Бритва Оккама

Юрий Мокрий

25 сентября 2017 г.

Введение

Бритва Оккама — методологический принцип, получивший название по имени английского монаха-францисканца, философа-номиналиста Уильяма Оккама.

В кратком виде гласит: *Не следует множить сущее без необходимости*.

В машинном обучении это принцип можно интерпретировать следующим образом: процесс обучения может быть представлен как поиск наиболее простой модели, соответствующей обучающей выборке.

Количество параметров как сложность модели

С точки зрения Байесовского подхода:

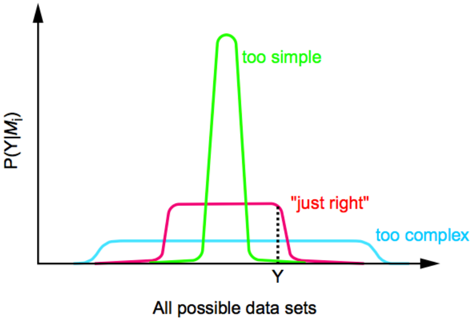
1. Вывести вероятности моделей разных размеров при условии наших данных и использовать их для подбора модели.

2. Не подбирать размер модели, просто взять большую модель.

Подход 1

$$P(M_i|Y) = \frac{P(Y|M_i)P(M_i)}{P(Y)}$$

В предположении, что априорное распределение на модели равномерное, можно просто подбирать модель, используя обоснованность.



Модель Фурье. Постановка.

Модель:

$$y(x) = a_0 + \sum_{d=1}^{D} a_d \cos(dx) + b_d \sin(dx)$$

Параметры - $a_0, a_1, b_1 \dots a_D, b_D$.

Априорное распределение на веса:

$$p(w|S, c) \propto \exp(-\frac{S}{2}(c_0 a_0^2 + \sum_{d=1}^{D} c_d (a_d^2 + b_d^2)))$$

Модель Фурье. Вывод.

Правдоподобие:

$$p(y|x, w, \tau) \propto \prod_{n=1}^{N} \exp(-\frac{\tau}{2}(y_n - w^T \Phi_n)^2)$$

Априорные распределения на параметры (для удобства вычислений введём $C : S = C\tau$):

$$p(\tau) \propto \tau^{a_1 - 1} \exp(-\beta_1 \tau),\ p(C) \propto C^{a_2 - 1} \exp(-\beta_2 C)$$

Выведем обоснованность $p(y|x, C, c)$.

Модель Фурье. Обоснованность.

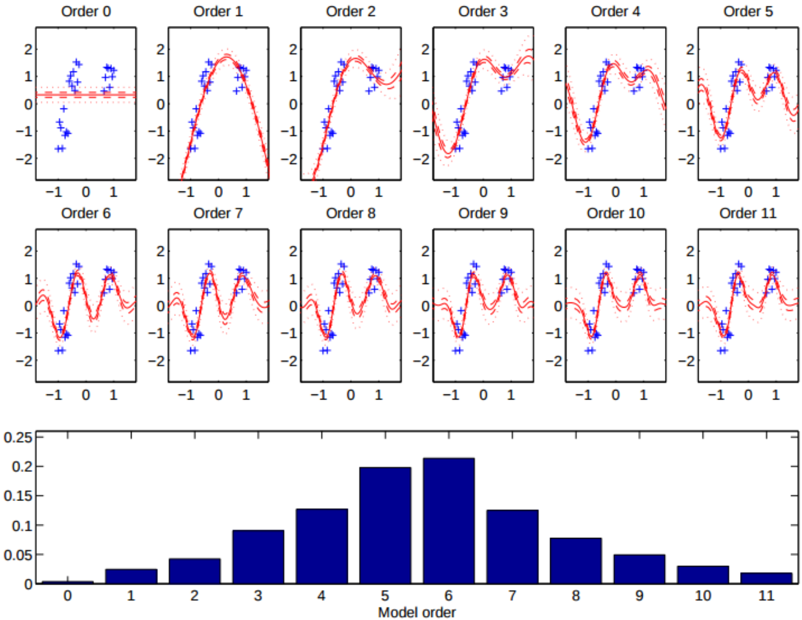
$$E(C, \mathbf{c}) = \iint p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \tau) p(\mathbf{w}|C, \tau, \mathbf{c}) p(\tau) p(C) d\tau d\mathbf{w} = \frac{\beta_1^{\alpha_1} \beta_2^{\alpha_2} \Gamma(\alpha_1 + N/2)}{(2\pi)^{N/2} \Gamma(\alpha_1) \Gamma(\alpha_2)}$$

$$\times |\mathbf{A}|^{1/2} \left[\beta_1 + \frac{1}{2} \mathbf{y}^\top (I - \Phi \mathbf{A}^{-1} \Phi^\top) \mathbf{y}\right]^{-\alpha_1 - N/2} C^{D+\alpha_2-1/2} \exp(-\beta_2 C) c_0^{1/2} \prod_{d=1}^{D} c_d,$$

$$A = \Phi^T \Phi + \mathrm{Diag}\{\bar{c}\}$$

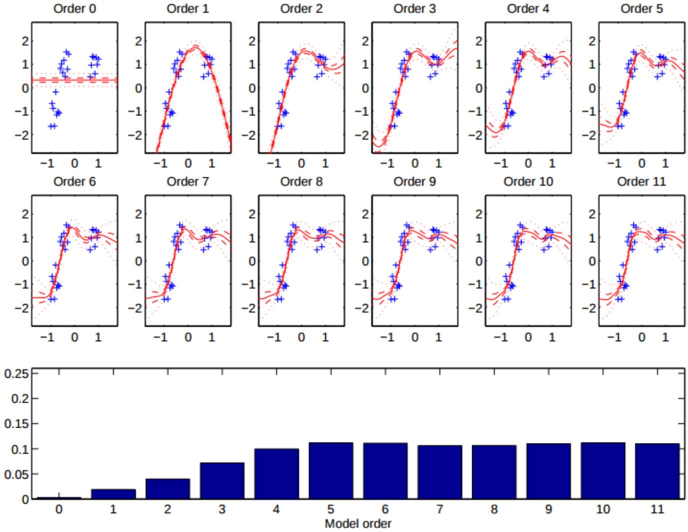
Данный функционал легко максимизируется по C, но с c всё сложно. Скажем, что $c = 1$.

Модель Фурье. $c = 1$



Модель Фурье. $c = d^\gamma$

Пусть $c_d = d^\gamma$ и $\gamma = 3$.



Параметр γ

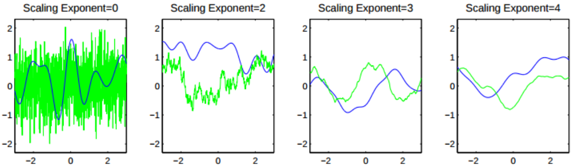
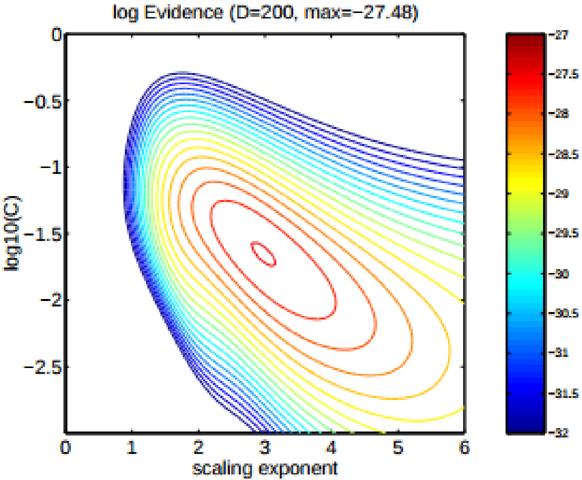


Figure 3: Functions drawn at random from the Fourier model with order $D = 6$ (dark) and $D = 500$ (light) for four different scalings; limiting behaviour from left to right: discontinuous, Brownian, borderline smooth, smooth.

Фактически, параметр γ определяет насколько мы поощряем «простые» функции по сравнению со «сложными». Чем больше γ, тем проще будут функции.

И снова Бритва Оккама



Ссылка на статью

http://mlg.eng.cam.ac.uk/zoubin/papers/occam.pdf