

Задача 1

1а) Загрузите данные по одной объясняемой переменной y (вида 0/1) и двум объясняющим x_1, x_2 . Оцените модель логит. Постройте кривую ROC для имеющихся данных (не прогнозов).

1б) Разбейте точки x_1, x_2 на кластеры. Сами выберите число кластеров и параметры. Изобразите кластеры на графике. Создайте категориальную переменную принадлежности кластеру.

1в) Оцените модель логит, добавив в 1-ю модель созданную категориальную переменную. Постройте аналогичную кривую ROC и добавьте на тот же график ROC по 1-й модели.

1г) По 2-й модели запишите формулу, взяв конкретные коэффициенты из таблицы, по которой можно рассчитать прогноз вероятности события $y=1$ для некоторых x_1, x_2 и 2-го по счету кластера.

1д) Постройте по 2-й модели таблицу сопряженности (contingency она же confusion) для уровня вероятности 0.5.

1е) Как вы думаете, какая модель более точно описывает данные? Объясните.

Задача 2

2а) Постройте каким-либо методом прогноз ряда, взяв последние 24 наблюдения в качестве тестового периода, а все предыдущие в качестве тренировочного периода. Прогноз должен быть осмысленным, но не обязательно точным. Объясните свой метод словами.

2б) Постройте график фактического значения ряда и прогноза. Рассчитайте ошибки прогноза и нарисуйте их график (+ график нуля).

2в) Рассчитайте показатель MAD для вашего прогноза.

Задача 3

3а) Возьмите свой ряд значений фондового индекса и создайте ряд доходностей r_t . Постройте график ряда и убедитесь каким-либо способом, что имеет место кластеризация волатильности.

3б) Оцените волатильность доходностей по экспоненциальному сглаживанию Riskmetrics с какими-либо двумя разными параметрами. Постройте график двух рядов волатильности.

3в) Две оценки волатильности можно рассматривать как прогнозы корня из дисперсии $\hat{\sigma}_t$. Получите для двух прогнозов значения потерь по функции потерь

$$QLIKE_t = \frac{r_t^2}{\hat{\sigma}_t^2} + \ln(\hat{\sigma}_t^2).$$

(Первые несколько наблюдений лучше не учитывать). Сравните прогнозы по тесту Диболда–Мариано.