Модели авторегрессии

Волхонский Денис
dvolkhonskiy@gmail.com

ВШЭ

2 февраля 2017 г.

1 Необходимая теория

1. **Скользящее среднее (MA, Mean Average)**

   Имеем дискретные данные, хотим смоделировать какаю-либо зависимость. Имеем:

   - Оператор L: $Lx_{n+1} = x_n$ — оператор сдвига времени назад;
   - Белый шум $\varepsilon_n,\ n \geq 0,\ E\varepsilon_n = 0,\ D\varepsilon_n = \sigma^2,\ E\varepsilon_n\varepsilon_m = 0\ \forall n \neq m$;
   - Многочлен $\beta(L) = b_0 + b_1 L + \ldots + b_q L^q$

   Тогда скользящее среднее h_n порядка q называется $h_n = \mu + \beta(L) \cdot \varepsilon_n$. То есть мы имеем постоянное среднее плюч среднее значение последних q значений шума.

2. **Авторегрессионная модель (AR)**

   $h_n = \mu_n + \sigma \cdot \varepsilon_n$, где $\mu_n = a_0 + a_1 h_{n-1} + \ldots + a_p h_{n-p}$.

   Если задать $\alpha(L) = 1 - a_1 L - \ldots - a_p L^p$, тогда $\alpha(L) h_n = a_0 + \sigma\varepsilon_n$.

   Условие стационарности AR:

   - AR(1): $-1 < a_1 < 1$
   - AR(2): $-1 < a_2 < 1,\ a_1 + a_2 < 1,\ a_2 - a_1 < 1$

3. **Авторегрессионное скользящее среднее (ARMA, AutoRegression Mean Average)**

   ARMA(p, q) — смесь моделей AR и MA: $h_n = \mu_n + \sigma\varepsilon_n$, где

   $$\mu_n = (a_0 + a_1 h_{n-1} + \ldots + a_p h_{n-p}) + (b_1 + \varepsilon_{n-1} + b_2 \varepsilon_{n-2} + \ldots + b_q \varepsilon_{n-q})$$

   Согласно теореме Вольда, любой стационарный ряд может быть аппроксимирован моделью ARMA(p, q) (в том смысле, что если мы поставим вместо p и q бесконечность, то мы сможем точно описать любой стационарный ряд).

4. **Интегральная модель (ARIMA)**

   ARIMA(p, d, q). Данная модель представляет собой случай, когда в модели ARMA вместо h_n подставляется ряд d-х разностей h'_n. ARIMA(p, 0, q) = ARMA(p, q)

5. **Оценка коэффициентов в модели ARIMA(p, d, q)**
   - Как правило считают, что шум ε_n смоделирован гауссовским распределением. Тогда при заданных p, d, q параметры модели оцениваются методом максимального правдоподобия. То есть критерием качества является логарифм правдоподобия (Log Likelihood);
   - d выбирается так, чтобы ряд был стационарным;
   - p, q нельзя выбирать из ММП, т.к. LL растёт с ростом p и q;
   - Используют автороррреляционную функцию и частично автокорреляционную функцию

6. **(Авторегрессионная модель условной неоднородности, AutoRegressive Conditional Heteroskedastic model) ARCH**

   Пусть $\varepsilon_n \sim \mathcal{N}(0,1)$. Тогда условно-гауссовская модель $h_n = \sigma_n \varepsilon_n$, где волатильность

$$\sigma^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i h_{n-i}^2,\ \alpha_0 > 0,\ \alpha_i \geq 0$$

7. **Получение стационарности для ряда**
   - Преобразование Бокса-Кокса для нормализации дисперсии:

$$\hat{h_n} = \begin{cases} \ln h_n,\ \lambda = 0 \\ \frac{h_n^\lambda - 1}{\lambda},\ \lambda \neq 0 \end{cases}$$

     Параметр λ Выбирается так, чтобы минимизировать дисперсию или максимизировать правдоподобие модели

     Что важно:
     – Если значения ряда отрицательные, и преобразование невозможно, нужно прибавить к ряду константу;
     – Можно округлять значения λ, чтобы упростить интерпретацию;
     – Как правило преобразование слабо влияет на прогноз и сильно на предсказательный интервал
   - Дифференцирование ряда (может применяться неоднократно):

$$\hat{h_n} = h_n - h_{n-1}$$

8. **Выбор модели**
   - $AIC = 2L + 2(p + q + k + 1)$, где $k = 1$, если в ряде есть константа, $k = 0$, если константы нет;
   - $BIC = -2L + (\log T - 2)(p + q + k + 1)$, где T — длина ряда;
   - Автокорреляционная функция и частичная автокорреляционная функция;
   - Метрика качества (например, MAE или MSE) на прошлых данных;