Wprowadzenie do uczenia statystycznego

Wstęp – uczenie maszynowe a statystyka

„Machine learning is just **glorified statistics**”

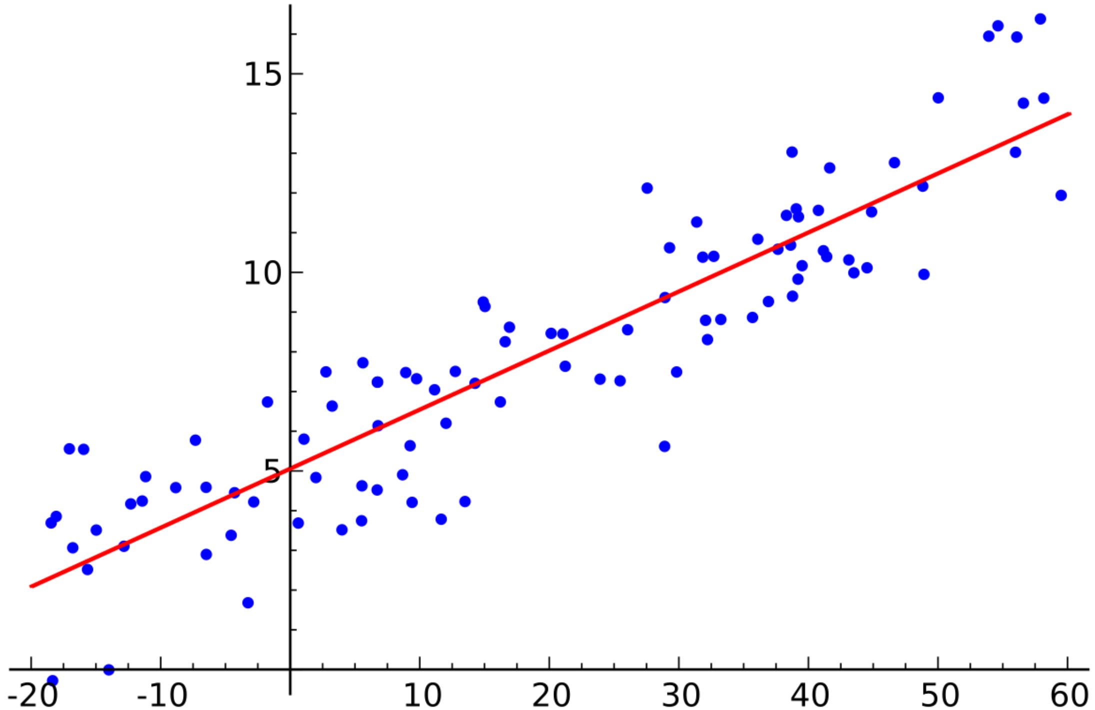
Robert Tibshirani

Wstęp – uczenie maszynowe a statystyka

- Celem tego wykładu jest zrozumienie podstaw uczenia maszynowego jako dziedziny wiedzy.
- Zacznijmy od próby zrozumienia różnic pomiędzy statystyką a uczeniem maszynowym.

Wstęp – uczenie maszynowe a statystyka

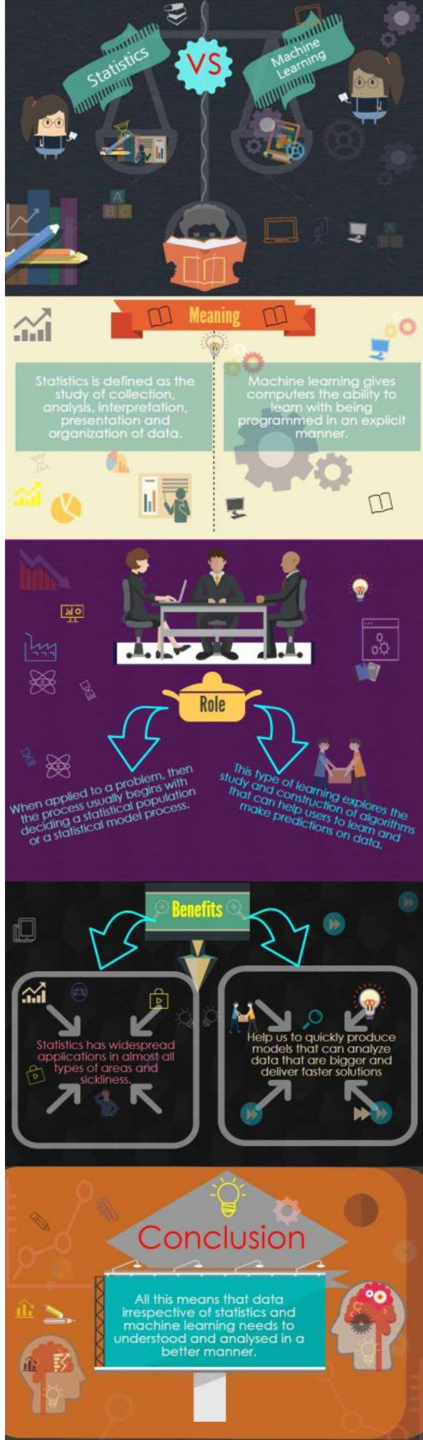
- Zastanówmy się nad regresją liniową – podstawowym modelem w obu dziedzinach.
- W jaki sposób będzie o nim myślał statystyk, a jak ekspert uczenia maszynowego?



Wstęp – uczenie maszynowe a statystyka

- W dużym uproszczeniu możemy powiedzieć, że celem statystyka jest **wnioskowanie** na podstawie zebranych danych na temat zależności pomiędzy nimi zachodzących.
- Natomiast w przypadku uczenia maszynowego celem jest **przewidywanie** tego jak będą zachowywać nieznane dane bazując na tych które już znamy.

Wstęp – uczenie maszynowe a statystyka



Typologia uczenia maszynowego

- 3 podstawowe podejścia do uczenia maszynowego:
  - Uczenie nadzorowane
  - Uczenie nienadzorowane
  - Uczenie ze wzmocnieniem

Uczenie statystyczne

- Oczekiwania od modeli:
  - maksymalizacja jakości prognoz
  - Krótki czas na przygotowanie modelu (automatyzacja procesu)

Uczenie statystyczne

Pierwotna definicja uczenia statystycznego (Vapnik, 1999):

Dla zadanej klasy funkcji $\mathcal{F} = \{\alpha \in \Lambda : f(x, \alpha)\}$, procesu generującego dane (X, Y) oraz funkcji straty $\mathrm{L}(y, \hat{y})$ należy rozwiązać problem:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}}(E(L\,(Y, f(X, \alpha)))$$

Na podstawie próby: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Uczenie statystyczne

Pierwotna definicja uczenia statystycznego (Vapnik, 1999):

Dla zadanej klasy funkcji $\mathcal{F} = \{\alpha \in \Lambda: f(x, \alpha)\}$, procesu generującego dane (X, Y) oraz funkcji straty $\mathrm{L}(y, \hat{y})$ należy rozwiązać problem:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}}(E(L(Y, f(X, \alpha)))$$

Na podstawie próby: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Aktualna definicja operacyjna (James et.al, 2013)

Zestaw narzędzi pozwalających na modelowanie i rozumienie złożonych zbiorów danych.

Twierdzenie Vapnika – problem klasyfikacji

- Zadana klasa funkcji dopuszczalnych $\mathcal{F}$.
- Dla $\mathcal{F}$ można wyznaczyć **wymiar Vapnika-Chervonenkisa** $h(\mathcal{F})$ mierzący jej zdolność dopasowania się do danych.
- Dysponujemy n-elementową próbą estymacyjną.
- Wybieramy funkcję $f \in \mathcal{F}$ minimalizującą błąd na danych estymacyjnych R_e.
- Chcemy oszacować błąd prognozy R_p:

Twierdzenie Vapnika – problem klasyfikacji

Twierdzenie (Vapnik, 1995):

Dla dowolnego rozkładu (X, Y) z prawdopodobieństwem $1 - q$ zachodzi zależność:

$$R_p \leq R_e + \underbrace{\sqrt{\frac{h(\mathcal{F})\,(1 + \ln(2n/h(\mathcal{F}))) - \ln(q/4)}{n}}}_{\varepsilon}$$

Wymiar Vapnika-Chervonenkisa

- Dla zadanego zbioru przykładów $C = \{c_1, c_2, c_3, \dots, c_n\} \in X$ możemy zdefiniować klasyfikator f jako funkcję która dla każdego podzbioru $C' \subseteq C$ (**każdej klasy**) pozwala przypisać mu odpowiednią etykietę:

$$f(c) = \begin{cases} 1 & c \in C' \\ 0 & c \in 1 - C' \end{cases}$$

Wymiar Vapnika-Chervonenkisa

- Dla zadanego zbioru przykładów $C = \{c_1, c_2, c_3, \dots, c_n\} \in X$ możemy zdefiniować klasyfikator f jako funkcję która dla każdego podzbioru $C' \subseteq C$ (**każdej klasy**) pozwala przypisać mu odpowiednią etykietę:
$$f(c) = \begin{cases} 1 & c \in C' \\ 0 & c \in 1 - C' \end{cases}$$
- Mówimy, że klasa funkcji $\mathcal{F}$ **rozdziela** (*shatters*) zbiór C jeżeli istnieje taka funkcja $f \in \mathcal{F}$, że dla przyporządkowania $\mathcal{F}_C = \{(f(c_1), f(c_2), \dots, f(c_n)) | f \in \mathcal{F}\}$ moc zbioru $|\mathcal{F}_C| = 2^{|C|}$

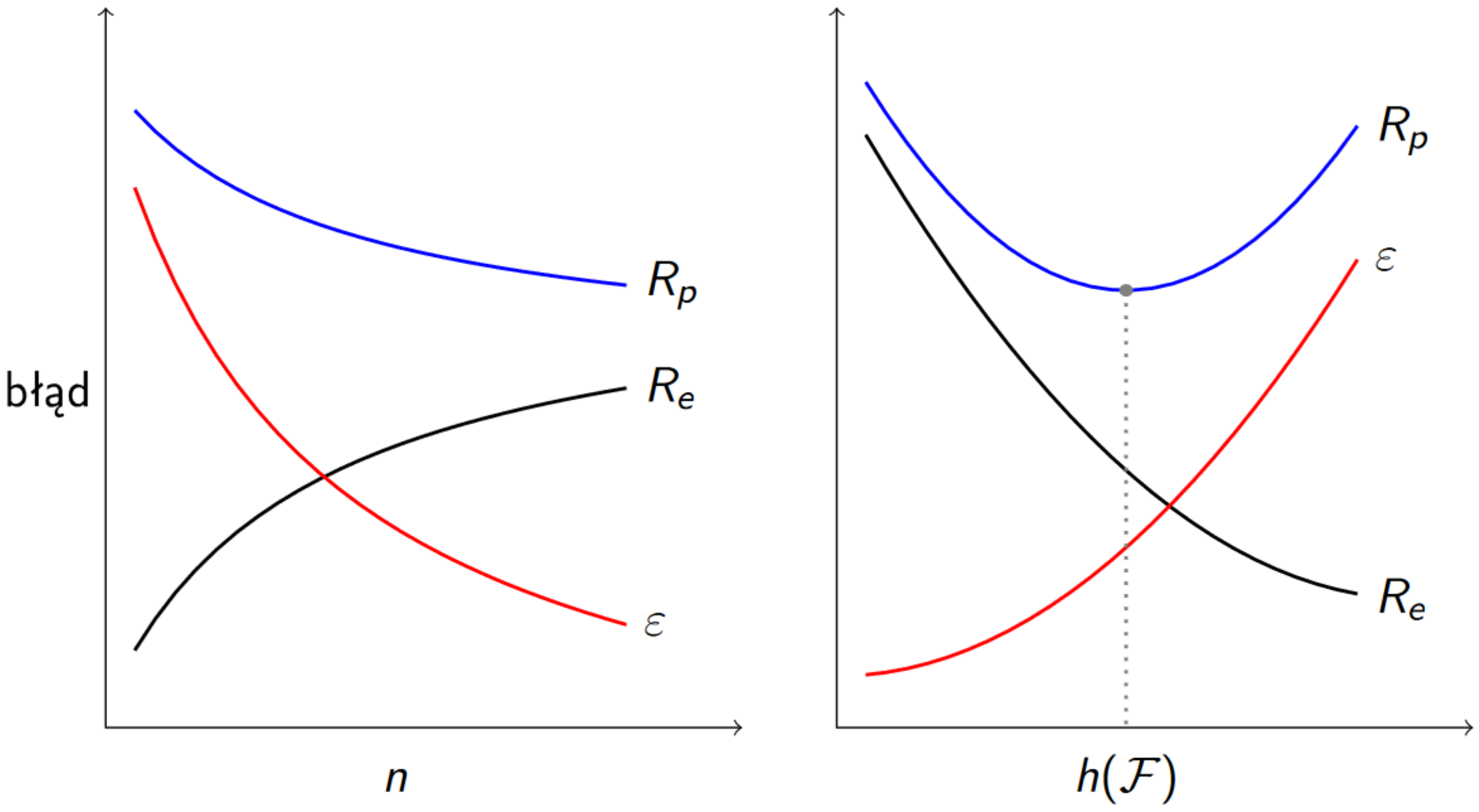
Wymiar Vapnika-Chervonenkisa

- Dla zadanego zbioru przykładów $C = \{c_1, c_2, c_3, \dots, c_n\} \in X$ możemy zdefiniować klasyfikator f jako funkcję która dla każdego podzbioru $C' \subseteq C$ (**każdej klasy**) pozwala przypisać mu odpowiednią etykietę:
$$f(c) = \begin{cases} 1 & c \in C' \\ 0 \; c \in 1 - C' \end{cases}$$
- Mówimy, że klasa funkcji $\mathcal{F}$ **rozdziela** (*shatters*) zbiór C jeżeli istnieje taka funkcja $f \in \mathcal{F}$, że dla przyporządkowania $\mathcal{F}_C = \{(f(c_1), f(c_2), \dots, f(c_n)) | \, f \in \mathcal{F}\}$ moc zbioru $|\mathcal{F}_C| = 2^{|C|}$
- Intuicyjnie oznacza to, że funkcja f poprawnie przyporządkowuje etykiety na wszystkie 2^n sposobów niezależnie od tego w jaki sposób są one początkowo przydzielone (jaka próbka C została wylosowana)

Wymiar Vapnika-Chervonenkisa

- Wymiar Vapnika-Chervonenkisa $h(\mathcal{F})$ klasy funkcji $\mathcal{F}$ definiujemy jako rozmiar największego zbioru $C \in X$ jaki jesteśmy w stanie rozdzielić za pomocą $\mathcal{F}$.
- Gdy możemy rozdzielić dowolny zbiór (nasz klasyfikator jest idealny) wtedy $h(\mathcal{F}) = \infty$.

Twierdzenie Vapnika – problem klasyfikacji



Twierdzenie Vapnika – problem klasyfikacji

- Wybieramy rodzinę zagnieżdżonych klas funkcji:

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$$

$$\Downarrow$$

$$h(\mathcal{F}_1) \leq h(\mathcal{F}_2) \leq h(\mathcal{F}_3) \leq \ldots$$

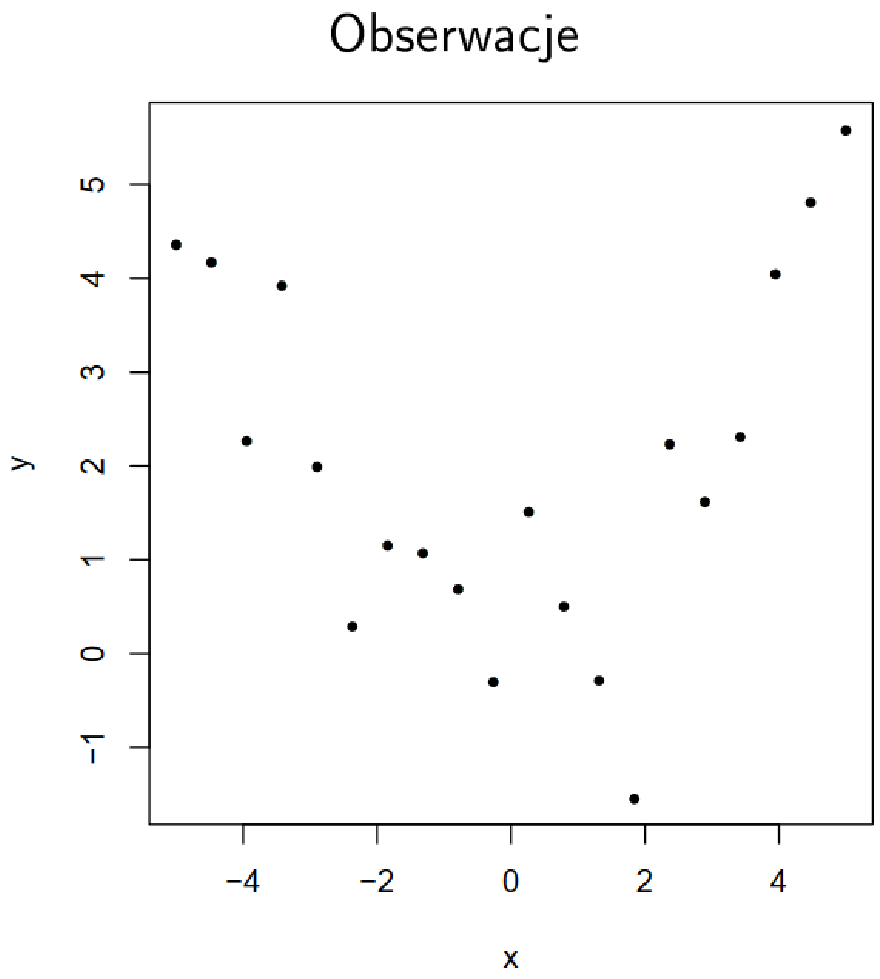
- Wyznaczamy:

$$R_e(\mathcal{F}_1) \geq R_e(\mathcal{F}_2) \geq R_e(\mathcal{F}_3) \geq \ldots$$

$$\varepsilon(\mathcal{F}_1) \leq \varepsilon(\mathcal{F}_2) \leq \varepsilon(\mathcal{F}_3) \leq \ldots$$

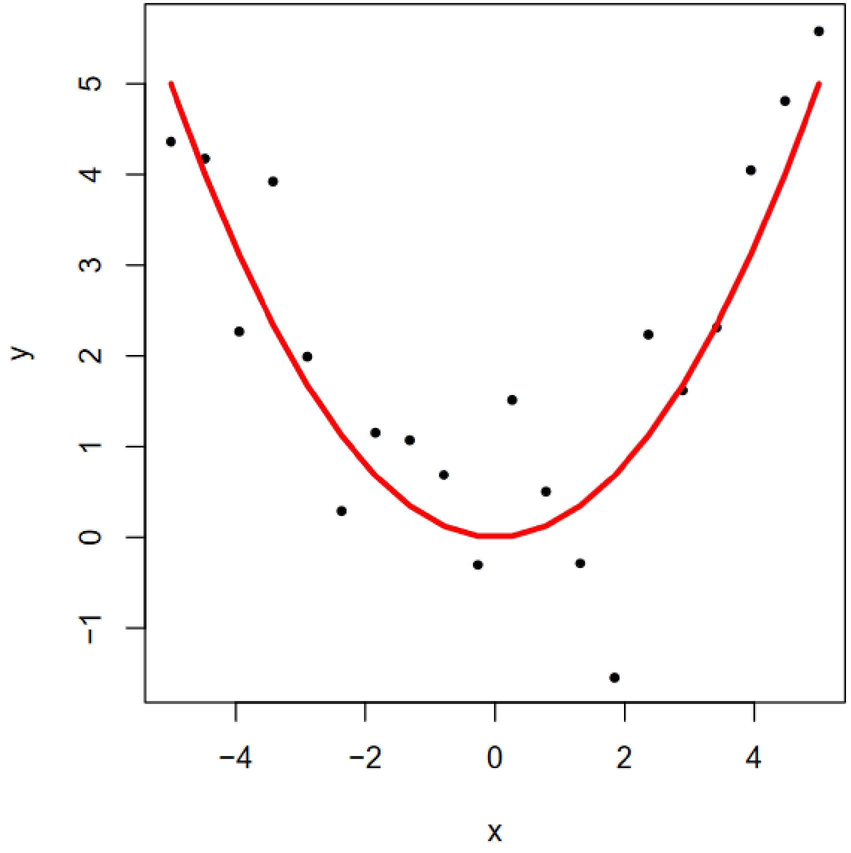
- Wybieramy model oszacowany na podstawie klasy funkcji $\mathcal{F}_i$ minimalizującej oszacowanie R_p.

Twierdzenie Vapnika – problem klasyfikacji



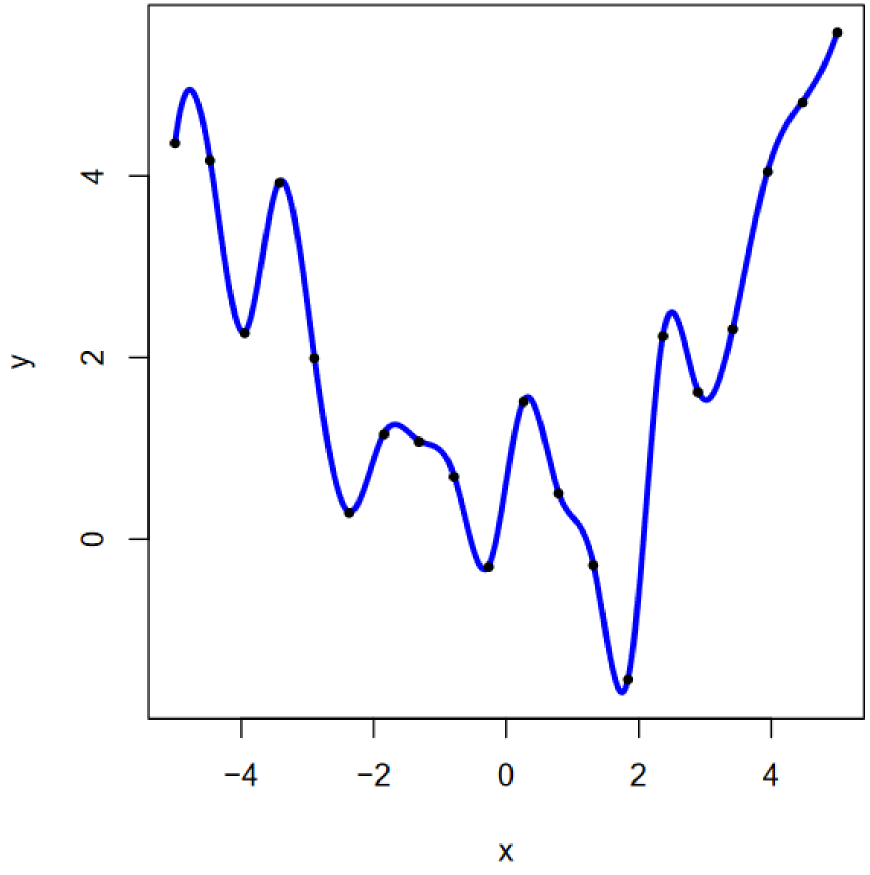
Twierdzenie Vapnika – problem klasyfikacji

proces generujący dane: $y = x^2/5 + \varepsilon$, gdzie $\varepsilon \sim N(0, 1)$



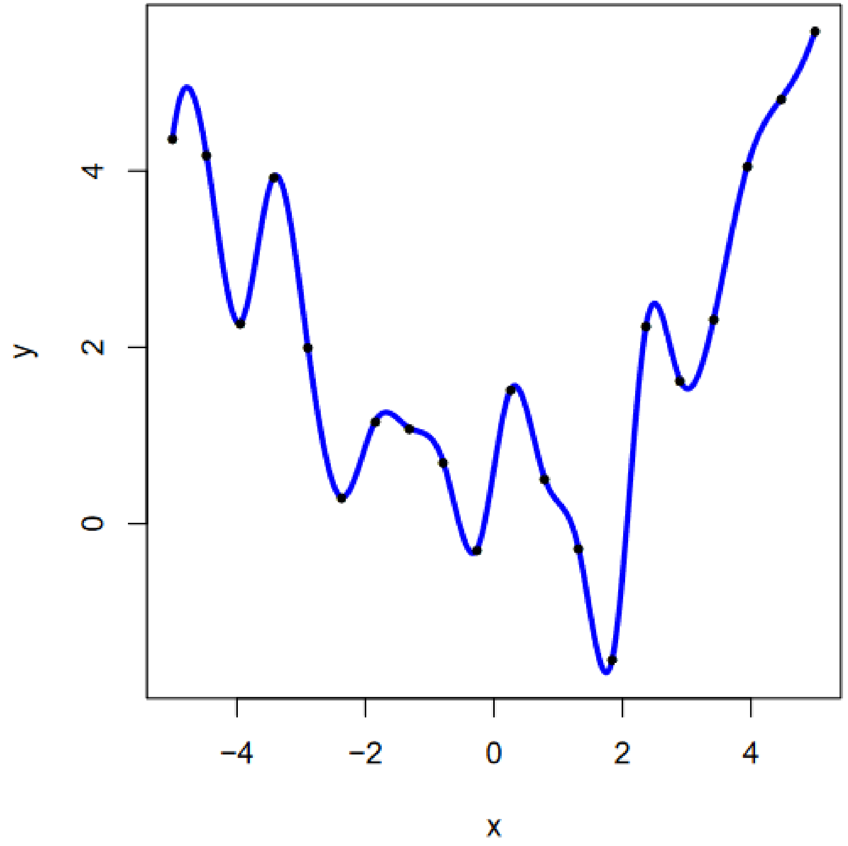
Twierdzenie Vapnika – problem klasyfikacji

Dwukrotnie różniczkowalna funkcja f: $\sum_{i=1}^{n}(f(x_i) - y_i)^2 \to \min$



Twierdzenie Vapnika – problem klasyfikacji

Dwukrotnie różniczkowalna funkcja f: $\sum_{i=1}^{n}(f(x_i) - y_i)^2 \to \min$

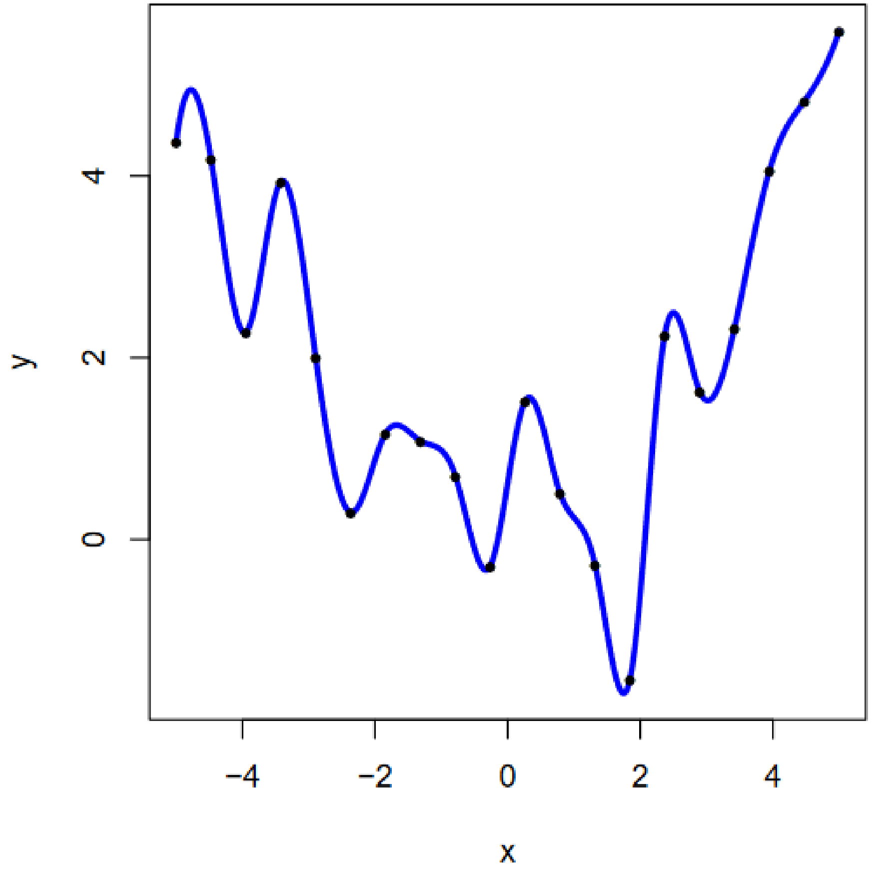


zagnieżdżona klasa funkcji:
wygładzane funkcje sklejane (Hastie et al., 2001)

Twierdzenie Vapnika – problem klasyfikacji

Dwukrotnie różniczkowalna funkcja f:

$\sum_{i=1}^{n}(f(x_i) - y_i)^2 \rightarrow \min$, p.w. $\int_D [f''(x)]^2 dx \leq \delta$

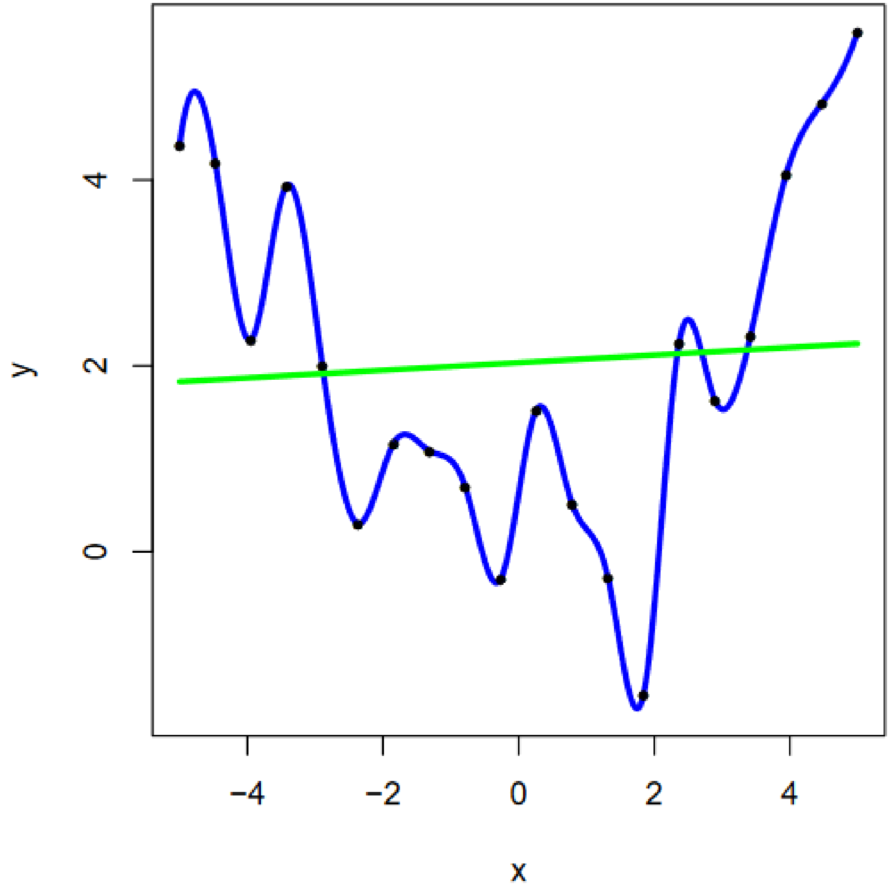


niebieski: $\delta \rightarrow +\infty$

Twierdzenie Vapnika – problem klasyfikacji

Dwukrotnie różniczkowalna funkcja f:

$$\sum_{i=1}^{n}(f(x_i) - y_i)^2 \to \min, \quad \text{p.w.} \quad \int_D [f''(x)]^2 dx \le \delta$$

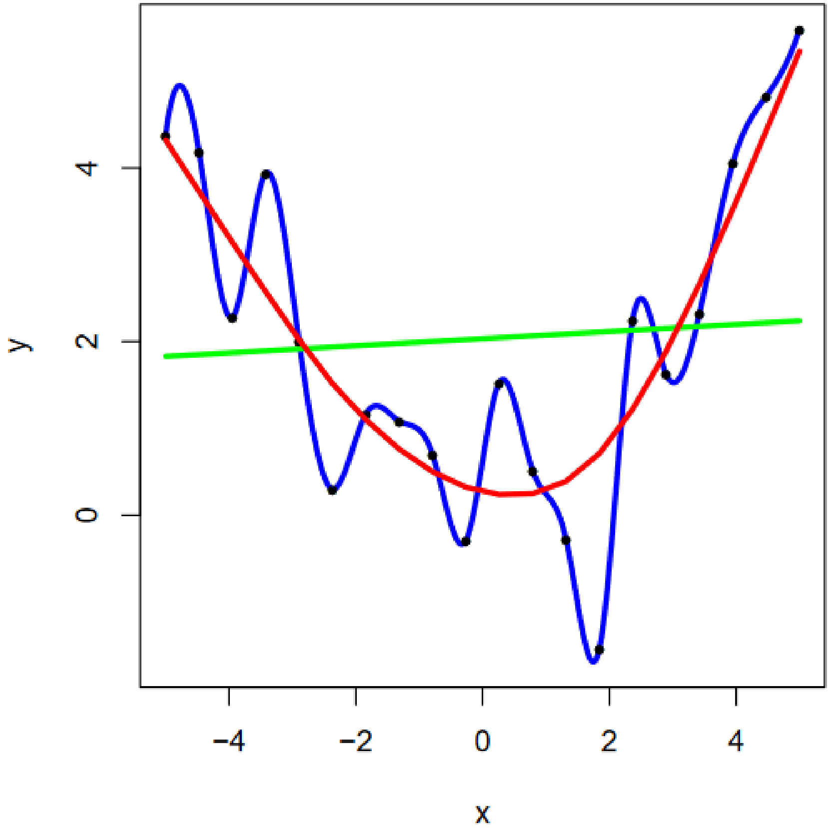


niebieski: $\delta \to +\infty$, zielony: $\delta = 0$

Twierdzenie Vapnika – problem klasyfikacji

Dwukrotnie różniczkowalna funkcja f:

$\sum_{i=1}^{n}(f(x_i) - y_i)^2 \to \min$, p.w. $\int_D [f''(x)]^2 dx \leq \delta$



niebieski: $\delta \to +\infty$, zielony: $\delta = 0$, czerwony: δ optymalne

Twierdzenie Vapnika – problem klasyfikacji

- Ograniczenia twierdzenia Vapnika:
  - Trudność z wyznaczeniem wartości $h(\mathcal{F})$ dla złożonych klas funkcji
  - Nierówność z twierdzenia jest bardzo konserwatywna
- W praktyce stosujemy zwykle procedury alternatywne:
  - kryteria informacyjne (AIC, BIC,. . .)
  - Zbiór walidacyjny
  - Walidacja krzyżowa
  - bootstrapping

Bibliografia

- Hastie T., Tibshirani R., and Friedman J., The Elements Of Statistical Learning, Springer, 2001
- James G., Witten D., Hastie T., and Tibsirani R., An Introduction to Statistical Learning, 2013
- Vapnik V., The Nature of Statistical Learning Theory, Springer, NewYork, 1995
- Vapnik V., An Overview of Satistical Learning Theory, IEEE Transactions on Neural Networks, 10(5), s.988-999, 1999