Линейные модели

Сергей Николенко

Казанский Федеральный Университет, 2014

Outline

1. Линейная регрессия
   - Линейная регрессия

В предыдущей серии...

- Теорема Байеса:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}.$$

- Две основные задачи байесовского вывода:
  1. найти апостериорное распределение на гипотезах/параметрах:

$$p(\theta \mid D) \propto p(D|\theta)p(\theta)$$

     (и/или найти гипотезу максимального правдоподобия $\arg\max_\theta p(\theta \mid D)$);
  2. найти апостериорное распределение исходов дальнейших экспериментов:

$$p(x \mid D) \propto \int_{\theta\in\Theta} p(x \mid \theta)p(D|\theta)p(\theta)\mathrm{d}\theta.$$

Метод наименьших квадратов

- Линейная модель: рассмотрим линейную функцию

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{p} x_j w_j = \mathbf{x}^\top \mathbf{w}, \quad \mathbf{x} = (1, x_1, \ldots, x_p).$$

- Таким образом, по вектору входов $\mathbf{x}^\top = (x_1, \ldots, x_p)$ мы будем предсказывать выход y как

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{j=1}^{p} x_j \hat{w}_j = \mathbf{x}^\top \hat{\mathbf{w}}.$$

Метод наименьших квадратов

- Как найти оптимальные параметры $\hat{\mathbf{w}}$ по тренировочным данным вида $(\mathbf{x}_i, y_i)_{i=1}^N$?
- Метод наименьших квадратов: будем минимизировать

$$\mathrm{RSS}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

- Как минимизировать?

Метод наименьших квадратов

- Можно на самом деле решить задачу точно – записать как

$$\mathrm{RSS}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}),$$

  где $\mathbf{X}$ – матрица $N \times p$, продифференцировать по $\mathbf{w}$, получится

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

  если матрица $\mathbf{X}^\top \mathbf{X}$ невырожденная.
- Замечание: $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$ называется *псевдообратной матрицей Мура–Пенроуза* (Moore–Penrose pseudo-inverse) матрицы $\mathbf{X}$; это обобщение понятия обратной матрицы на неквадратные матрицы.
- Много ли нужно точек, чтобы обучить такую модель?

Метод наименьших квадратов

- Пример: задача классификации. Два класса; мы кодируем один ответ как $y = 1$, другой ответ как $y = 0$ и рисуем прямую $\mathbf{x}^\top \hat{\mathbf{w}} = \frac{1}{2}$.
- Мы видим, что данные разделяются не то чтобы совсем замечательно.
- Когда линейная модель работает хорошо, когда плохо?
- Предположим, что это была смесь нескольких нормальных распределений – что тогда?

Байесовская регрессия

- Теперь давайте поговорим о линейной регрессии по-байесовски.
- Основное наше предположение – в том, что шум (ошибка в данных) распределён нормально, т.е. переменная t, которую мы наблюдаем, получается как

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Иными словами,

$$p(t \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(t \mid y(\mathbf{x}, \mathbf{w}), \sigma^2).$$

- Здесь пока y – любая функция.

Байесовская регрессия

- Чтобы не повторять совсем уж то же самое, мы рассмотрим не в точности линейную регрессию, а её естественное обобщение – линейную модель с базисными функциями:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})$$

(M параметров, $M-1$ базисная функция, $\phi_0(\mathbf{x}) = 1$).

Байесовская регрессия

- Базисные функции ϕ_i – это, например:
  - результат feature extraction;
  - расширение линейной модели на нелинейные зависимости (например, $\phi_j(x) = x^j$);
  - локальные функции, которые существенно не равны нулю только в небольшой области (например, гауссовские базисные функции $\phi_j(\mathbf{x}) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$);
  - ...

Байесовская регрессия

- Рассмотрим набор данных $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ со значениями $\mathbf{t} = \{t_1, \ldots, t_N\}$.
- Будем предполагать, что данные взяты независимо по одному и тому же распределению:

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \sigma^2\right).$$

- Прологарифмируем (опустим $\mathbf{X}$, т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right)^2.$$

Байесовская регрессия

- Прологарифмируем (опустим $\mathbf{X}$, т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n) \right)^2.$$

- И вот мы получили, что для максимизации правдоподобия по $\mathbf{w}$ нам нужно как раз минимизировать среднеквадратичную ошибку!

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n) \right) \boldsymbol{\phi}(\mathbf{x}_n).$$

Байесовская регрессия

- Решая систему уравнений $\nabla \ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = 0$, получаем то же самое, что и раньше:

$$\mathbf{w}_{ML} = \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \mathbf{t}.$$

- Здесь $\mathbf{\Phi} = (\phi_j(\mathbf{x}_i))_{i,j}$.

Байесовская регрессия

- Теперь можно и относительно σ^2 максимизировать правдоподобие; получим

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top_{ML}\boldsymbol{\Phi}(\mathbf{x}_n)\right)^2,$$

т.е. как раз выборочная дисперсия имеющихся данных вокруг предсказанного значения.

Оверфиттинг

- Пример регрессии с базисными функциями:
  - обучающая выборка – несколько точек $\mathbf{x} = \{x_1, \ldots, x_N\}$ с наблюдениями неизвестной функции $\mathbf{t} = \{t_1, \ldots, t_N\}$;
  - мы хотим понять, что это была за функция $y(x)$, и предсказать новые значения $y(x)$;
  - давайте будем просто искать функцию в виде многочлена:

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_d x^d = \sum_{j=0}^{d} w_j x^j.$$

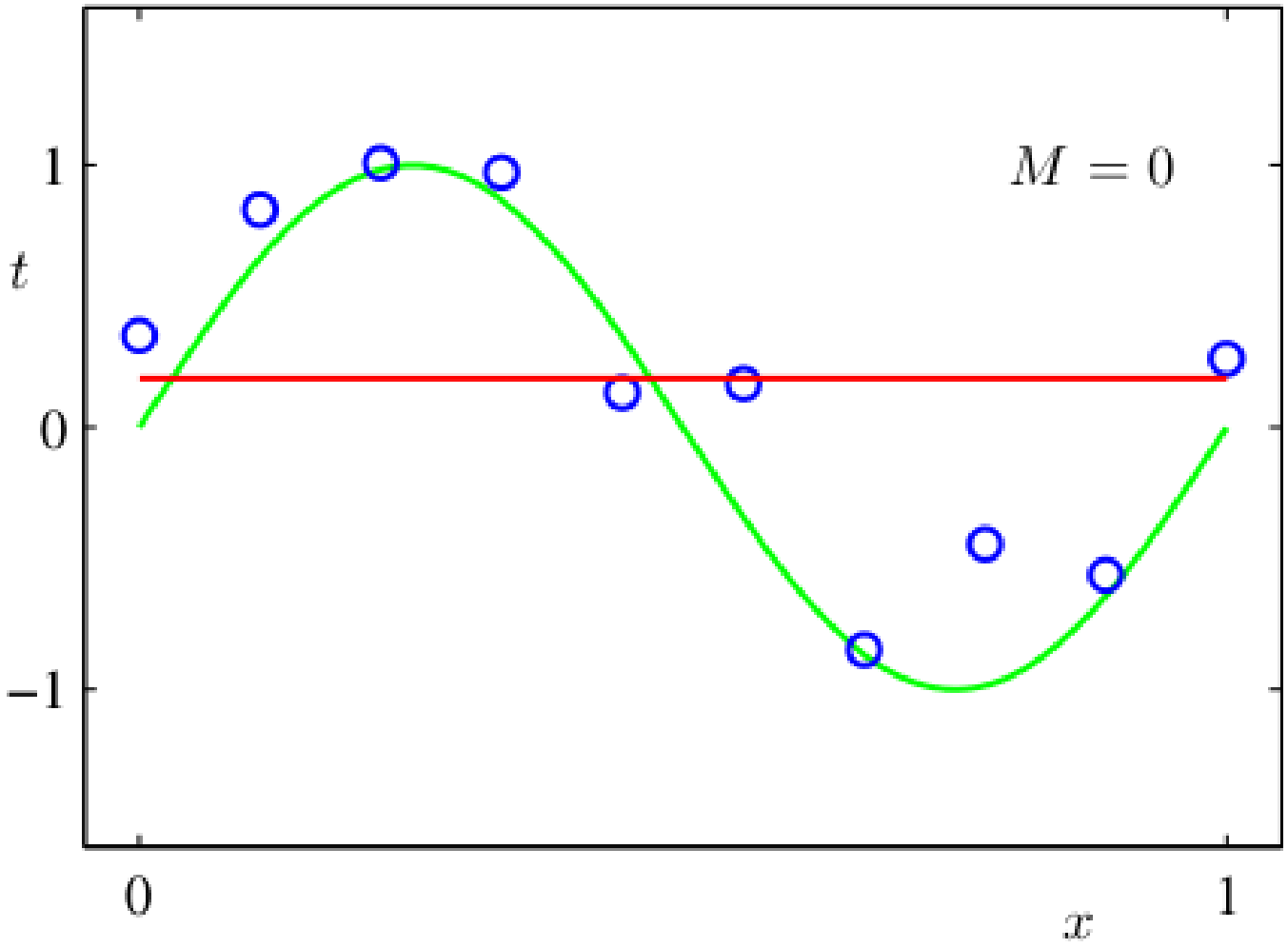
Оверфиттинг

- Прежде всего заметим, что такую задачу все наверняка решали и безо всякого байесовского вывода:
  - выберем функцию ошибки – обычно выбирают среднеквадратическое отклонение

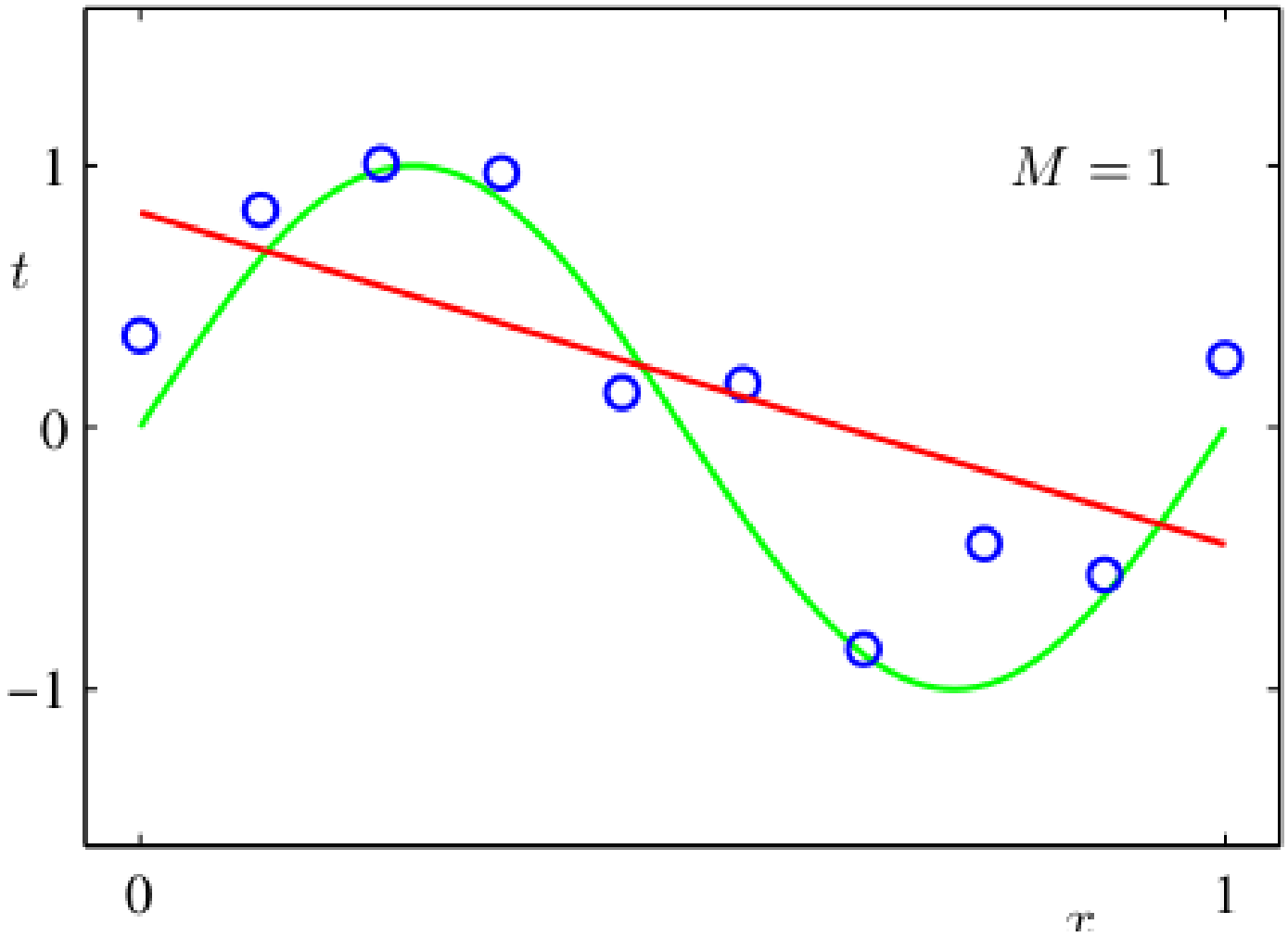
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n, \mathbf{w}) - t_n\right)^2;$$

  - будем её минимизировать – возьмём производные и найдём минимум (или будем к нему двигаться градиентным спуском);
  - в результате получатся оптимальные коэффициенты многочлена данной степени.

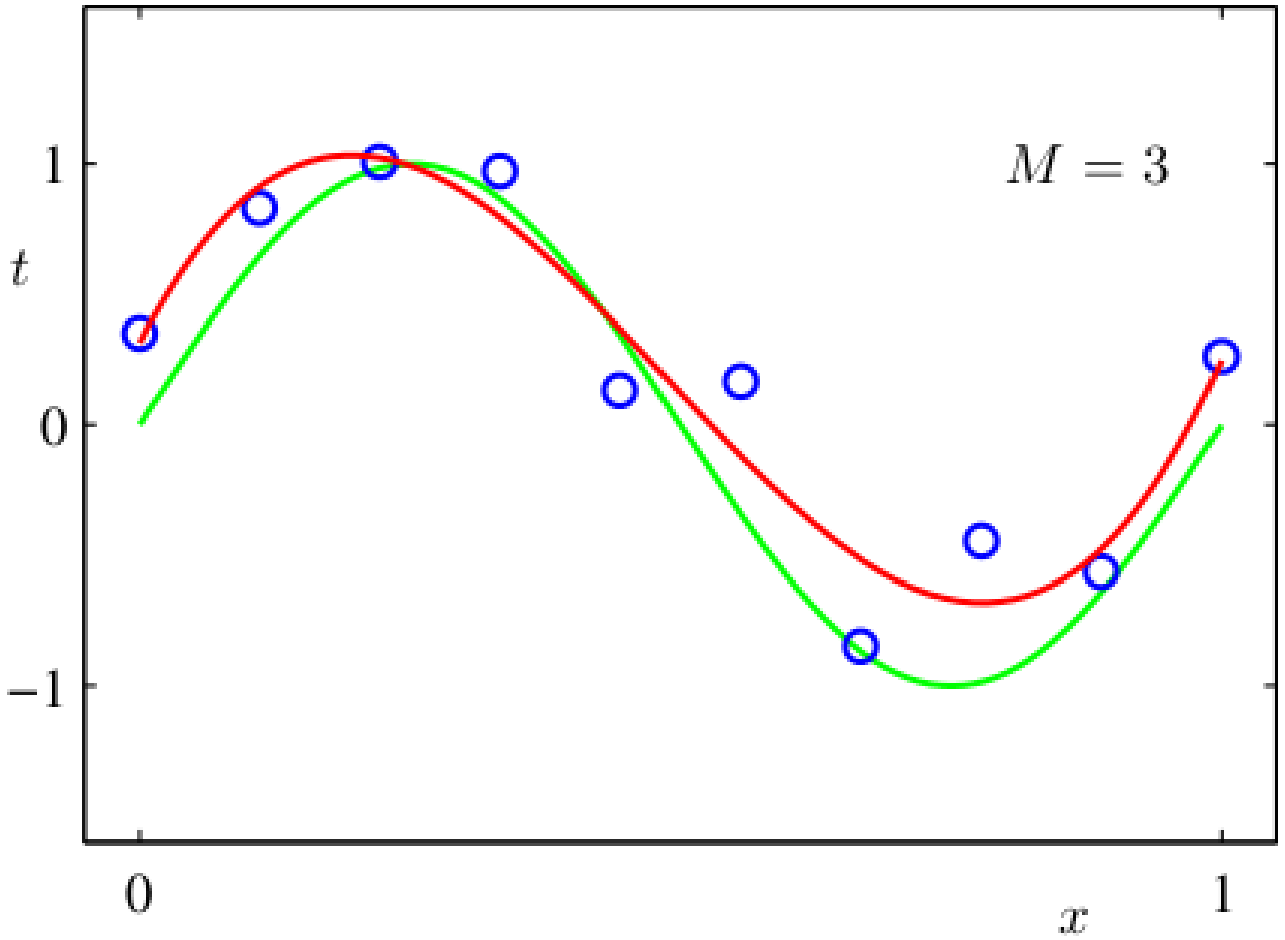
Полиномиальная аппроксимация



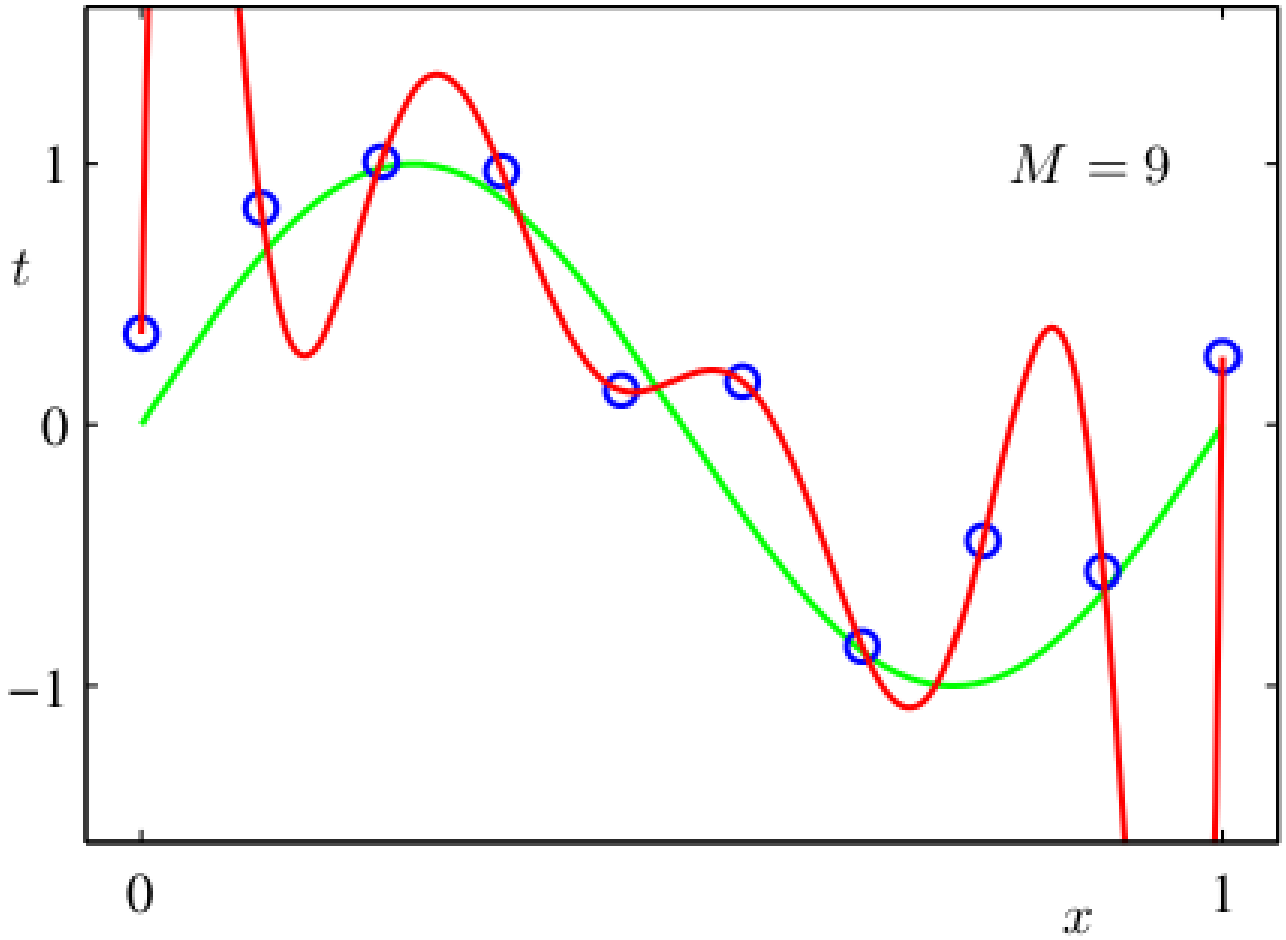
Полиномиальная аппроксимация



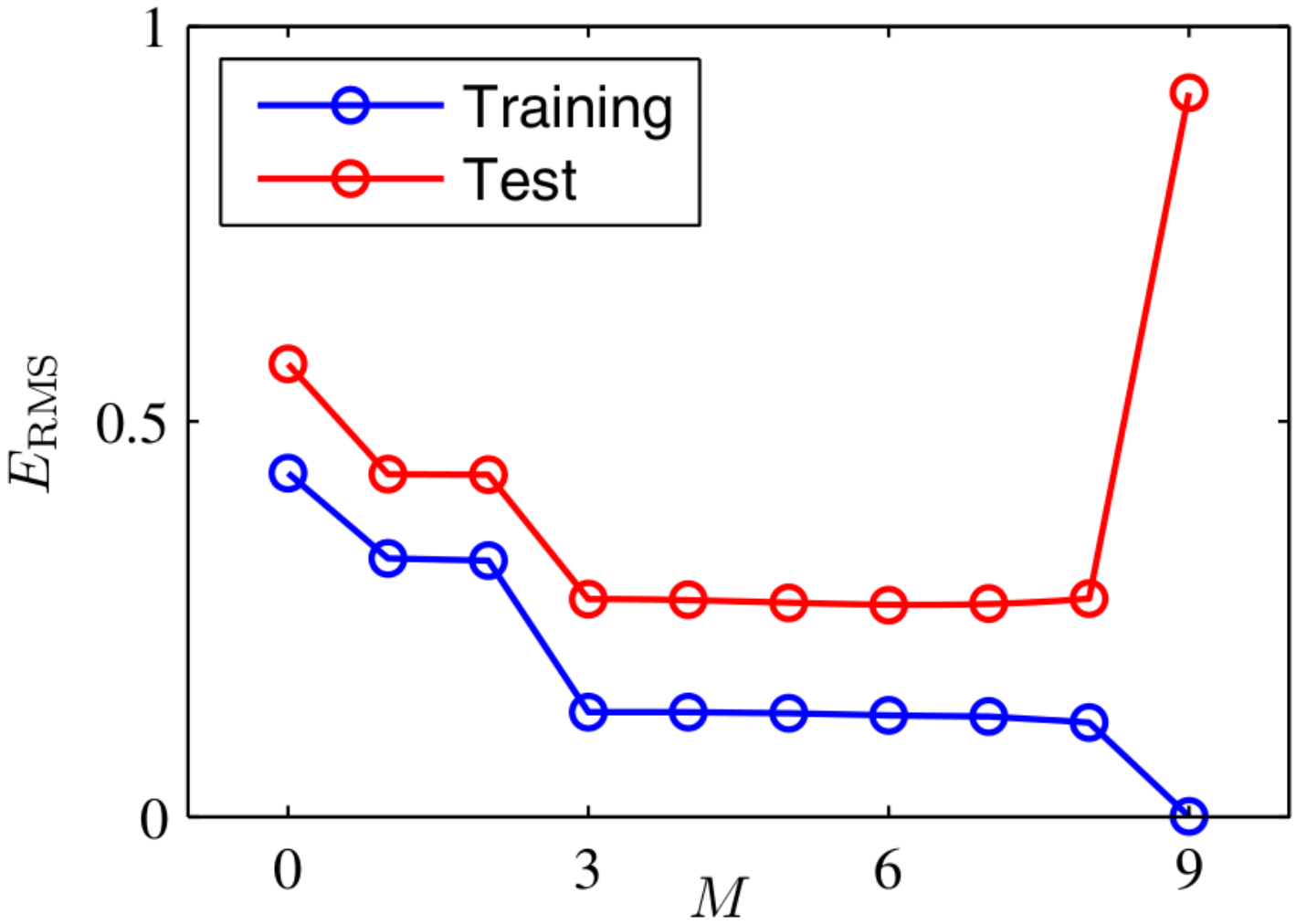
Полиномиальная аппроксимация



Полиномиальная аппроксимация



Значения RMS



Регуляризация

- Обнаружилась проблема – чем больше степень многочлена, тем, конечно, точнее им можно приблизить данные, но в какой-то момент начнётся оверфиттинг; какая модель лучше?

Thank you!

Спасибо за внимание!