Автор:

Дмитриев Егор Андреевич

студент

ФГАОУ ВО «Самарский национальный исследовательский университет им. академика С.П. Королева»

г. Самара, Самарская область

ЛИНЕЙНАЯ РЕГРЕССИЯ

Аннотация: *в статье рассматривается один из методов машинного обучения – линейная регрессия.*

Ключевые слова: *регрессия, функционал, машинное обучение.*

Основные определения

Определение 1. Регрессия – математическое выражение, отражающее зависимость математического ожидания одной случайной величины от других случайных величин.

Определение 2. Функционал – отображение $a: X \to R$, где X – метрическое пространство признаков классифицируемого объекта, а R – множество действительных чисел.

Определение 3. Машинное обучение – нахождение отображения, в частности алгоритма классификации, который строится по множеству, называемому обучаемой выборкой, а качество обучения проверяется по множеству, называемому тестовой выборкой.

Введение

Регрессия – это модель, с помощью которой мы получаем наш функционал для оптимизации. Естественно, построение моделей одна из ключевых составляющих в машинном обучении. Используя разные модели, мы подбираем функционал, с помощью которого будет достигаться минимальное отклонение от желаемого результата.

Постановка проблемы

Необходимо ознакомиться с принципами работы линейной регрессии.

Описание метода линейной регрессии

Суть задачи классификации заключается в построении алгоритма классификации по обучаемой выборке. Пусть у нас имеются классы, представленные в виде чисел. Тогда, используя функцию регрессионной зависимости f, имеем модель данных:

$$y(x_i) = f(x_i, a) + \varepsilon_i$$

где ε_i – случайные величина с некоррелированным гауссовским шумом с нулевым математическим ожиданием, имеющие плотность распределения:

$$\frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{\varepsilon_i^2}{2\sigma_i^2}}$$

Используя метод максимального правдоподобия:

$$L(\varepsilon_1, \dots, \varepsilon_l|\alpha) = \prod_{i=1}^{l} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{\varepsilon_i^2}{2\sigma_i^2}} \to max$$

$$-\ln\big(L(\varepsilon_1, \dots, \varepsilon_l|\alpha)\big) = const(a) + \frac{1}{2}\sum_{i=1}^{l} \frac{1}{\sigma_i^2}(f(x_i, a)-y_i)^2 \to min_a$$

Теперь рассмотрим метод наименьших квадратов:

$$Q(a, X^l) = \sum_{i=1}^{l} w_i(f(x_i, a)-y_i)^2 \to min_a$$

Как данный метод очень похож на метод правдоподобия, при условии, что все невязки ε_i имеют нормальное распределение, некоррелированны и с одинаковой дисперсией.

Перейдем к рассмотрению многомерной линейной регрессии. В данном случае функция f выглядит таким образом:

$$f(x_i, a) = \sum_{j=1}^{n} a_j f(x_j)$$

В матричном виде функционал запишется как:

$$Q(a, X^l) = ||Fa-y||^2 \to min_a$$

где F – матрица l*n, объект – признаки.

Решением для данной задачи будет являться векторов параметров a, который равен:

$$a = (F^t F)^{-1} F^t y$$

В силу линейной возможной линейной зависимости столбцов матрицы F, вектор параметров a по норме может получится очень большим, тем самым отображение будет очень чувствительно к шумам. Для того, чтобы бороться с увеличением весов введем дополнительное слагаемое в целевой функции:

$$Q(a, X^l) = ||Fa\text{-}y||^2 + \frac{1}{2\sigma}|a| \to min_a$$

где $\tau = \frac{1}{2\sigma}$ – параметр регуляризации.

Заключение

В данной работе был рассмотрены принципы работы линейной регрессии.

Список литературы

1. Айвазян С.А. Прикладная статистика: классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989.

2. Дрейпер Н. Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит. – М.: Вильямс, 2007.

3. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979.