SVM

- Данные $(w_1, y_1), \ldots, (w_N, y_N)$, $w_i \in \mathbb{R}^2$, $y_i \in \{+1, -1\}$
- Идея - отделить $+1$ от -1 гиперплоскостью в пространстве $\mathbb{W}$

$$w \cdot x - b = 0$$

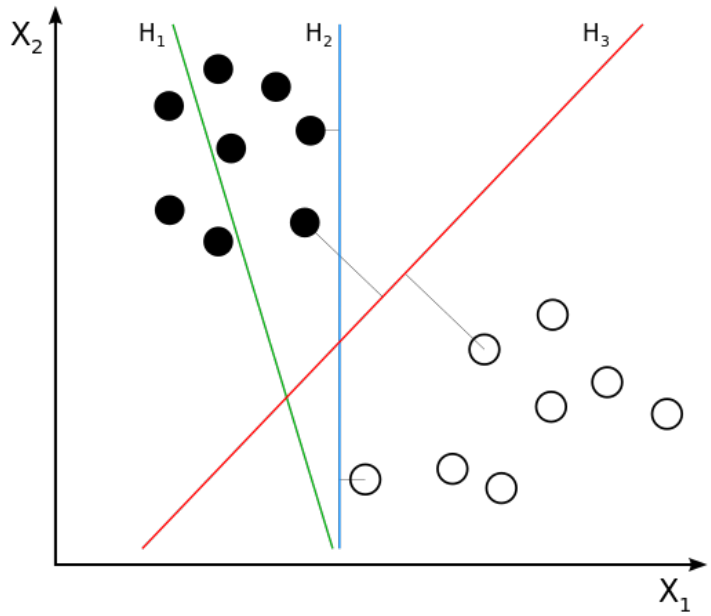
$$w_i \cdot x - b = 1, \quad y_i = 1$$
$$w_i \cdot x - b = -1, \quad y_i = -1$$

$$\begin{cases} ||x||_2 \to \min \\ y_i(w_i \cdot x - b) \geq 1 \end{cases}$$

$$F(x) = \frac{1}{2}||x||_2 -$$

$$- \sum_{i=1}^{N} \alpha_i (y_i(w_i \cdot x - b) - 1) \to$$

$$\to \min_x \max_{\alpha_i \geq 0}$$



Напоминание. Модель

- Модель данных:

$$y = f(w) + \varepsilon,$$

где $y \in \mathbb{Y}$ — выход модели, $w \in \mathbb{W}$ — выход модели, $f : \mathbb{W} \to \mathbb{Y}$ — искомая зависимость, ε — шум

- Нам дано:
  - выборка $D = \{(w_1, y_1), \ldots, (w_N, y_N)\}$,
  - множество функций H ($H = \{h_x(w) : \quad x \in X\}$

Напоминание. Задача

Мы хотим:

1. найти f
2. найти $g \in H$, *объясняющую* выборку D:

$$g(w_i) \sim y_i$$

3. минимизировать функционал качества $F(h|D)$:

$$F(h) = F(h|D) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, h(w_i)) \to \min_{h \in H},$$

где $L(\cdot, \cdot) : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$ — расстояние ($L(a, b) = (a - b)^2$) Полученная ошибка

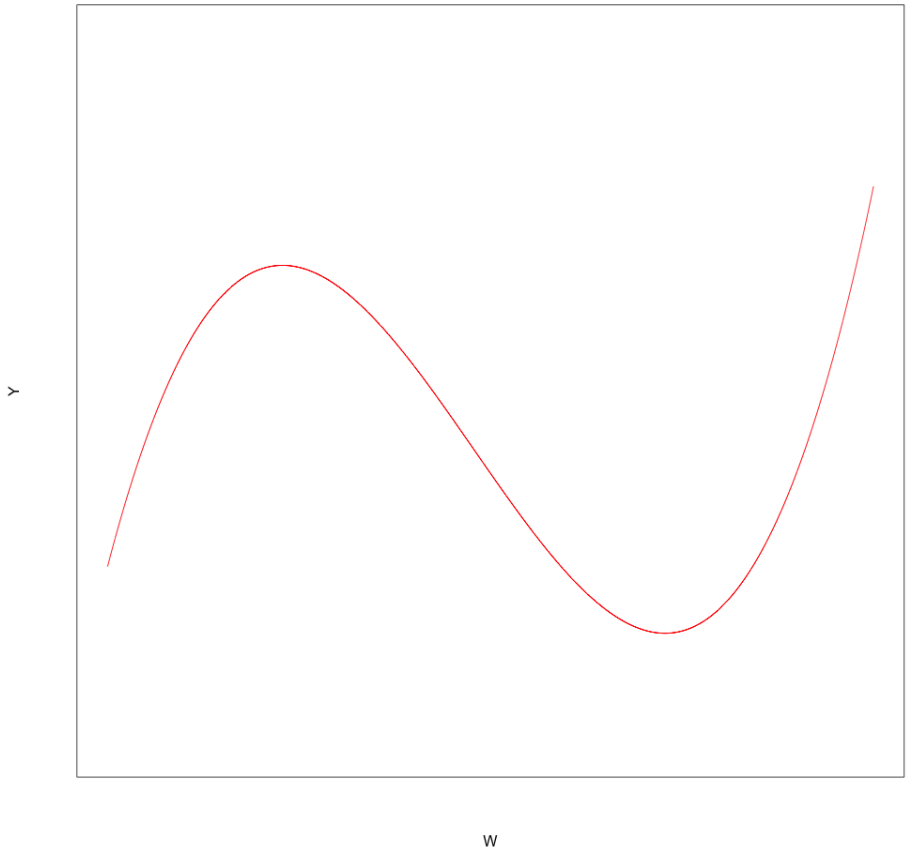
$$error = F(g), \quad \text{где } g = \arg\min_{h \in H} F(h)$$

4. Если *error* маленькая, то все хорошо.

... или не хорошо?

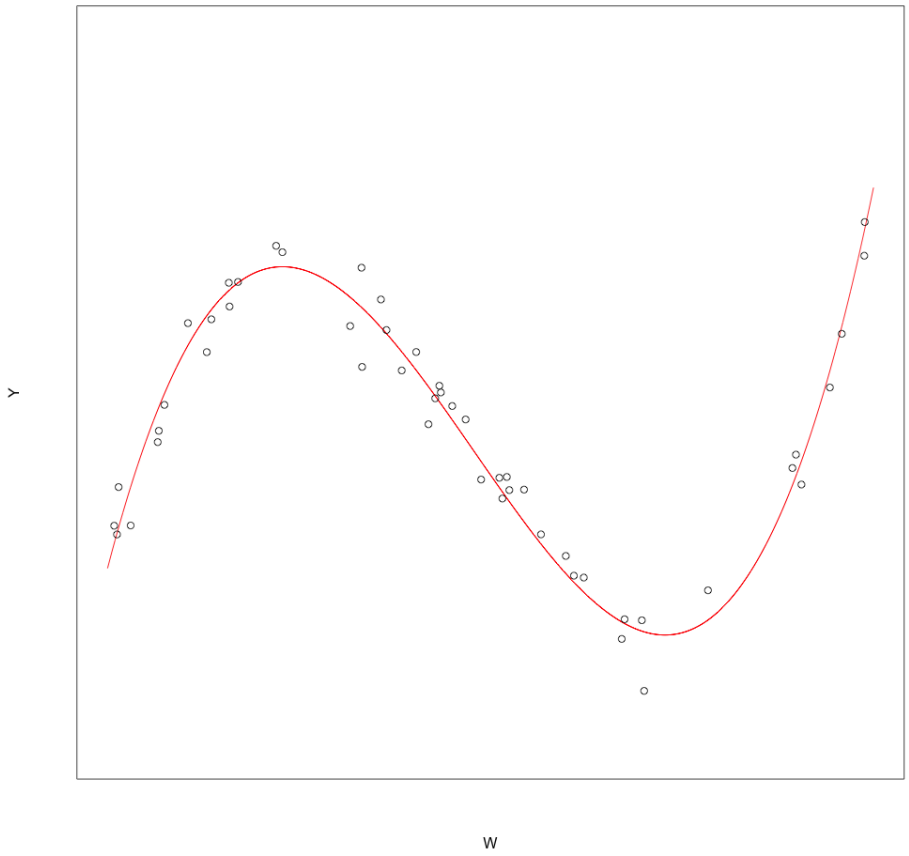
Переобучение. Пример

Искомая функция $f(w)$



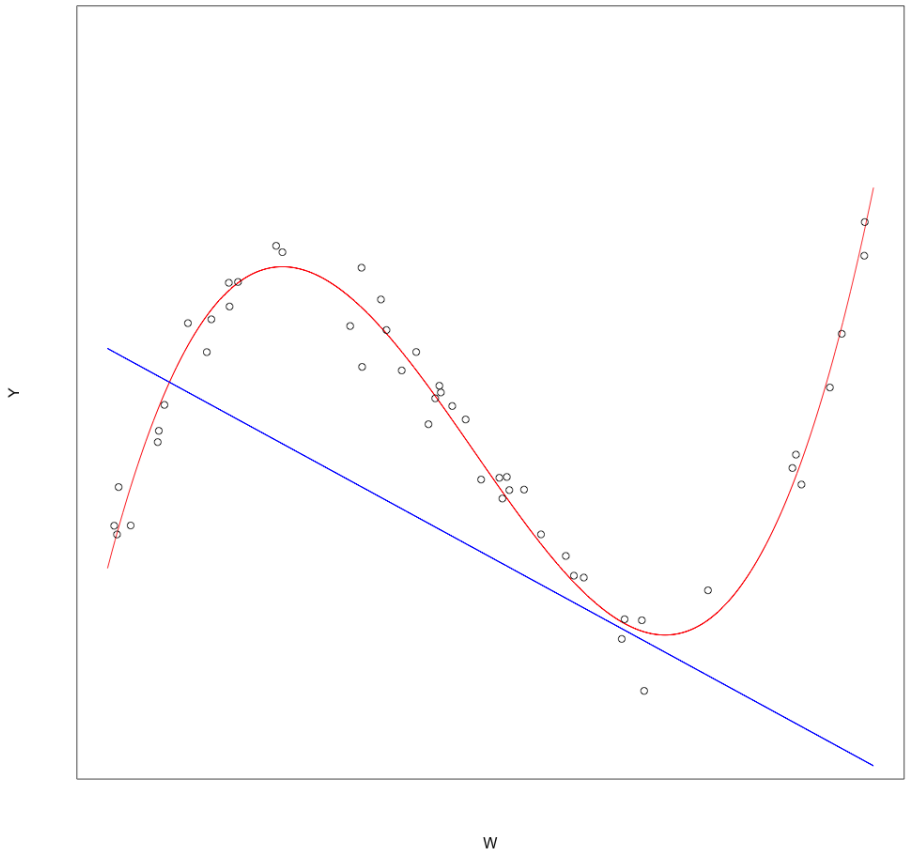
Переобучение. Пример

Добавим шум $\varepsilon \rightsquigarrow$ получим данные D



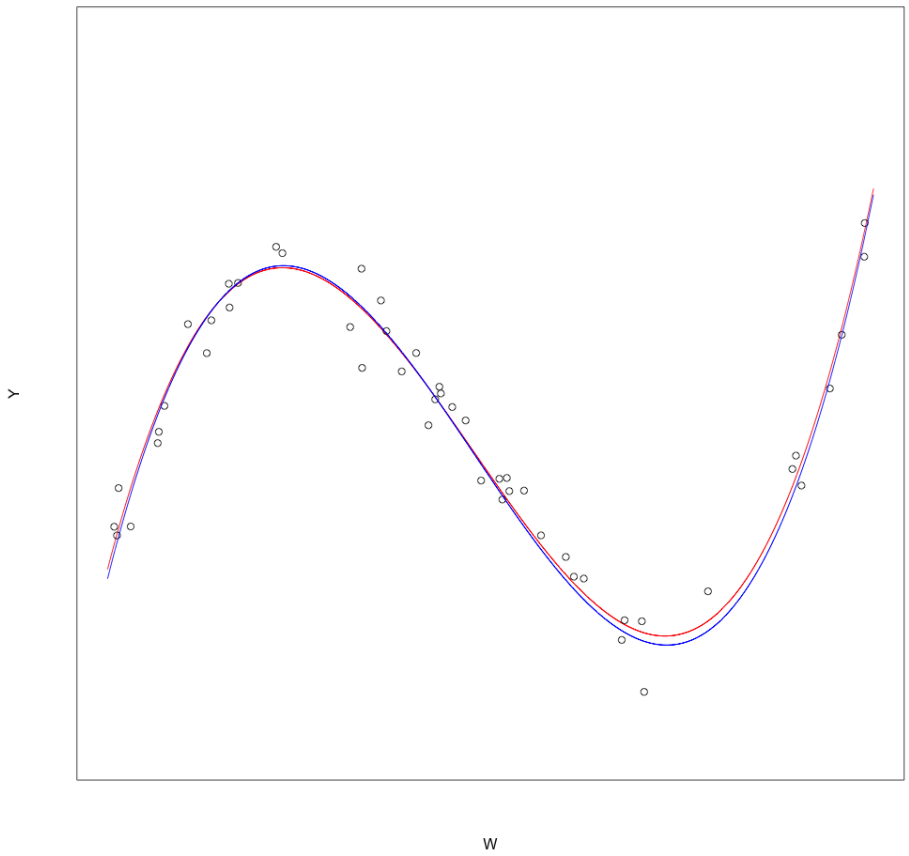
Переобучение. Пример

Попробуем приблизить линейными функциями



Переобучение. Пример

Попробуем приблизить полиномами 3–ей степени

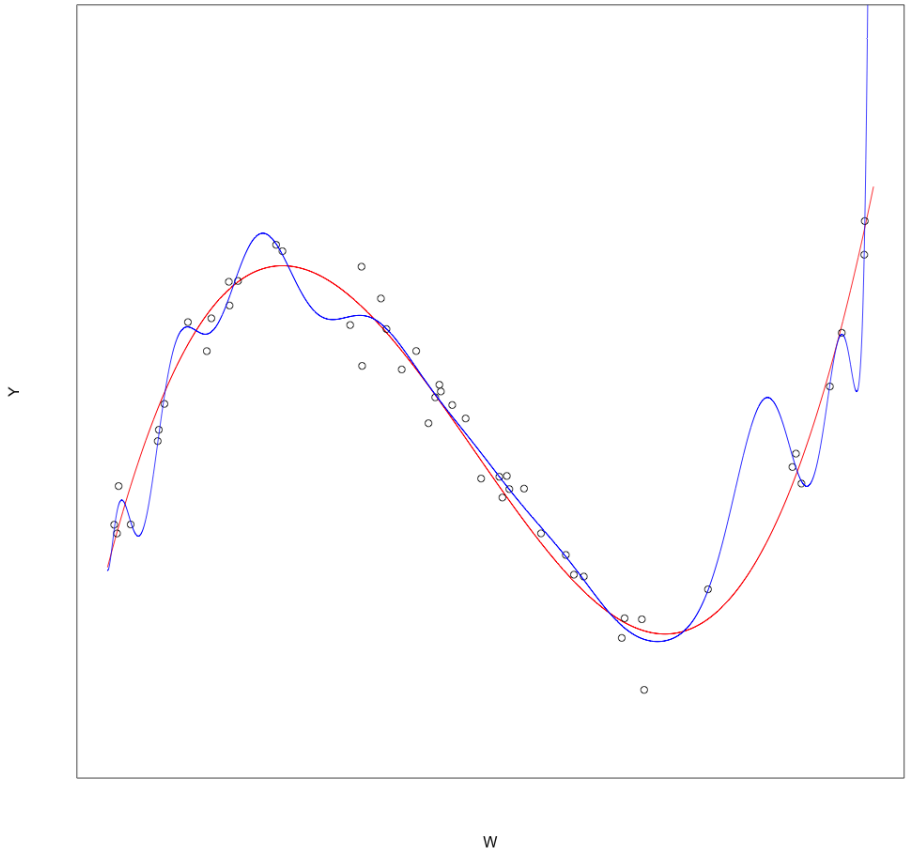


Успех!

Больше — лучше?

Переобучение. Пример

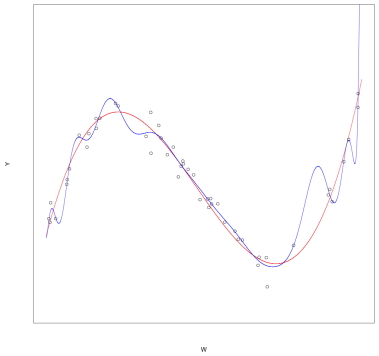
Полином 30–й степени

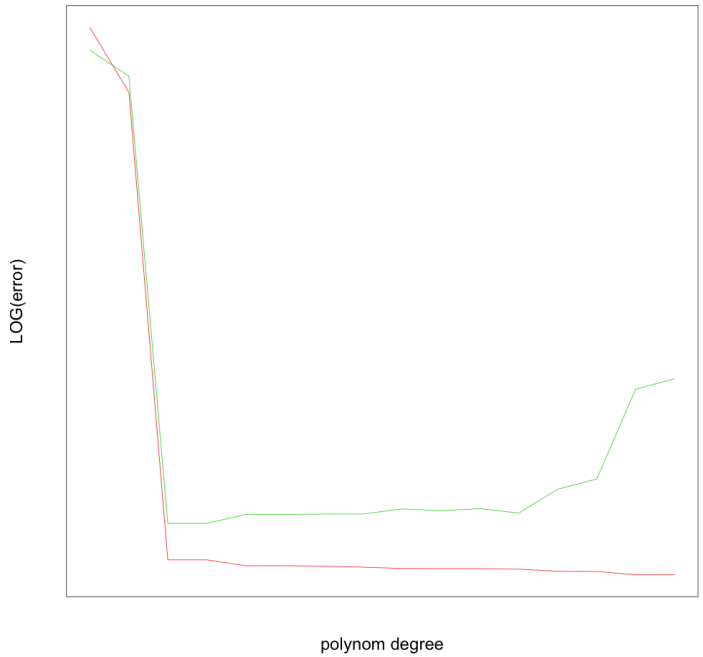


“Перегнули палку”

Переобучение. Неформальное определение

Переобучение (переподгонка, overfitting)
— явление, когда модель хорошо объясняет примеры из обучающей выборки, но значительно хуже объясняет не участвовавшие в обучении примеры (тестовую выборку).





Что делать?

Переобучение. Причины

- Почему так происходит?
- Нужно не только *запоминать* но и *обобщать*.
- Недостаточно

$$F(h|D) \to \min_{h \in H} = error$$

- *error* может не соответствовать тому, как будет вести себя g “в реальности”

. . . что делать?

Переобучение. Методы борьбы

Основные подходы

1. Валидация
   — не только учимся, но и проверяем (валидируем) чему научились
2. Регуляризация
   — добавляем *разумные ограничения*

Валидация. Делим выборку

Валидация бывает разная
Основная идея: учимся на одной части D, а потом проверям качество на другой части D.
Классический подход: делить выборку D на две части:

$$D = D_{train} + D_{test}$$

- D_{train} — тренировочная выборка, на которой обучаем
- D_{test} — тестовая выборка, на которой проверям
- $D_{train} \cap D_{test} = \emptyset$

D_{test} — помогает оценить *error* вне выборки D

Тестовая выборка

1. Ищем наилучшую $h \in H$ на D_{train}

$$g = \arg\min_{h \in H} F(h|D_{train})$$

$$error_{train} = F(g|D_{train}) = \min_{h \in H} F(h|D_{train})$$

2. Проверяем g на D_{test}:

$$error_{test} = F(g|D_{test})$$

3. $error_{test}$ — оценка ошибки вне нашей выборки D

$\rightsquigarrow$ пусть $error_{test} >> error_{train}$, что делать?

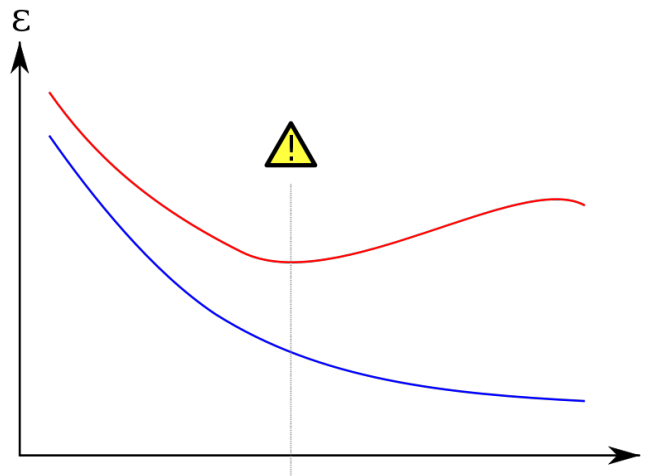
Валидация. Идея

$$D = D_{train} + D_{test} + \mathbf{D_{val}}$$

Минимизация F — итерационный процесс:

$$F(h_0) > F(h_1) > F(h_2) > \ldots > F(h_T)$$

1. На каждой t–итерации меряем $F(h_t|D_{train})$ и $F(h_t|D_{val})$



2. Строим рафик $F(h_t|D_{train})$ и $F(h_t|D_{val})$
3. Выбираем наилучшую итерацию $t_{best} \rightsquigarrow g = h_{t_{best}}$

Валидация. Итог

- больше D_{train} — лучше учимся
- меньше D_{val} и D_{test} — хуже отображают реальную ошибку
- обычно $D_{train} + D_{val} + D_{test} \sim 60\% + 20\% + 20\%$
- или $D_{val} = 20\%$ от D
- часто $D_{train} > D_{val} > D_{test}$

Регуляризация. Идея

- Много параметров $\rightsquigarrow$ сложная модель
- Большие значения параметров $\rightsquigarrow$ сложная модель
- Высокая сложность модели $\rightsquigarrow$ переобучение
- $\rightsquigarrow$ введем штраф за сложность!
  Ограничение

$$F_{reg}(h_x) = F(h_x) + penalty(x)$$
$$g = \arg\min_{h_x} F_{reg}(h_x)$$

Варианты $penalty(x)$:

- $\lambda||x||_0$ — ограничиваем количество параметров
- $\lambda||x||_2$ — ограничиваем значения параметров
- $\lambda||x||_1$ — ограничиваем значения параметров (если параметров много)

Общая идея

Состав ошибки $error = overfiterror + underfiterror$

- $|D_{train}| \nearrow \Rightarrow overfiterror \searrow$ и $underfiterror \searrow$
- сложность модели $\nearrow \Rightarrow overfiterror \nearrow$ и $underfiterror \searrow$

Следовательно

- Чем больше данных D_{train}— тем лучше
- С моделью нужно не перемудрить
- Постоянно проверяем на D_{val}
- Переобучение возникает рано — используем регуляризацию
- “Реальную” ошибку оцениваем по D_{test}

Обучение без учителя

- Кластеризация
- Снижение размерности
  - Principle Component Analysis (PCA)
  - Self-organizing Kohonen maps (SOM)
  - Random Projections
  - Compressive Sensing
  - ...
- Остальное: Hidden Markov Models, structured learning, ...

Конец

Спасибо за внимание!