффициентом можно применять любые градиентные методы: метод переменной метрики, сопряженные градиенты и так далее.

Подытожим

Мы рассмотрели один из самых популярный алгоритмов - логистическая регрессия. Преимуществами этого подхода является работа не с метками классов, а с вероятностями меток классов, что позволяет более точечно настраивать алгоритм. Также этот алгоритм в силу своей простоты не требует объемных вычислительных мощностей. Но в то же время этот подход хорошо работает только с переменными одного типа и время обучения модели нелинейно начинает расти при высокой размерности признакового пространства.