Машинное обучение. Лабораторная работа №3

Алексей Романенко
alexromsput@gmail.com

16 апреля 2015 г.

1 Исследование параметров регуляризации

Изучаем Regression на примере методов из`sklearn.linear_model.`: `linear_regression`, `La`
https://www.wakari.io/sharing/bundle/aromanenko/RegressionExample.

Данные http://lib.stat.cmu.edu/datasets/houses.zip.

1. Скачайте данные, сохраните в структуре `numpy.array`. Изучите признаки объекты (сколько регрессоров, сколько объектов в обучающей выборке, есть ли пропуски в данных, есть ли выборсы)?

2. Обучите метод `linear_regression`. Выведите ошибку на обучающей выборке и на тестовой. Рассчитайте число обусловленности матрицы.

3. Обучите метод `Ridge`. Подберите оптимальное значение параметра регуляризации (на обучающей выборке). Выведите значение функции потерь MSE (mean squared error) на тестовой и обучающей выборке, а также число обусловленности матрицы в зависимости от параметра регуляризации.

4. Обучите метод `Lasso`. Подберите оптимальное значение параметра регуляризации (на обучающей выборке). Выведите значение функции потерь MSE (mean squared error) на тестовой и обучающей выборке, а также число обусловленности матрицы в зависимости от параметра регуляризации.

5. Отобразите веса регрессоров в зависимости от параметра регуляризации для методов `Ridge` и `LASSO` (см. пример
   https://www.wakari.io/sharing/bundle/aromanenko/RegressionExample).

6. * Решите задачу регрессии методом `LARSLasso`, выведите оптимальные веса регрессоров, объясните разницу между полученными результатами в этом пункте и в пунктах 2.–4.

2 Прогнозирование временных рядов

Изучаем авторегрессионные методы прогнозирования временных рядов (см. ML_MIPT_10.ipnb https://drive.google.com/folderview?id=0B9ZsO9o9XXqNRXpyWG05V

Данные consumption_train.csv https://drive.google.com/folderview?id=0B9ZsO9o9XXc

1. Скачайте данные, сохраните в структуре `pandas.DataFrame`. Изучите ряд «EnergyCons»: длина истории, наличие созонности (какая сезонность наблюдается), наличие трендов).

2. Постройте прогноз с отсрочкой $h = 1$ скользящим методом: $\hat{x}_{t+1} = x_{t+1-168}$ (напомню, 168 - количество часов в неделе). Оцените точность прогноза по критериям $MAPE$ и R^2 на обучающей выборке и на тестовой.

3. Настройка ширины окна K при авторерессионном прогнозе. Для отсрочки пронгнозирования $h = 1$ постройте график зависимости $MAPE$ от ширина окна K в модели авторегрессии. Начиная с какой ширина окна точность прогноза изменяется незначительно? Повторите настройку ширина окна относительно критерия R^2. Сильно ли отличаются оптимальные значения K при $MAPE$ и R^2?

4. Повторите шаги из предыдущего пукнта для отсрочки прогноза $h = 168$.

5. Выполните отбор признаков в авторегрессионной модели для $K = 168$ и $h = 1$ (можно использовать, как методы отбора признаков (ADD-DELL), так и `LASSO`). Повторите процедуру для отсрочки $h = 168$ и $K = 672$.

6. * Изучите методы $ARMA$ и $ARIMA$ на примере пакета `statsmodels`. Вот небольшая статья о методах и их настройке http://conference.scipy.org/proceedi
Вот несколько примеров с запуском методов: http://statsmodels.sourceforge.net/deve
http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/tsa_arma

   Настройте параметры $ARMA$ и постройте прогноз для $h = 1$ и $h = 168$.

3 Литература

[1] *Воронцов К. В.* Математические методы обучения по прецедентам (теория обучения машин) // http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf