

Линейная регрессия.

1. Загрузите данные из набора Forest Fires (файл forestfires.csv) о лесных пожарах в Португалии. Задача состоит в том, чтобы с помощью линейной регрессии научиться предсказывать координату area (площадь пожара) в виде линейной комбинации других данных.

Преобразование данных. Чтобы работать с числовыми координатами, нечисловые координаты (month, day) нужно перевести в числовые. Для простоты можно заменить координату month на индикатор летнего сезона, а координату day не использовать вообще. По желанию можно сделать преобразование другим способом. Так же желательно добавить координату, тождественно равную единице.

Разбейте выборку на две части в соотношении 7:3 (перемешав её с помощью `random.shuffle`). По первой части постройте регрессионную модель. Примените модель ко второй части выборки и посчитайте по ней среднеквадратичную ошибку.

Сделайте для area преобразование $f(x) = \ln(x + c)$ и постройте для нее новую регрессионную модель. Посчитайте среднеквадратичную ошибку для преобразованных значений. При каком c предсказания получаются лучше всего?

При выбранном c сделайте разбиение выборки в соотношении 7:3 разными способами (перемешивая каждый раз). Сильно ли зависит качество от способа разбиения? Сделайте выводы.

2. Пусть $X_i = \beta_1 + i\beta_2 + \varepsilon_0 + \dots + \varepsilon_i, i = 0, 1, \dots, n$ – расстояния, которое проехал трамвай за i секунд по показанию датчика. Здесь β_1 – начальное расстояние, β_2 – скорость трамвая, ε_0 – ошибка начального показания датчика. Трамвай едет с постоянной скоростью, и через каждую секунду датчик фиксирует расстояние, которое проехал трамвай. Отсчет времени идет от предыдущего замера, причем отсчет происходит с ошибкой. Для $i = 1, \dots, n$ величина ε_i есть ошибка приращения расстояния, то есть $\varepsilon_i = \varepsilon_i^t \beta_2$, где ε_i^t – ошибка отсчета времени. Все ошибки ε_i независимы и распределены по закону $N(0, \sigma^2)$. Сведите задачу к линейной модели и найдите оценки наименьших квадратов для начального расстояния β_1 и скорости β_2 , а также несмещенную оценку для σ^2 , из которой выразите оценку дисперсии отсчета времени. Данные взять из файла на диске. Сделайте выводы.