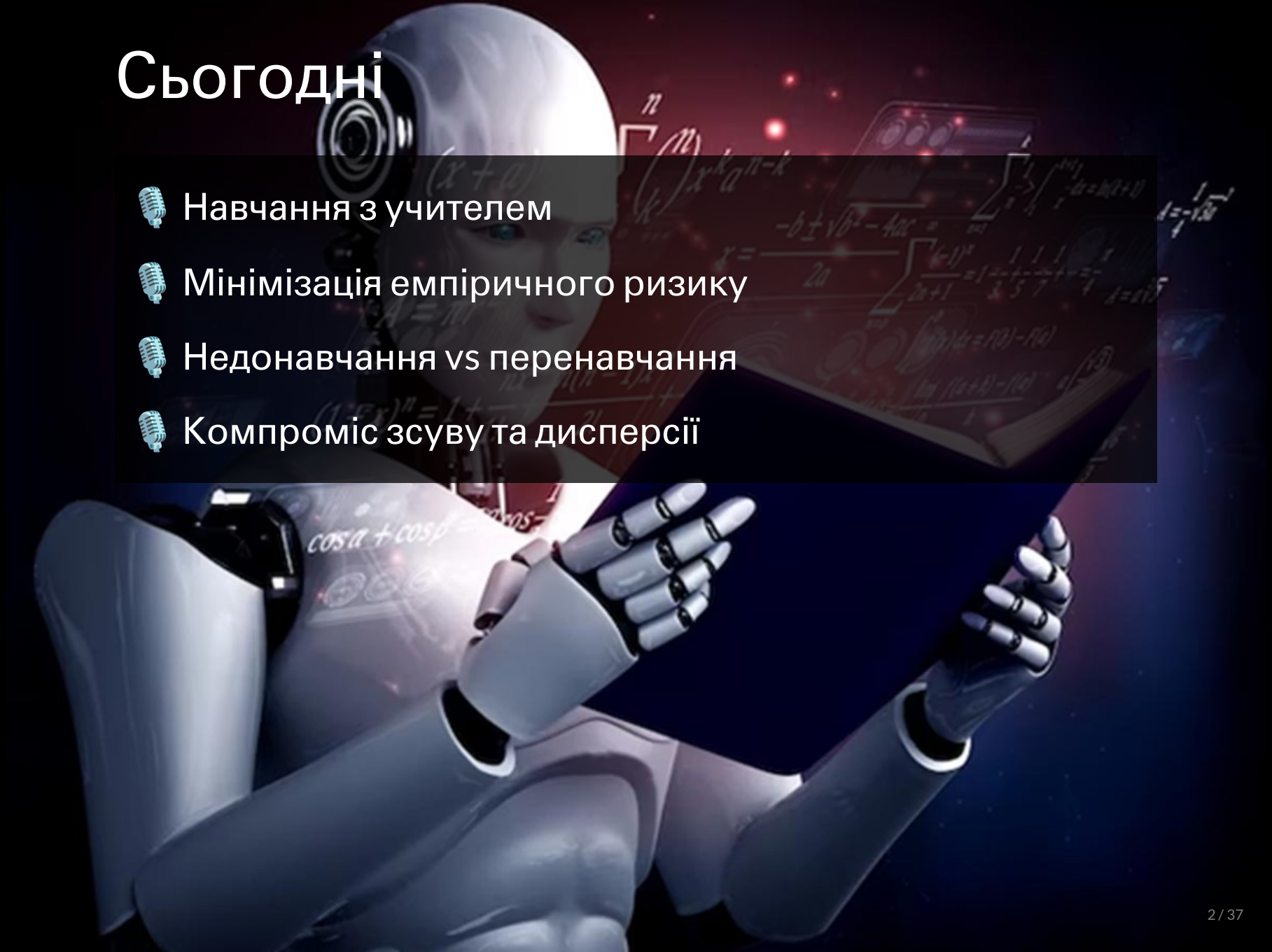
Машинне навчання

Лекція 2: Статистичне навчання

Кочура Юрій Петрович
iuriy.kochura@gmail.com
@y_kochura

Сьогодні

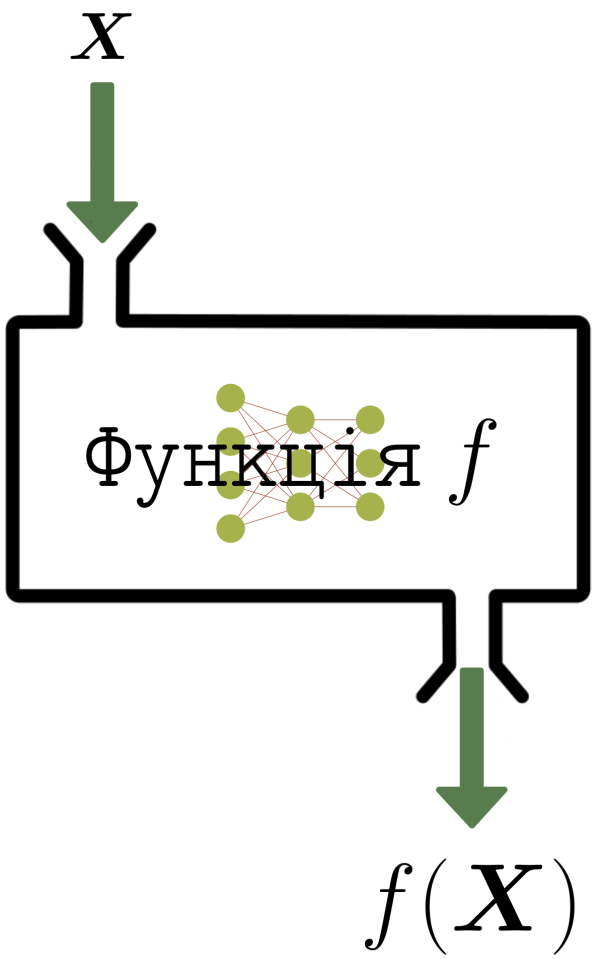


- Навчання з учителем
- Мінімізація емпіричного ризику
- Недонавчання vs перенавчання
- Компроміс зсуву та дисперсії

Навчання з учителем

Статистичне навчання

Модель



Навчання з учителем

Розглянемо невідомий спільний розподіл ймовірності $p_{X,Y}$

Припустимо, навчальна вибірка:

$$(\mathbf{X}^{(i)}, y^{(i)}) \sim p_{X,Y}$$

Навчання з учителем

Розглянемо невідомий спільний розподіл ймовірності $p_{X,Y}$

Припустимо, навчальна вибірка:

$$(\mathbf{X}^{(i)}, y^{(i)}) \sim p_{X,Y}$$

де $\mathbf{X}^{(i)} \in \mathcal{X}, y^{(i)} \in \mathcal{Y}, i = 1, ..., n$.

Навчання з учителем

Розглянемо невідомий спільний розподіл ймовірності $p_{X,Y}$

Припустимо, навчальна вибірка:

$$(\mathbf{X}^{(i)}, y^{(i)}) \sim p_{X,Y}$$

де $\mathbf{X}^{(i)} \in \mathcal{X}, y^{(i)} \in \mathcal{Y}, i = 1, \ldots, n$.

- У більшості випадків
  - $\mathbf{X}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)})$ — m-вимірний вектор ознак або дескрипторів,
  - $y^{(i)}$ — скаляр (наприклад, категорія або дійсне значення).

Навчання з учителем

Розглянемо невідомий спільний розподіл ймовірності $p_{X,Y}$

Припустимо, навчальна вибірка:

$$(\mathbf{X}^{(i)}, y^{(i)}) \sim p_{X,Y}$$

де $\mathbf{X}^{(i)} \in \mathcal{X}, y^{(i)} \in \mathcal{Y}, i = 1, \ldots, n$.

- У більшості випадків
  - $\mathbf{X}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)})$ — m-вимірний вектор ознак або дескрипторів,
  - $y^{(i)}$ — скаляр (наприклад, категорія або дійсне значення).
- Дані навчання згенеровано як незалежні та однаково розподілені випадкові величини

Навчання з учителем

Розглянемо невідомий спільний розподіл ймовірності $p_{X,Y}$

Припустимо, навчальна вибірка:

$$(\mathbf{X}^{(i)}, y^{(i)}) \sim p_{X,Y}$$

де $\mathbf{X}^{(i)} \in \mathcal{X}, y^{(i)} \in \mathcal{Y}, i = 1, \ldots, n$.

- У більшості випадків
  - $\mathbf{X}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)})$ — m-вимірний вектор ознак або дескрипторів,
  - $y^{(i)}$ — скаляр (наприклад, категорія або дійсне значення).
- Дані навчання згенеровано як незалежні та однаково розподілені випадкові величини
- Навчальна вибірка може мати будь-який кінцевий розмір n.

Навчання з учителем

Розглянемо невідомий спільний розподіл ймовірності $p_{X,Y}$

Припустимо, навчальна вибірка:

$$(\mathbf{X}^{(i)}, y^{(i)}) \sim p_{X,Y}$$

де $\mathbf{X}^{(i)} \in \mathcal{X}, y^{(i)} \in \mathcal{Y}, i = 1, \ldots, n$.

- У більшості випадків
  - $\mathbf{X}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)})$ — m-вимірний вектор ознак або дескрипторів,
  - $y^{(i)}$ — скаляр (наприклад, категорія або дійсне значення).
- Дані навчання згенеровано як незалежні та однаково розподілені випадкові величини
- Навчальна вибірка може мати будь-який кінцевий розмір n.
- У загальному випадку у нас немає жодної попередньої інформації про $p_{X,Y}$.

Висновок моделі

Навчання з учителем зазвичай стосується двох наступних типів задач логічного висновку моделі:

- Класифікація
  - Дано $(\mathbf{X}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^m \times \triangle^C$, де $i = 1, \ldots, n$
  - Ми хочемо оцінити для будь-якого нового $\mathbf{X}$:

$$\boxed{\arg\max_y p(Y = y|\mathbf{X})}$$

Висновок моделі

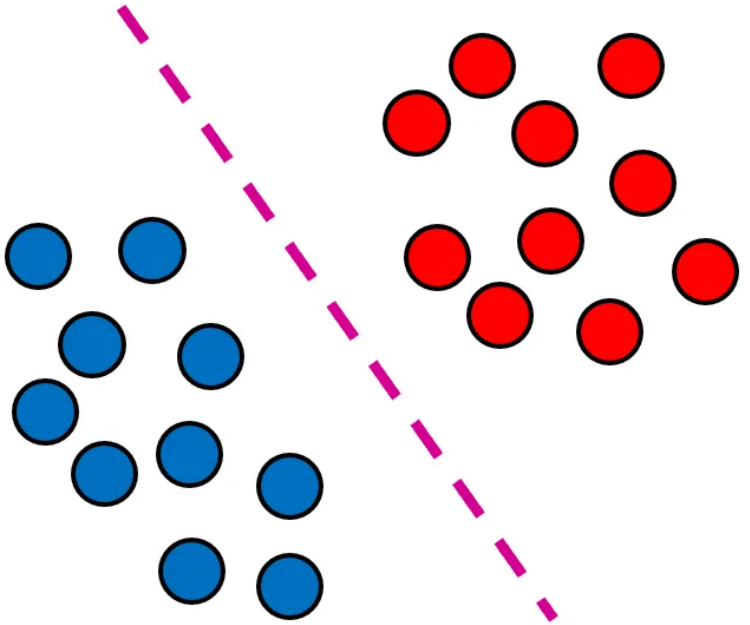
Навчання з учителем зазвичай стосується двох наступних типів задач логічного висновку моделі:

- Класифікація
  - Дано $(\mathbf{X}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^m \times \triangle^C$, де $i = 1, \ldots, n$
  - Ми хочемо оцінити для будь-якого нового $\mathbf{X}$:

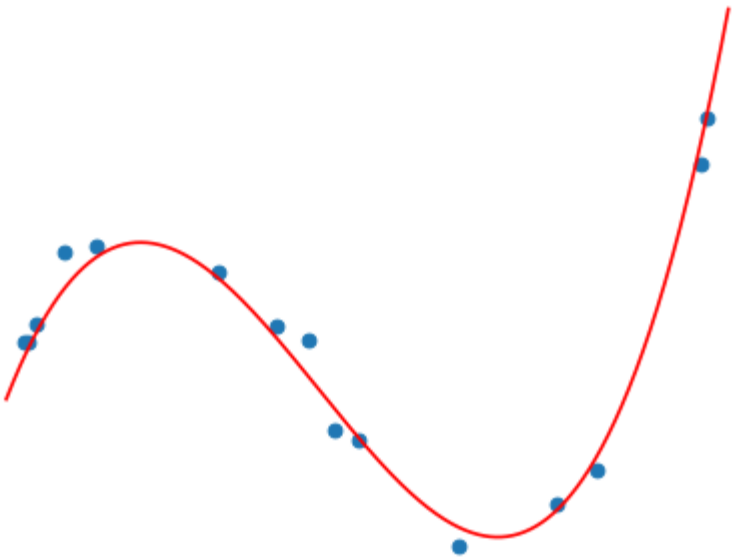
$$\boxed{\arg\max_y p(Y = y|\mathbf{X})}$$

- Регресія
  - Дано $(\mathbf{X}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^m \times \mathbb{R}$, де $i = 1, \ldots, n$
  - Ми хочемо оцінити для будь-якого нового $\mathbf{X}$:

$$\mathbb{E}\left[Y|\mathbf{X}\right]$$



Класифікація полягає у визначенні
межі (границі) рішення між об’єктами різних класів.



Регресія намагається оцінити в взаємозв'язок між скалярною (зазвичай неперервною) залежною змінною та однією або кількома незалежними змінними.

Приклад: Прогнозування цін на житло

Імовірнісна постановка задачі

Навчання з учителем можна формалізувати як імовірнісний висновок моделі, метою якого є оцінка умовного розподілу

$$p(Y = y|\mathbf{X})$$

для будь-якої нової пари $(\mathbf{X}, y)$.

Мінімізація емпіричного ризику

Традиційний підхід навчання з учителем полягає в мінімізації емпіричного ризику моделі.

Розглянемо функцію $f : \mathcal{X} \to \mathcal{Y}$, породжену деяким алгоритмом навчання. Якість прогнозів цієї моделі можна оцінити за допомогою функції втрат

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R},$$

де $\ell(y, f(\mathbf{X})) \geq 0$ вимірює, наскільки близьким є передбачення моделі $f(\mathbf{X})$ до епіричного значення y.

Приклади функцій втрат

Класифікація: $\ell(y, f(\mathbf{X})) = \mathbf{1}_{y \neq f(\mathbf{X})}$

Регресія: $\ell(y, f(\mathbf{X})) = (y - f(\mathbf{X}))^2$

Нехай $\mathcal{F}$ — простір гіпотез, тобто множина усіх функцій f, які можуть бути породжені вибраним алгоритмом навчання.

Ми шукаємо функцію $f \in \mathcal{F}$ з невеликим середнім ризиком (або помилкою узагальнення)

$$R(f) = \mathbb{E}_{(\mathbf{x},y)\sim p_{X,Y}} \left[\ell(y, f(\mathbf{X}))\right]$$

Це означає, що для заданого розподілу даних $p_{X,Y}$, і для заданого простору гіпотез $\mathcal{F}$, оптимальна модель

$$f_* = \arg\min_{f\in\mathcal{F}} R(f)$$

Оскільки $p_{X,Y}$ невідомий, неможливо оцінити середній ризик і визначити оптимальну модель.

Однак, якщо ми маємо незалежні та однаково розподілені навчальні дані $\mathbf{d} = \{(\mathbf{X}^{(i)}, y^{(i)}) | i = 1, \ldots, n\}$, ми можемо обчислити оцінку, емпіричного ризику (або помилку навчання)

$$\hat{R}(f, \mathbf{d}) = \frac{1}{n} \sum_{(\mathbf{X}^{(i)}, y^{(i)}) \in \mathbf{d}} \ell\left(y^{(i)}, f(\mathbf{X}^{(i)})\right)$$

Ця оцінка є незміщеною і може бути використана для знаходження достатньо хорошого наближення f_*. Це призводить до принципу мінімізації емпіричного ризику:

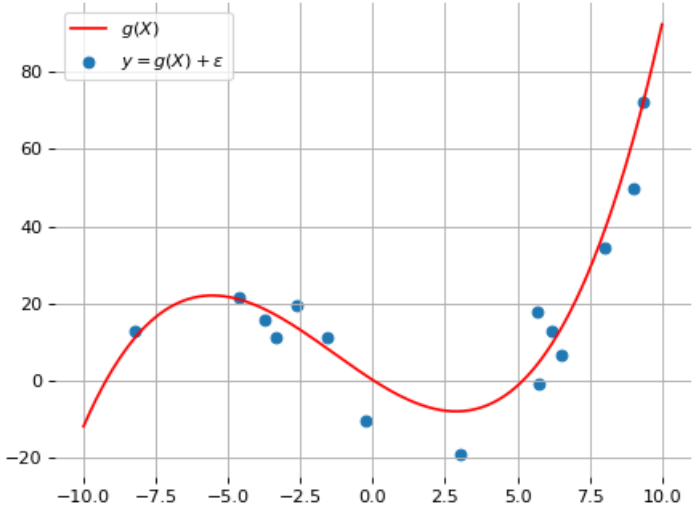
$$f_*^{\mathbf{d}} = \arg\min_{f \in \mathcal{F}} \hat{R}(f, \mathbf{d})$$

Більшість алгоритмів машинного навчання, включаючи нейронні мережі, реалізують мінімізацію емпіричного ризику.

За законом великих чисел, мінімізація емпіричного ризику збігається:

$$\lim_{n\to\infty} f_*^{\mathbf{d}} = f_*$$

Поліноміальна регресія



Розглянемо спільний розподіл імовірностей $p_{X,Y}$, породжений процесом генерування даних

$$(\mathbf{X}, y) \sim p_{X,Y} \Leftrightarrow \mathbf{X} \sim U[-10; 10], \varepsilon \sim \mathcal{N}(0, \sigma^2), y = g(x) + \varepsilon$$

де $\mathbf{X} \in \mathbb{R}, y \in \mathbb{R}$ та g — невідомий поліном степеня 3.

Наша мета — знайти функцію f, яка робить в середньому хороші прогнози для розподілу $p_{X,Y}$.

Розглянемо простір гіпотез $f \in \mathcal{F}$ для поліномів степеня 3, визначених через навчальні параметри $\mathbf{w} \in \mathbb{R}^4$, наступним чином

$$\hat{y} \triangleq f(\mathbf{X}; \mathbf{w}) = \sum_{d=0}^{3} w_d X^d$$

Для цієї задачі регресії ми використовуємо квадратичне відхилення як функцію втрат

$$\ell(y, f(x; \mathbf{w})) = (y - f(\mathbf{X}; \mathbf{w}))^2$$

щоб визначити, наскільки помилковими є прогнози.

Тому нашою метою є знаходження найкращого значення $\mathbf{w}_*$ такого, що

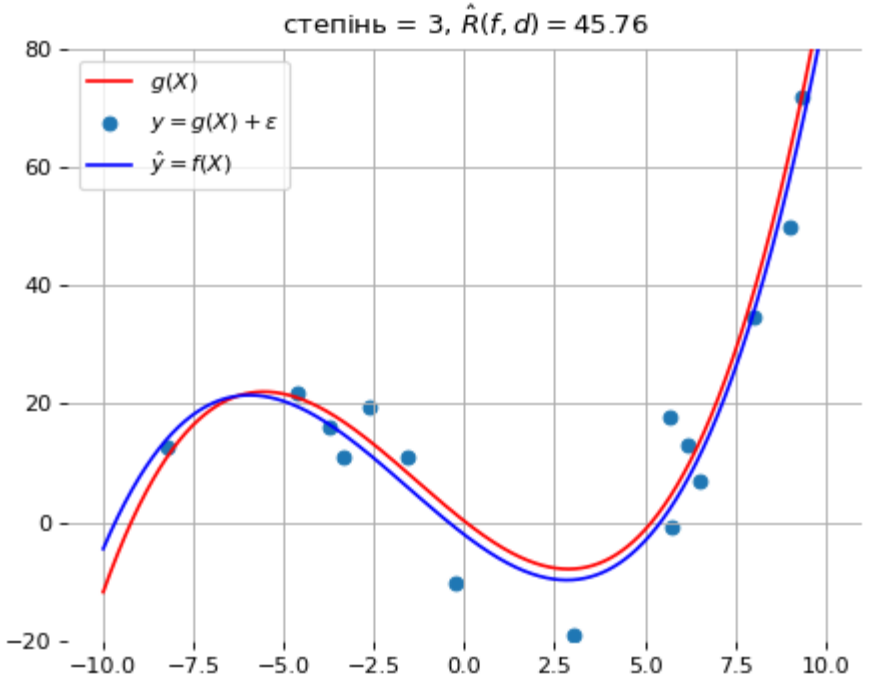
$$\begin{aligned}\mathbf{w}_* &= \arg\min_{\mathbf{w}} R(\mathbf{w}) = \\ &= \arg\min_{\mathbf{w}} \mathbb{E}_{(\mathbf{X}, y) \sim p_{X,Y}} \left[(y - f(\mathbf{X}; \mathbf{w}))^2\right]\end{aligned}$$

Для досить великої навчальної вибірки $\mathbf{d} = \{(\mathbf{X}^{(i)}, y^{(i)}) | i = 1, \dots, n\}$ принцип мінімізації емпіричного ризику говорить нам, що гарну оцінку $\mathbf{w}_*^{\mathbf{d}}$ з $\mathbf{w}_*$ можна знайти шляхом мінімізації емпіричного ризику:

$$
\begin{aligned}
\mathbf{w}_*^{\mathbf{d}} &= \arg\min_{\mathbf{w}} \hat{R}(\mathbf{w}, \mathbf{d}) = \\
&= \arg\min_{\mathbf{w}} \frac{1}{n} \sum_{(\mathbf{X}^{(i)}, y^{(i)}) \in \mathbf{d}} (y^{(i)} - f(\mathbf{X}^{(i)}; \mathbf{w}))^2 = \\
&= \arg\min_{\mathbf{w}} \frac{1}{n} \sum_{(\mathbf{X}^{(i)}, y^{(i)}) \in \mathbf{d}} (y^{(i)} - \sum_{d=0}^{3} w_d x^{d(i)})^2 = \\
&= \arg\min_{\mathbf{w}} \frac{1}{n} \left\| \underbrace{\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(n)} \end{pmatrix}}_{\mathbf{y}} - \underbrace{\begin{pmatrix} x^{0(1)} & \dots & x^{3(1)} \\ x^{0(2)} & \dots & x^{3(2)} \\ & \dots & \\ x^{0(n)} & \dots & x^{3(n)} \end{pmatrix}}_{\mathbf{X}} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix} \right\|^2
\end{aligned}
$$

Це звичайна регресія за методом найменших квадратів, для якої аналітичний розв'язок:

$$\mathbf{w}_*^{\mathbf{d}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

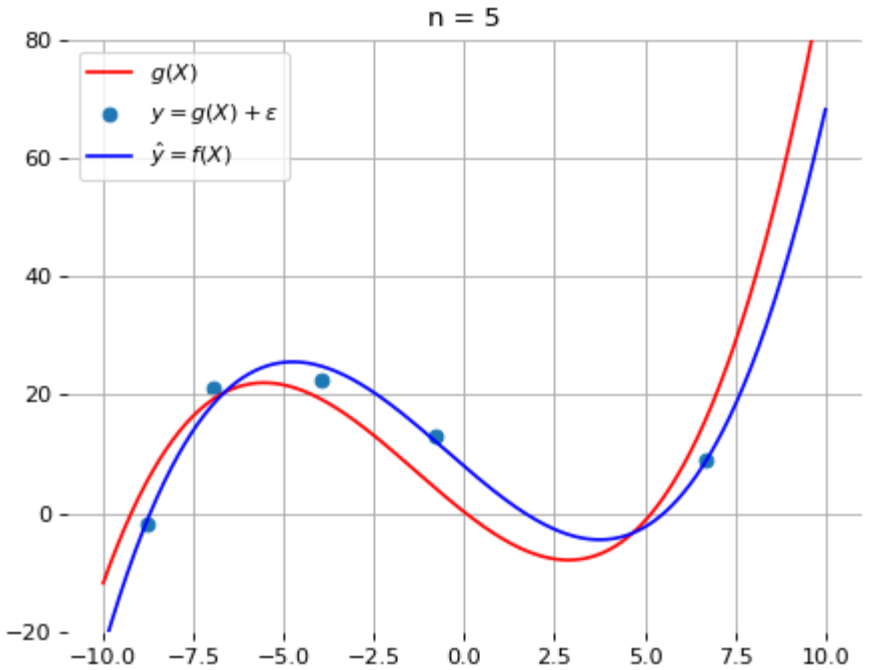


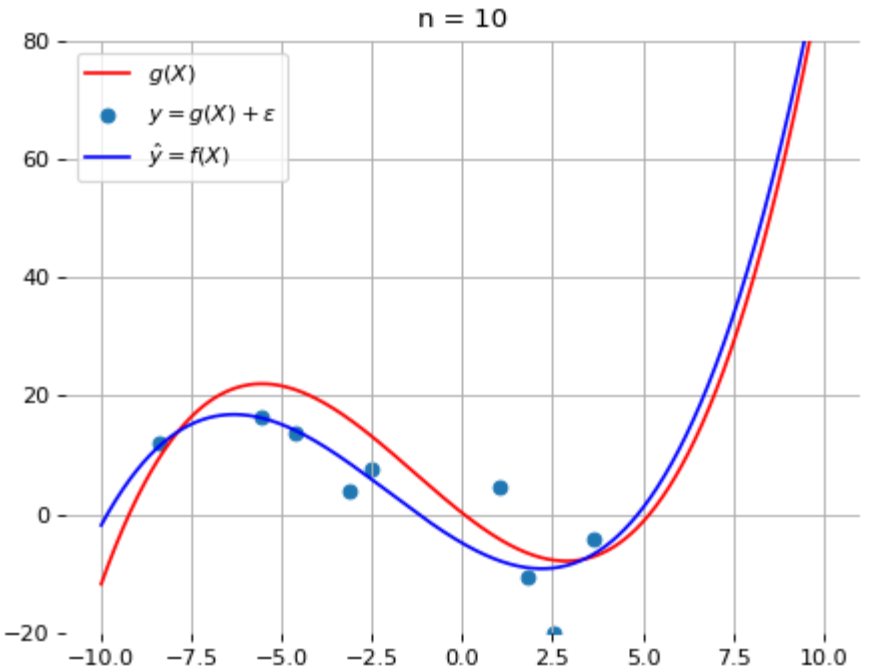
Найкращим мінімізатором ризику $\mathbf{w}_*$ у просторі наших гіпотез є g.

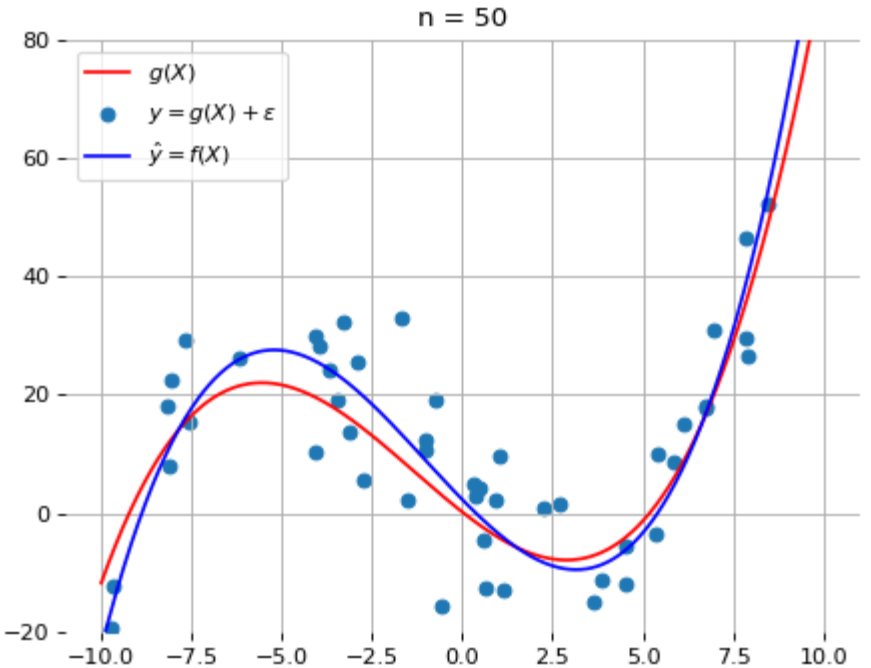
Отже, ми можемо перевірити наступне:

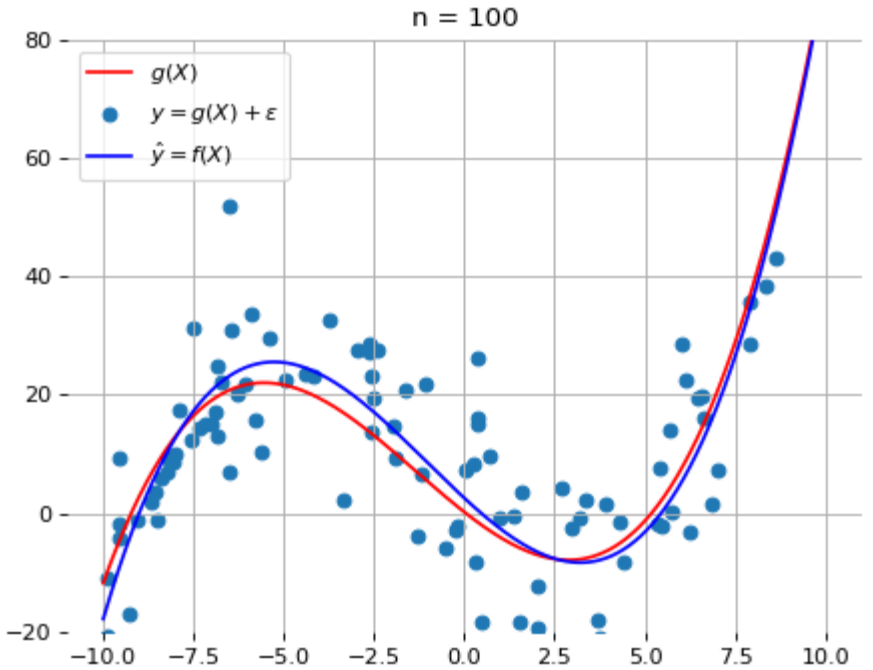
$$f(X; \mathbf{w}_*^{\mathbf{d}}) \to f(X; \mathbf{w}_*) = g(X)$$

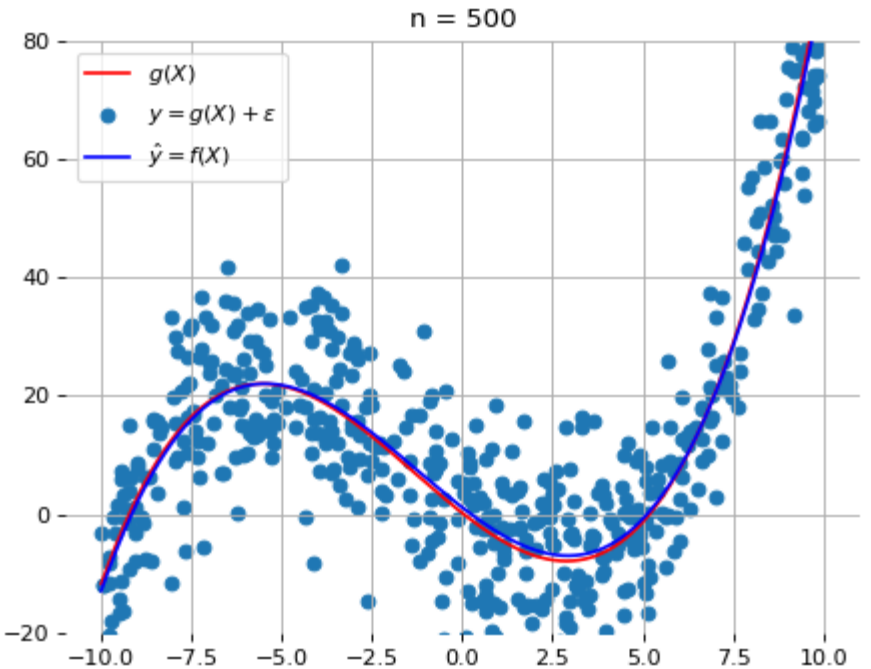
при $n \to \infty$


n = 5
$g(X)$
$y = g(X) + \varepsilon$
$\hat{y} = f(X)$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


n = 10
$g(X)$
$y = g(X) + \varepsilon$
$\hat{y} = f(X)$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

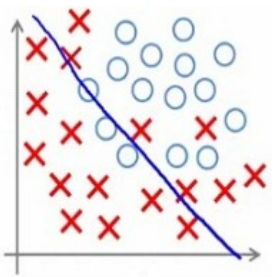

n = 50
$g(X)$
$y = g(X) + \varepsilon$
$\hat{y} = f(X)$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


n = 100
$g(X)$
$y = g(X) + \varepsilon$
$\hat{y} = f(X)$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


n = 500
$g(X)$
$y = g(X) + \varepsilon$
$\hat{y} = f(X)$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

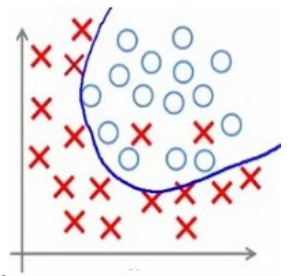
Недонавчання **vs** перенавчання

Що буде, якщо ми розглянемо простір гіпотез $\mathcal{F}$, у якому функції-кандидати f або надто «прості», або надто «складні» по відношенню до справжнього розподілу даних?

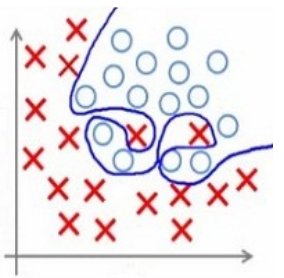


Недонавчання

Занадто проста модель, щоб пояснити розкид

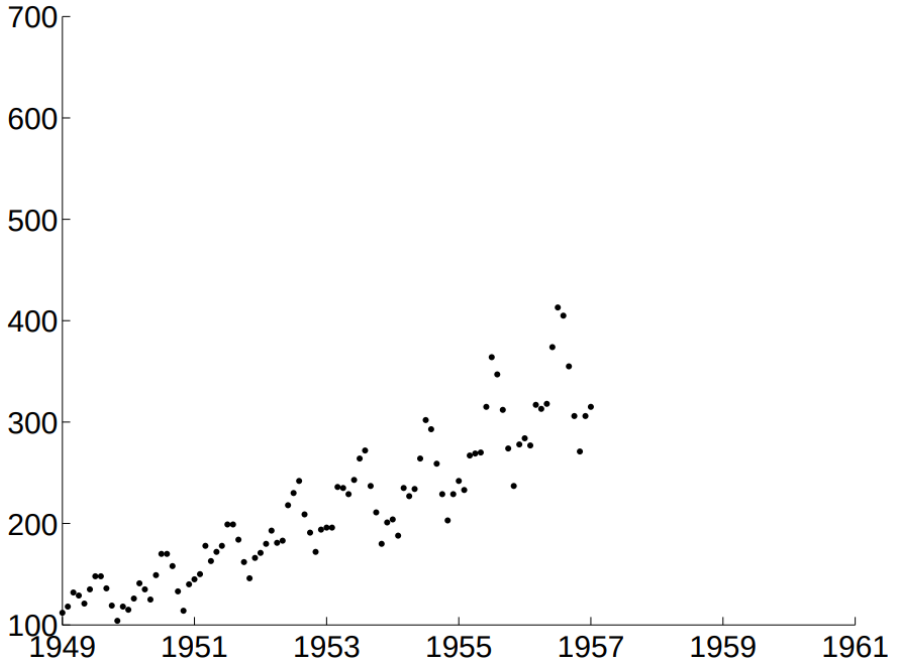


Належний рівень навчання



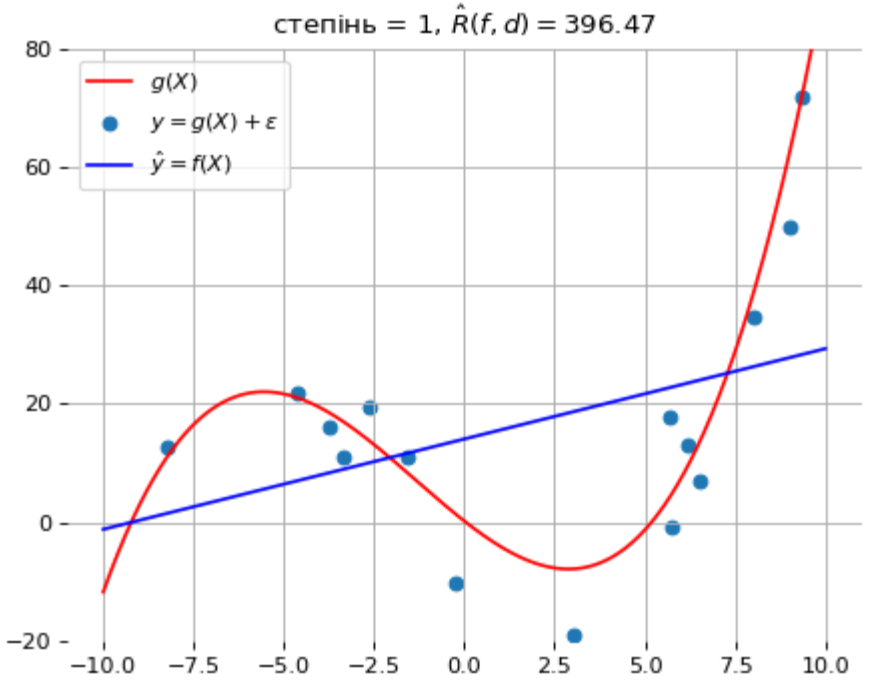
Перенавчання

Занадто складна модель, враховує шум, замість взаємозв'язку, що лежить в основі даних

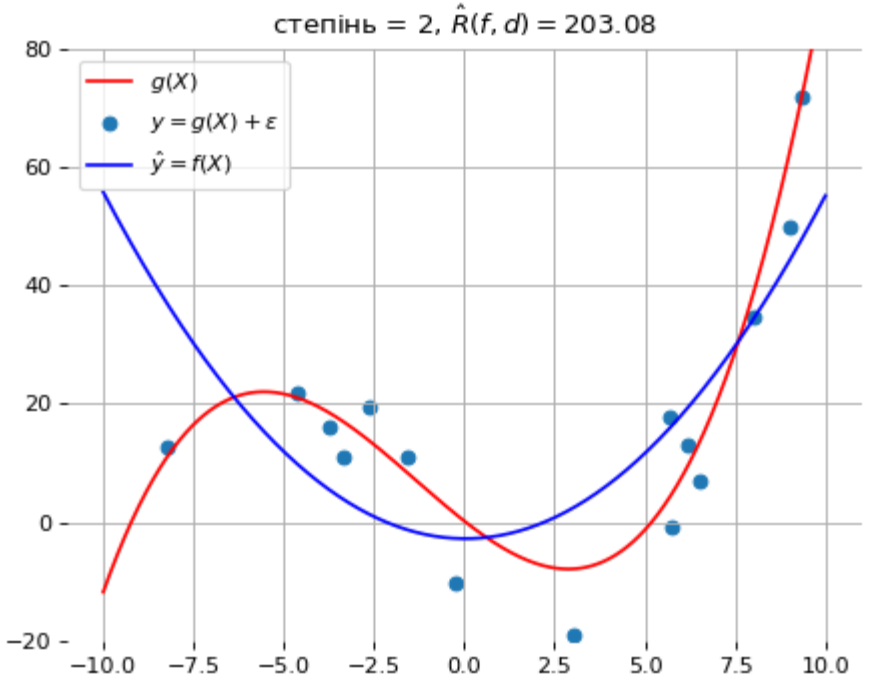


Яку модель Ви б обрали?

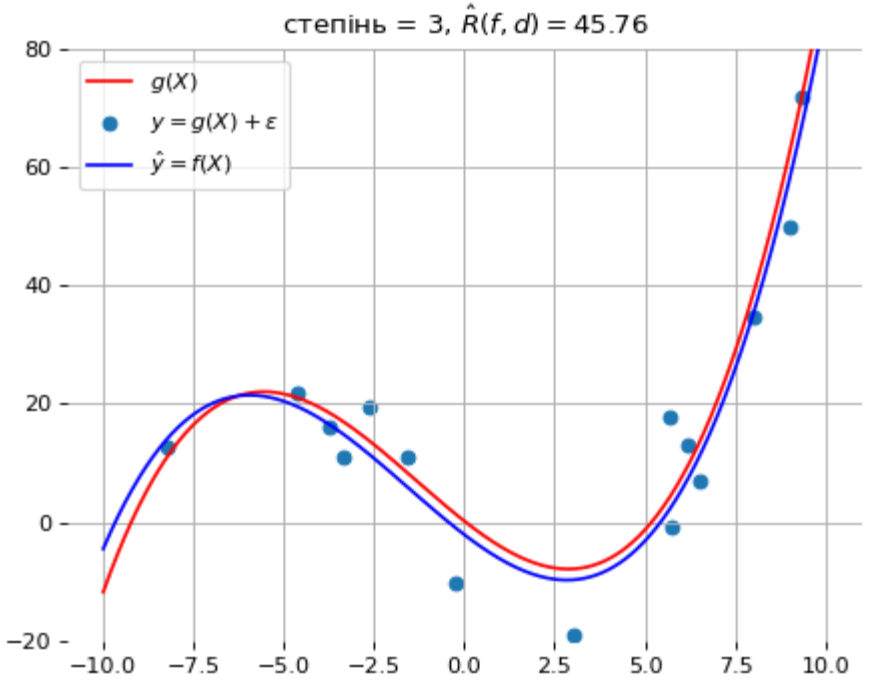
$f_1(x) = w_0 + w_1 x$ $f_2(x) = \sum_{j=0}^{3} w_j x^j$ $f_3(x) = \sum_{j=0}^{10^4} w_j x^j$



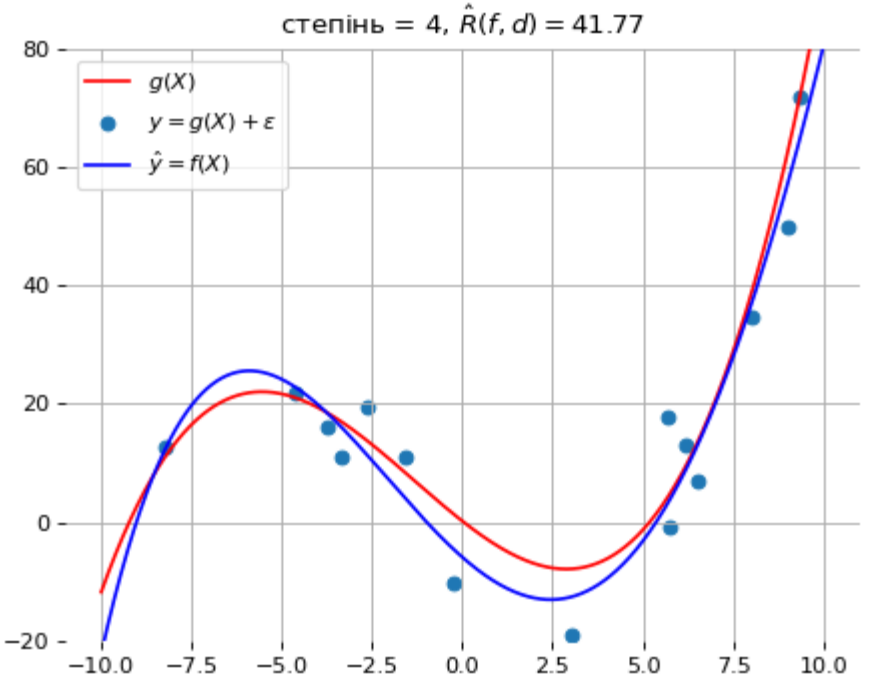
$\mathcal{F}$ = поліноми степеня 1



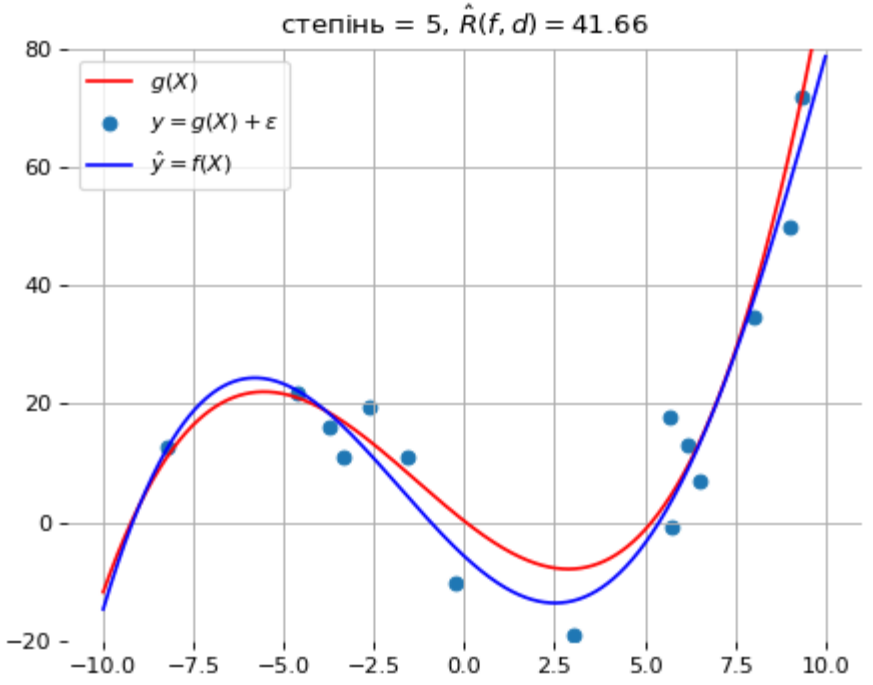
$\mathcal{F}$ = поліноми степеня 2



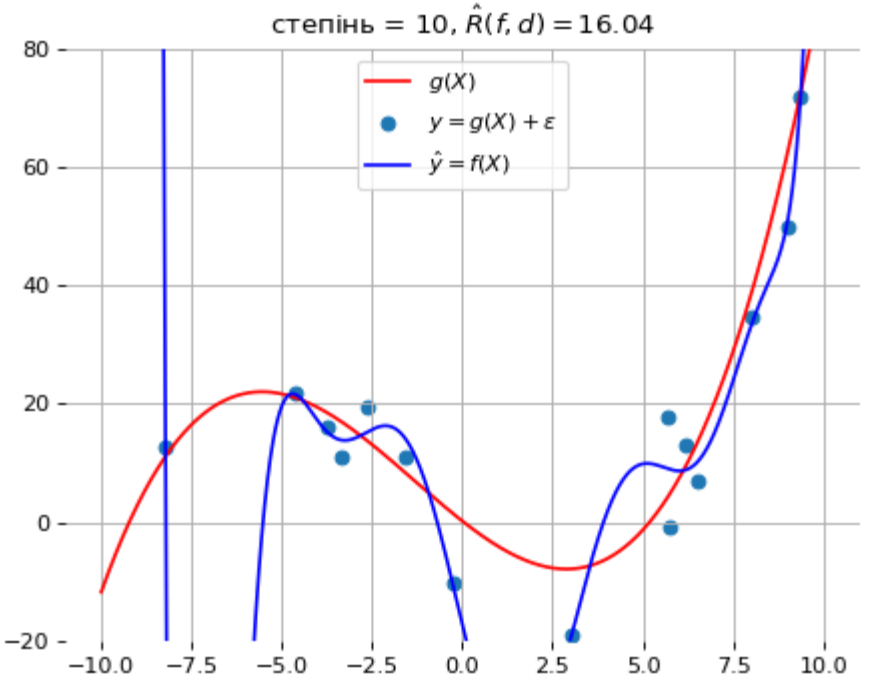
$\mathcal{F}$ = поліноми степеня 3



$\mathcal{F}$ = поліноми степеня 4



$\mathcal{F}$ = поліноми степеня 5



$\mathcal{F}$ = поліноми степеня 10

Степінь поліному VS помилка навчання

Нахай $\mathcal{Y}^{\mathcal{X}}$ — множина усіх функцій $f : \mathcal{X} \to \mathcal{Y}$.

Ми визначаємо ризик Байєса як мінімальний очікуваний ризик для всіх можливих функцій

$$R_B = \min_{f \in \mathcal{Y}^{\mathcal{X}}} R(f),$$

і називаємо оптимальною моделлю Байєса модель f_B, яка досягає цього мінімуму.

Жодна модель f не може працювати краще, ніж f_B.

Потужність простору гіпотез, породженого алгоритмом навчання, інтуїтивно представляє здатність знайти хорошу модель $f \in \mathcal{F}$ для будь-якої функції, незалежно від її складності.

На практиці потужністю можна управляти за допомогою гіперпараметрів алгоритму навчання. Наприклад:

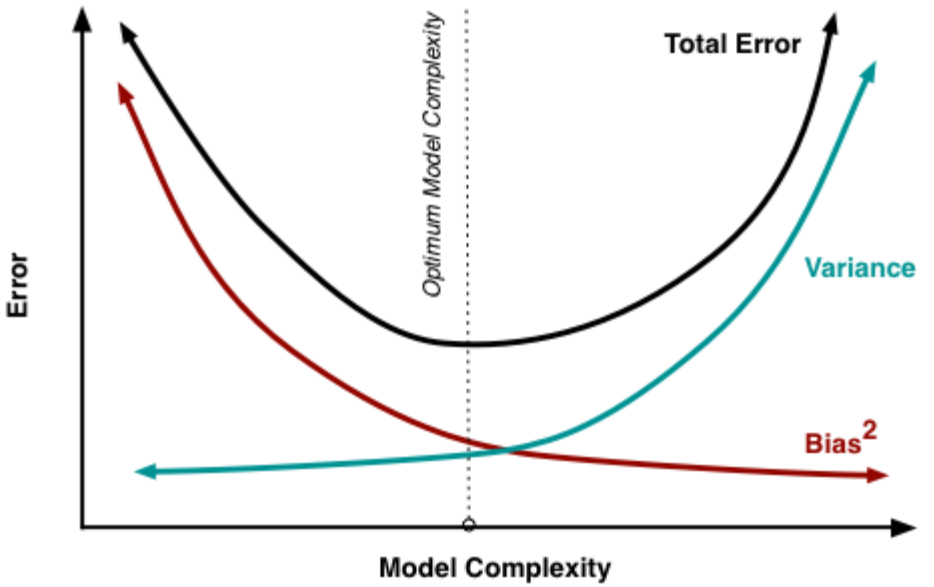
- Степенем полінома;
- Числом прихованих шарів нейронної мережі;
- Кількістю ітерацій навчання (епохами);
- Виразами регуляризації.

- Якщо потужність $\mathcal{F}$ дуже мала, тоді $f_B \notin \mathcal{F}$ і $R(f) - R_B$ дуже велика для будь-якої $f \in \mathcal{F}$, включаючи f_* and $f_*^{\mathbf{d}}$. Кажуть, що такі моделі f недостатньо відповідають даним (недонавчені).
- Якщо потужність $\mathcal{F}$ дуже велика, тоді $f_B \in \mathcal{F}$ і $R(f_*) - R_B$ дуже мала. Однак через високу потужність простору гіпотез мінімізатор емпіричного ризику $f_*^{\mathbf{d}}$ міг як завгодно добре відповідати навчальним даним, так що

$$R(f_*^{\mathbf{d}}) \geq R_B \geq \hat{R}(f_*^{\mathbf{d}}, \mathbf{d}) \geq 0$$

  У цій ситуації $f_*^{\mathbf{d}}$ стає надто спеціалізованою до справжнього процесу генерування даних (розподілу даних) , і значне зменшення емпіричного ризику (часто) відбувається за рахунок збільшення очікуваного ризику . У цій ситуації кажуть, що $f_*^{\mathbf{d}}$ враховує шум, замість взаємозв'язку, що лежить в основі даних (перенавчена модель).

Наша мета полягає в тому, щоб налаштувати потужність простору гіпотез таким чином, щоб знайти оптимальну складність моделі та досягти якомога меншого очікуваного ризику.



У випадку перенавчання,

$$R(f_*^{\mathbf{d}}) \geq R_B \geq \hat{R}(f_*^{\mathbf{d}}, \mathbf{d}) \geq 0$$

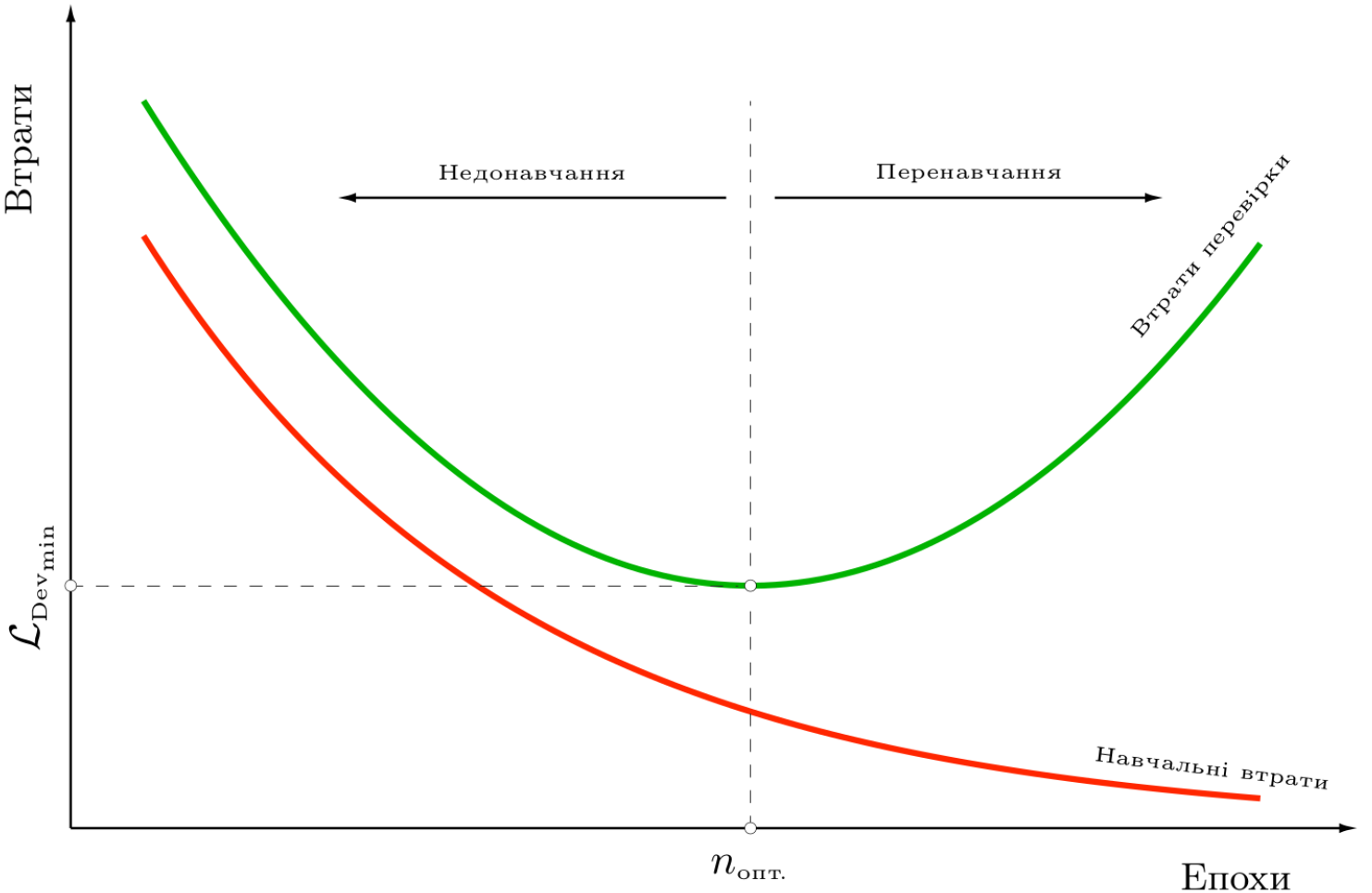
Це свідчить, що емпіричний ризик $\hat{R}(f_*^{\mathbf{d}}, \mathbf{d})$ є поганою оцінкою для очікуваного ризику $R(f_*^{\mathbf{d}})$.

Тим не менш, незміщену оцінку очікуваного ризику можна отримати шляхом оцінки $f_*^{\mathbf{d}}$ на тестових даних $\mathbf{d}_{\text{test}}$ незалежно від навчальної вибірки $\mathbf{d}$:

$$\hat{R}(f_*^{\mathbf{d}}, \mathbf{d}_{\text{test}}) = \frac{1}{n} \sum_{(\mathbf{X}^{(i)}, y^{(i)}) \in \mathbf{d}_{\text{test}}} \ell(y^{(i)}, f_*^{\mathbf{d}}(\mathbf{X}^{(i)}))$$

Ця величина помилки тестування може бути використана для оцінки фактичної продуктивності моделі. Однак не слід використовувати тестову вибірку одночасно з навчальною вибіркою на етапі вибору моделі.

Степінь полінома VS. Помилки


Втрати
Недонавчання
Перенавчання
Втрати перевірки
$\mathcal{L}_{\mathrm{Dev}_{\min}}$
Навчальні втрати
$n_{\text{опт.}}$
Епохи

Компроміс зсуву та дисперсії

Розглянемо фіксовану точку x і прогноз $\hat{y} = f_*^{\mathbf{d}}(x)$ емпіричного мінімізатора ризику для довільного фіксованого x.

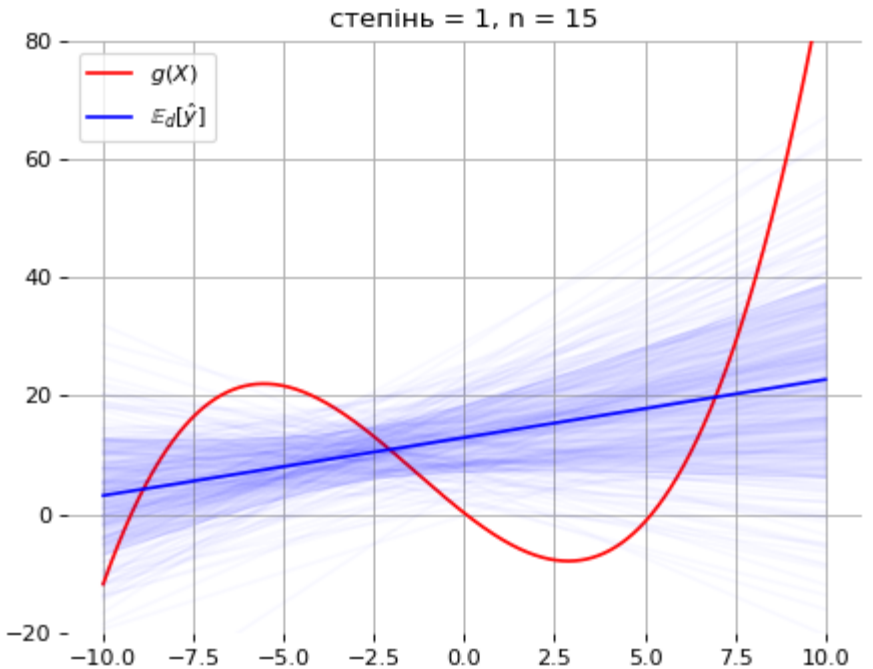
Тоді локальний очікуваний ризик $f_*^{\mathbf{d}}$ дорівнює:

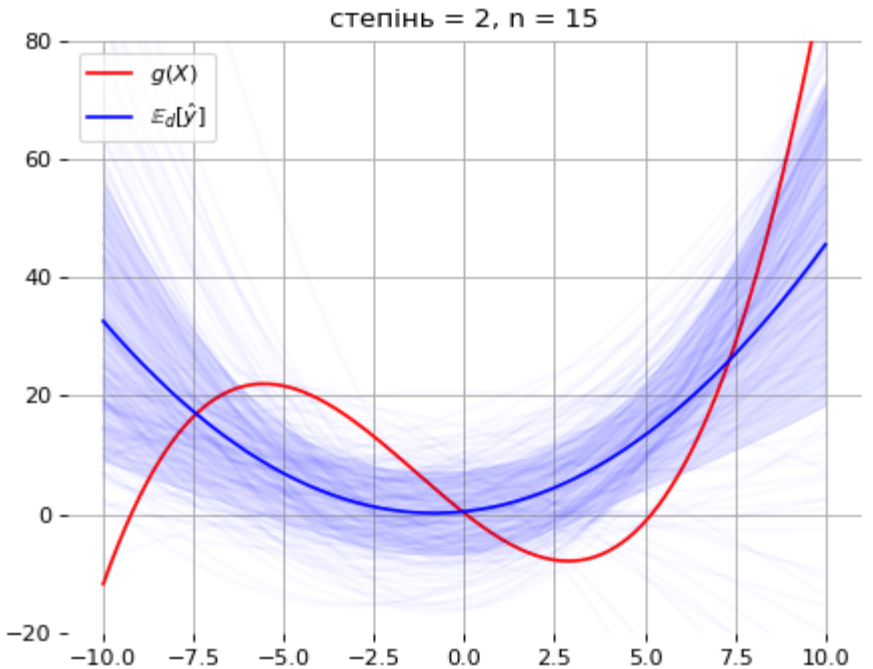
$$\begin{aligned} R(f_*^{\mathbf{d}}|x) &= \mathbb{E}_{y\sim p_{Y|x}}\left[(y - f_*^{\mathbf{d}}(x))^2\right] = \\ &= \mathbb{E}_{y\sim p_{Y|x}}\left[(y - f_B(x) + f_B(x) - f_*^{\mathbf{d}}(x))^2\right] = \\ &= \mathbb{E}_{y\sim p_{Y|x}}\left[(y - f_B(x))^2\right] + \mathbb{E}_{y\sim p_{Y|x}}\left[(f_B(x) - f_*^{\mathbf{d}}(x))^2\right] = \\ &= R(f_B|x) + (f_B(x) - f_*^{\mathbf{d}}(x))^2 \end{aligned}$$

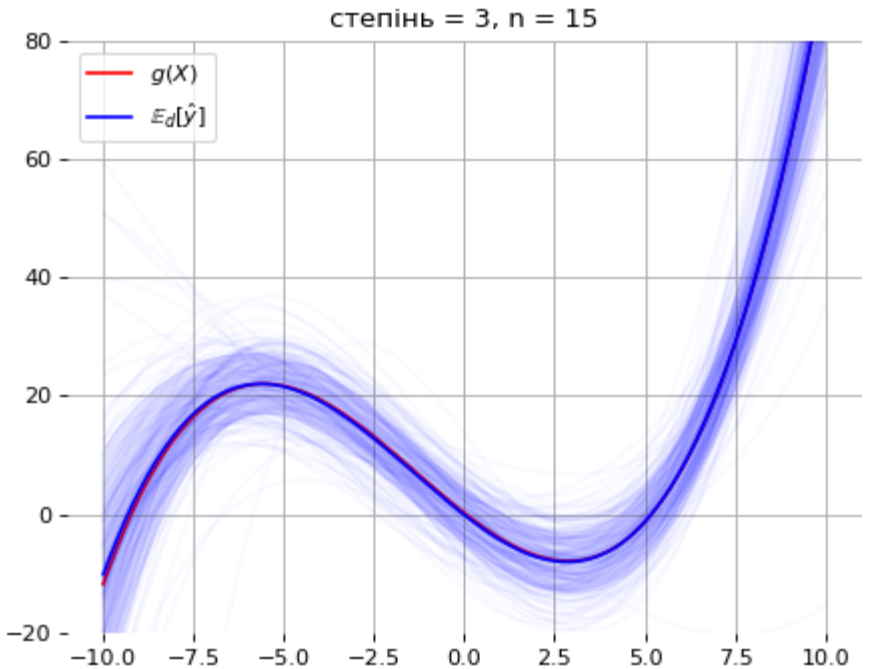
де

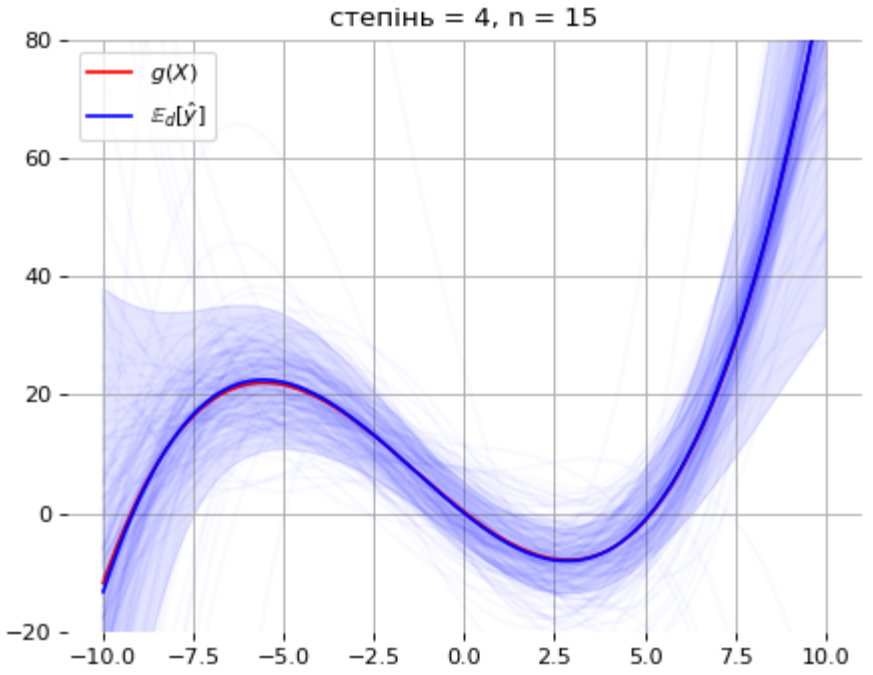
- $R(f_B|x)$ – локальний очікуваний ризик моделі Байєса. Цей тип помилки не можа жодним чином зменшити.
- $(f_B(x) - f_*^{\mathbf{d}}(x))^2$ представляє розбіжність між f_B та $f_*^{\mathbf{d}}$.

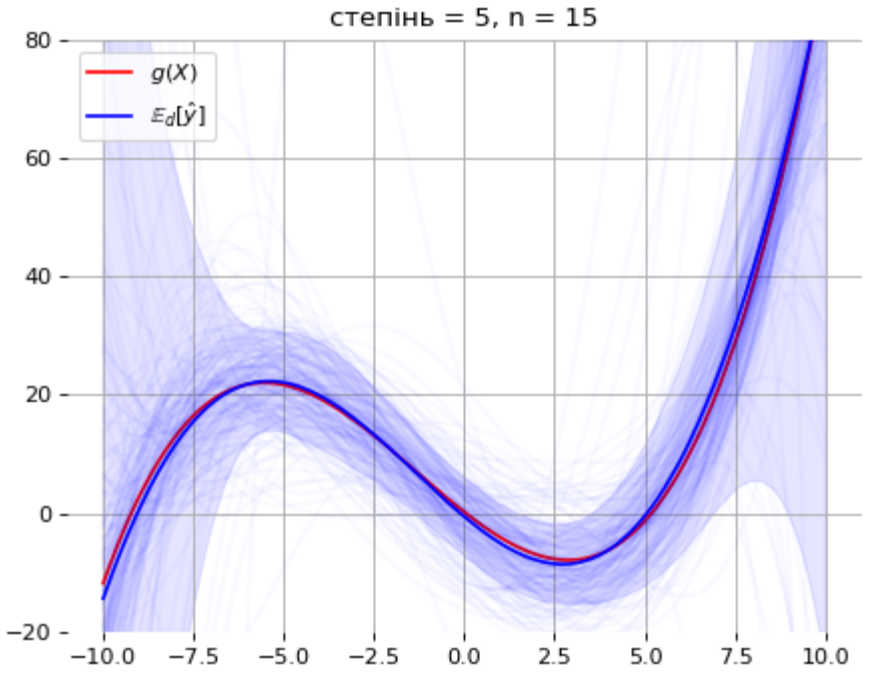
Якщо $\mathbf{d} \sim p_{X,Y}$ розглядається як випадкова змінна, тоді $f_*^{\mathbf{d}}$ — також є випадковою змінною разом із її прогнозами $\hat{y}$.


степінь = 1, n = 15
$g(X)$
$\mathbb{E}_d[\hat{y}]$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


степінь = 2, n = 15
$g(X)$
$\mathbb{E}_d[\hat{y}]$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


степінь = 3, n = 15
$g(X)$
$\mathbb{E}_d[\hat{y}]$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


степінь = 4, n = 15
$g(X)$
$\mathbb{E}_d[\hat{y}]$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


степінь = 5, n = 15
$g(X)$
$\mathbb{E}_d[\hat{y}]$
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Формально середній локальний очікуваний ризик зведеться:

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{d}}\left[R(f_*^{\mathbf{d}}|x)\right] \\
&= \mathbb{E}_{\mathbf{d}}\left[R(f_B|x) + (f_B(x) - f_*^{\mathbf{d}}(x))^2\right] \\
&= R(f_B|x) + \mathbb{E}_{\mathbf{d}}\left[(f_B(x) - f_*^{\mathbf{d}}(x))^2\right] \\
&= \underbrace{R(f_B|x)}_{\text{noise}(x)} + \underbrace{(f_B(x) - \mathbb{E}_{\mathbf{d}}\left[f_*^{\mathbf{d}}(x)\right])^2}_{\text{bias}^2(x)} + \underbrace{\mathbb{E}_{\mathbf{d}}\left[(\mathbb{E}_{\mathbf{d}}\left[f_*^{\mathbf{d}}(x)\right] - f_*^{\mathbf{d}}(x))^2\right]}_{\text{var}(x)}
\end{aligned}
$$

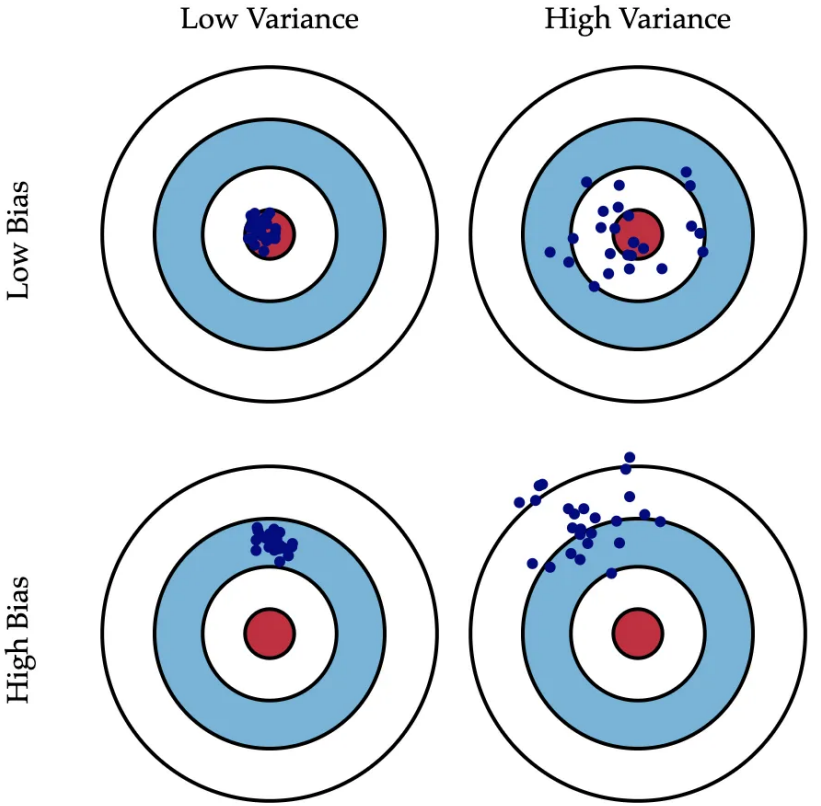
Цей запис відомий як компроміс зсуву-дисперсії.

- Доданок шуму кількісно визначає частину очікуваного ризику, яку неможливо зменшити.
- Доданок зсуву визначає розбіжність між середньою моделлю та моделлю Байєса.
- Доданок дисперсії визначає мінливість прогнозів.

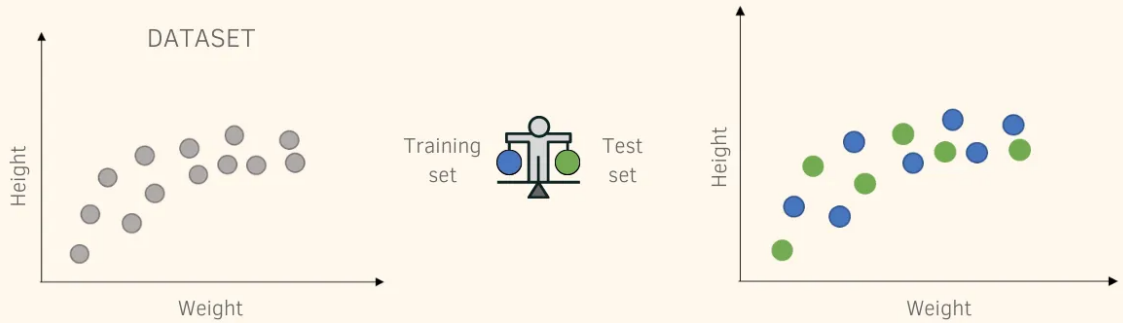
Компроміс зсуву та дисперсії

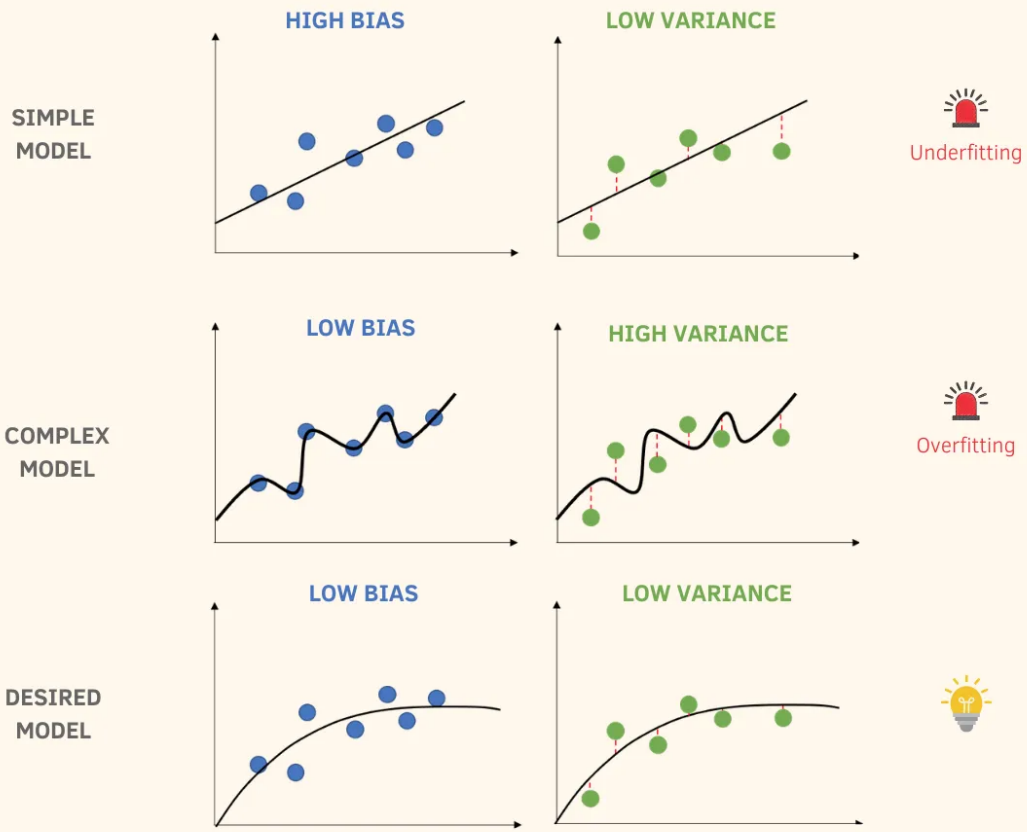
- Зменшення потужності $\mathcal{F}$ призводить до того, що $f_*^{\mathbf{d}}$ починає у середньому гірше відповідати даним, що збільшує зсув.
- Збільшення потужності $\mathcal{F}$ призводить до того, що $f_*^{\mathbf{d}}$ сильно пристосовується до навчальних даних, що збільшує дисперсію.

Інтуїція



BIAS AND VARIANCE TRADEOFF





Medium

Image created by @cristianefragata

Демо

Кінець