Problem przeuczenia

Systemy uczące się - laboratorium

Mateusz Lango

Zakład Inteligentnych Systemów Wspomagania Decyzji
Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”,
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

Na ostatnich zajęciach: zasada MLE

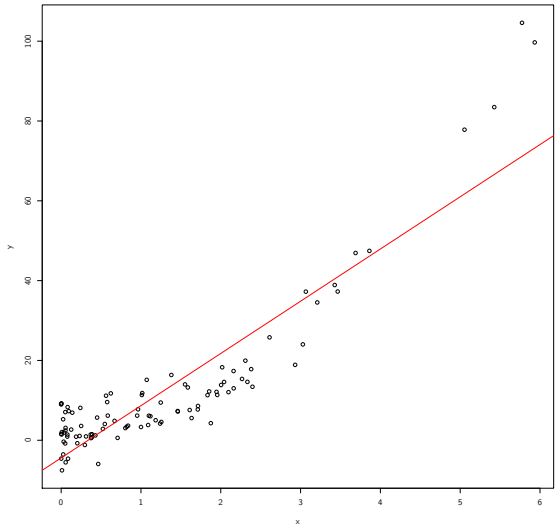
Estymacja maksymalnej wiarygodności

$$\max \sum_{i=1}^{n} \ln P(\vec{x_i}, y_i)$$

Dwa rodzaje modeli statystycznego uczenia maszynowego:

- modele dyskryminacyjne $P(y|x)$
- modele generatywne $P(x, y)$

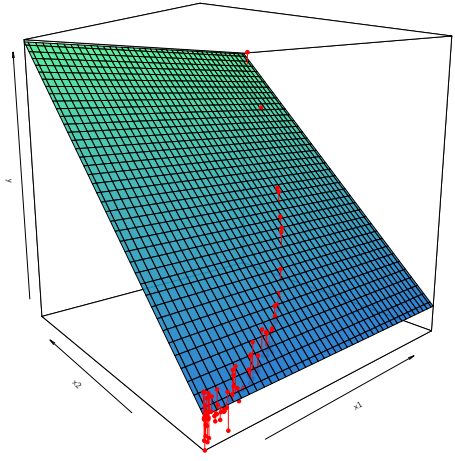
Cechy wielomianowe

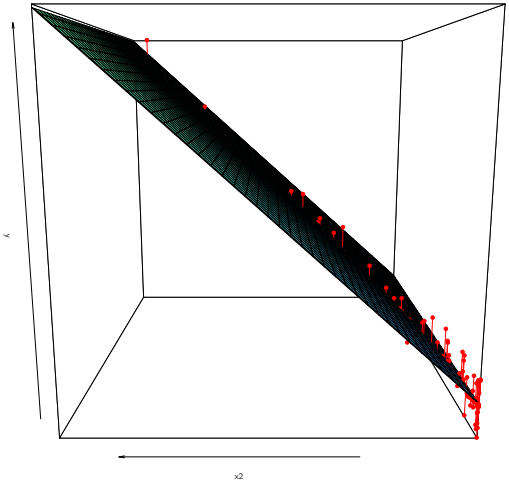


Problem

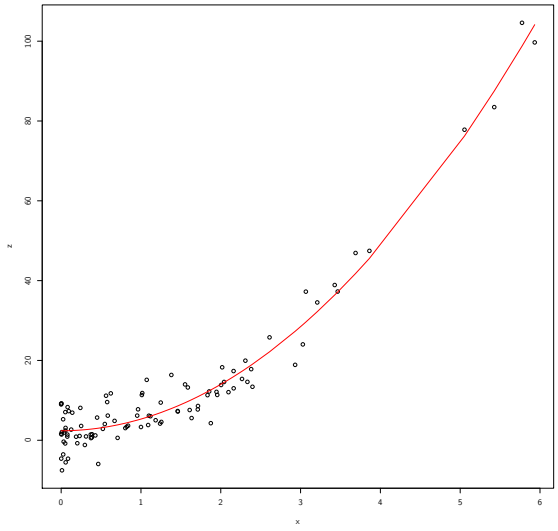
Mając do dyspozycji regresję liniową obserwujemy, że linia prosta nie jest wystarczająca do zamodelowania wiedzy w danych. Co możemy zrobić?
Podpowiedź: regresja liniowa jest liniowa w wagach (a nie cechach!)

Cechy wielomianowe





Regresja wielomianowa



Regresja ma postać:

$$h(x) = b + w_1 x_1 + w_2 x_2$$

ale poprzez sposób konstrukcji zbioru danych uzyskujemy:

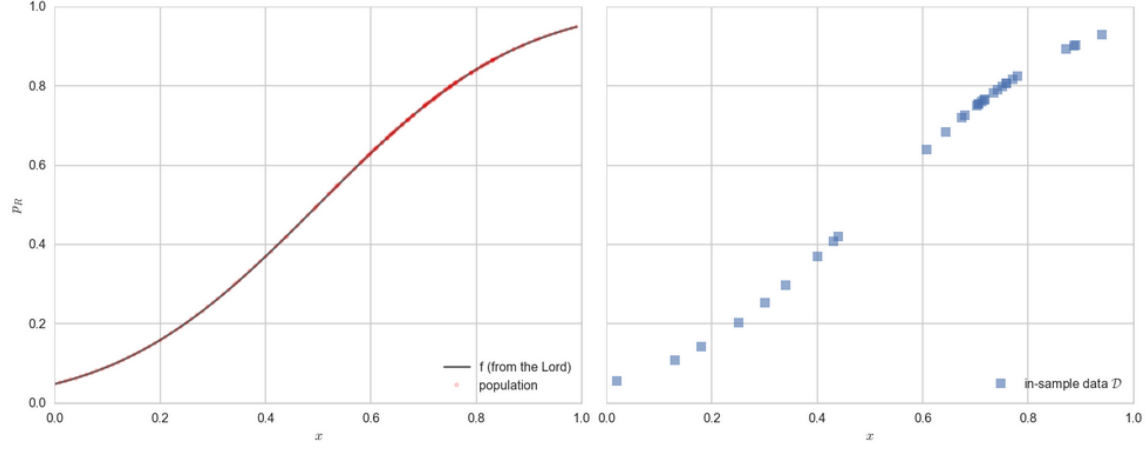
$$h(x) = b + w_1 x_1 + w_2 x_1^2$$

Theorem (Stone'a-Weierstrassa)

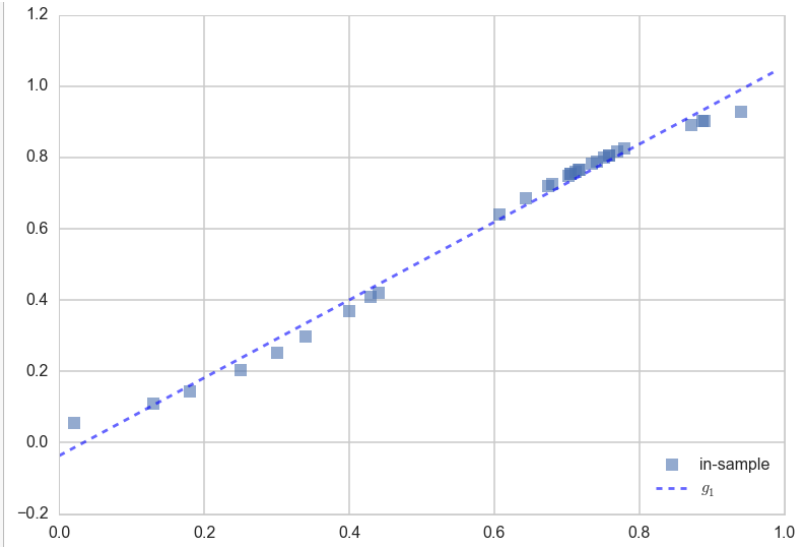
Każdą funkcję ciągłą o wartościach rzeczywistych na przedziale domkniętym można przybliżyć jednostajnie z dowolną dokładnością wielomianami.

Problem dopasowania: czy więcej znaczy lepiej?

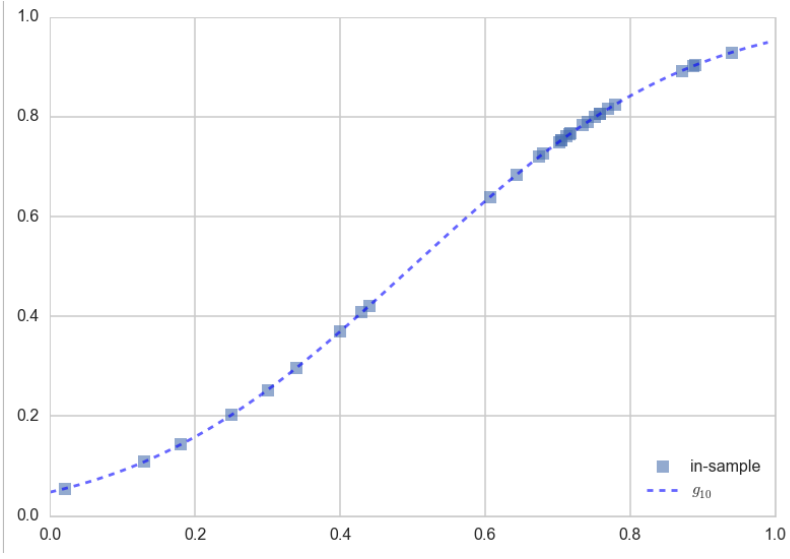
Ponieważ im więcej cech wielomianowych dodamy, tym lepiej możemy przybliżyć funkcję $f(x)$ to wydaje się, że powinniśmy dodawać jak najwięcej cech, aby uzyskać jak najlepszy model.



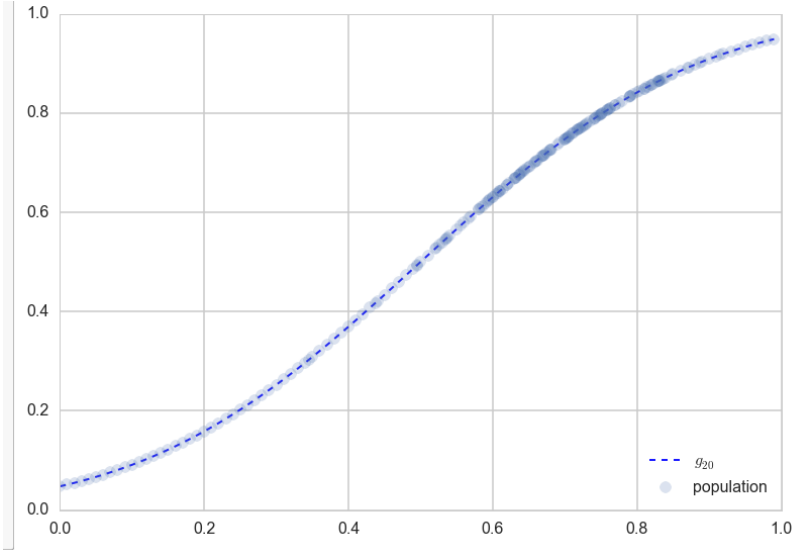
Problem dopasowania: czy więcej znaczy lepiej? - cechy liniowe



Problem dopasowania: czy więcej znaczy lepiej? - cechy wielomianowe rzędu 10

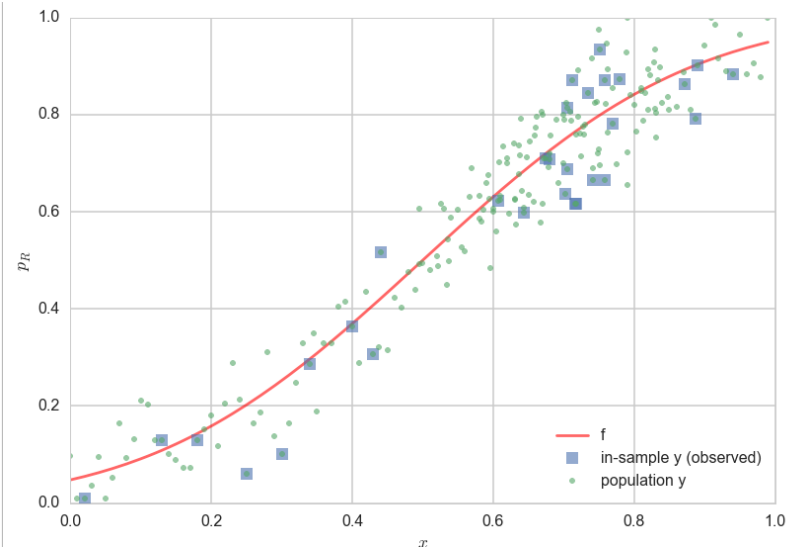


Problem dopasowania: czy więcej znaczy lepiej? - cechy wielomianowe rzędu 20



Obrana strategia zdaje się działać...

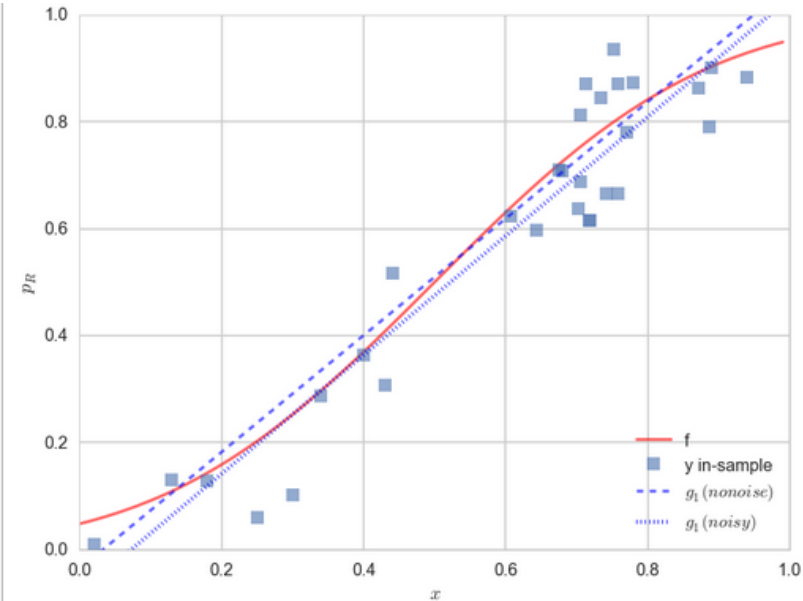
Problem dopasowania: czy więcej znaczy lepiej?



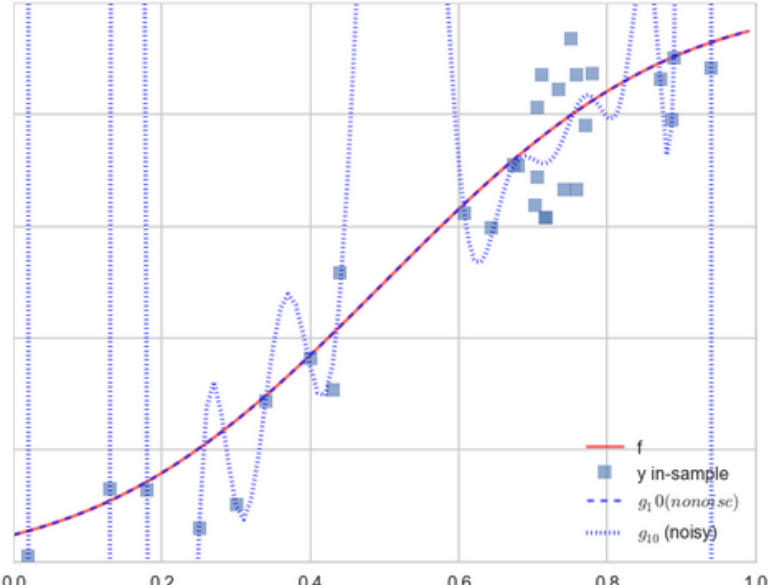
Jednak problem uczenia jest w rzeczywistości trudniejszy: w próbce obecny jest szum.

$$f(x) + \epsilon$$

Problem dopasowania: czy więcej znaczy lepiej? cechy liniowe

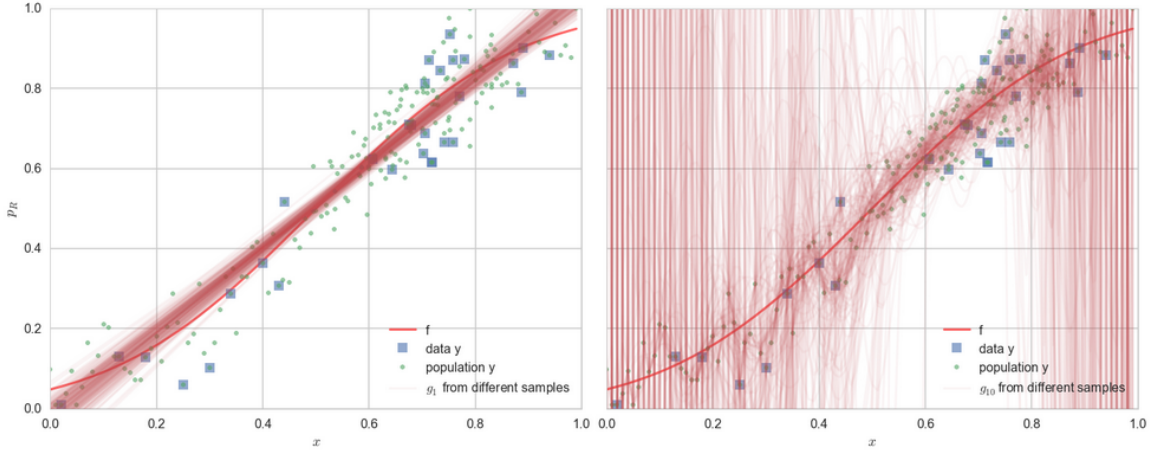


Problem dopasowania: czy więcej znaczy lepiej? cechy wielomianowe rzędu 10



Błąd uczący zdecydowanie się poprawił względem modelu liniowego, jednak...

Problem dopasowania: czy więcej znaczy lepiej? Porównane na wielu zbiorach



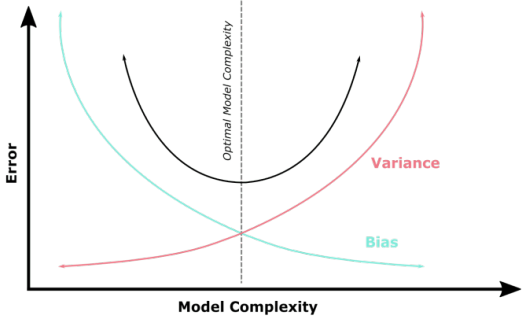
Dwa czynniki błędu

- błąd wynikający z ograniczeń klasy hipotez
  - jaki błąd popełnia hipoteza wybrana na podstawie wszystkich możliwych zbiorów uczących?
- błąd wynikający z trudności wybrania najlepszej hipotezy
  - jak bardzo model się zmieni przy zmianie danych uczących?
  - inaczej: jak bardzo model zamodelował charakterystykę próbki uczącej

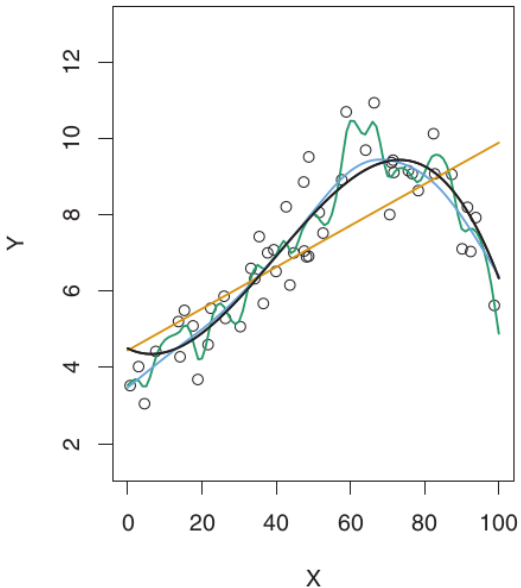
Dekompozycja obciążenie-wariancja

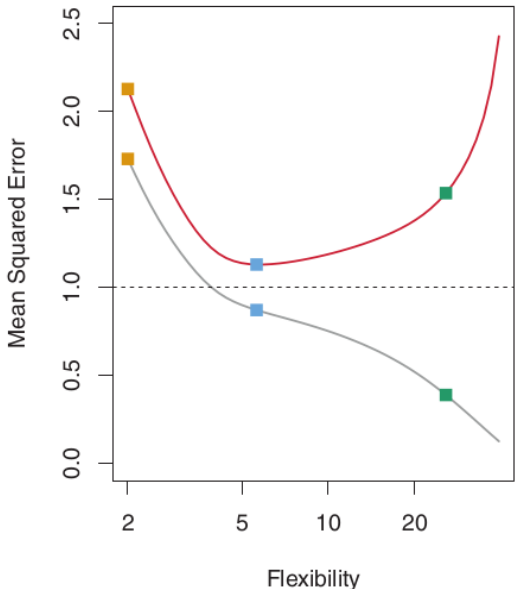
Theorem (Dekompozycja obciążenie-wariancja)

$$\mathbb{E}[(\hat{f}(X) - f(X))^2] = \mathbb{D}^2[\hat{f}(X)] + [Bias(\hat{f}(X))]^2 + \mathbb{D}^2[\epsilon]$$



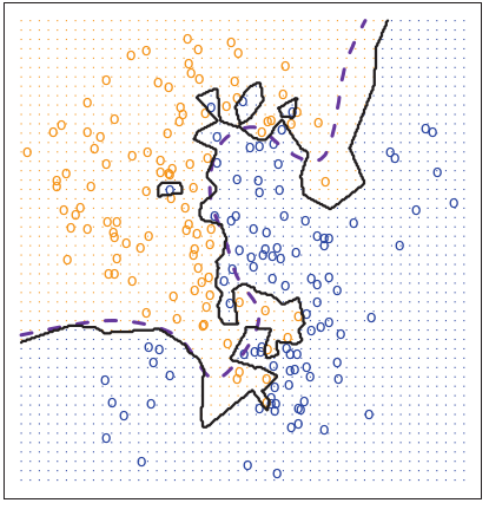
Dekompozycja obciążenie-wariancja



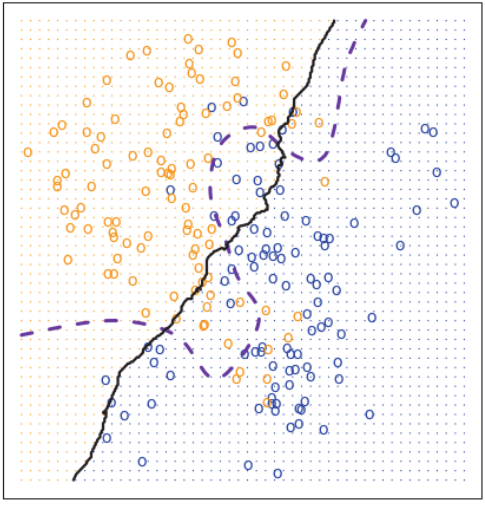


Dekompozycja obciążenie-wariancja

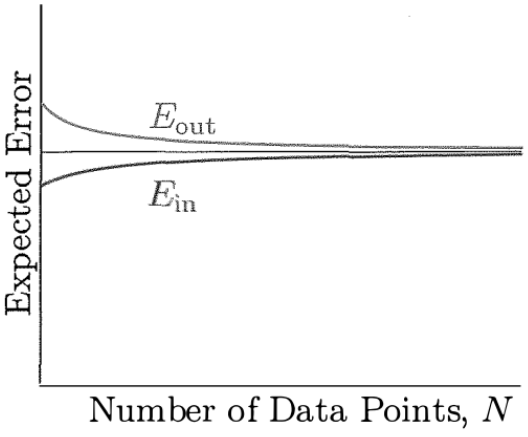
KNN: K=1



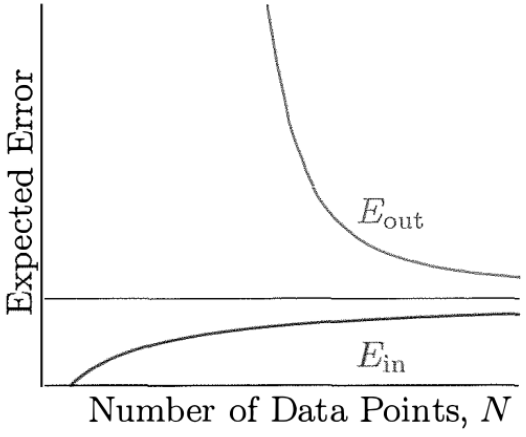
KNN: K=100



Diagnostyka modelu



Simple Model



Complex Model

Potrójny przetarg

W uczeniu maszynowym dochodzi do tzw. potrójnego przetargu:

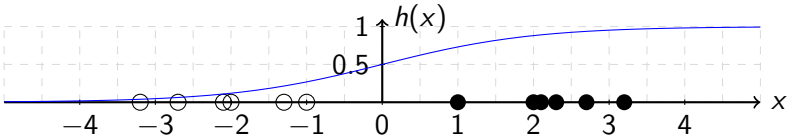
- złożoność (ang. *capacity*) klasy hipotez
- wielkość zbioru uczącego
- błąd na nowych danych (uogólnianie)

Regresja logistyczna - przykład uczenia

Rozważmy jednowymiarowy zbiór danych przedstawiony na ilustracji. Zgodnie z modelem regresji logistycznej

$$h(x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

gdzie $w = 1$ i $b = 0$. Czy to byłby ostateczny wynik uczenia zgodny z MLE?

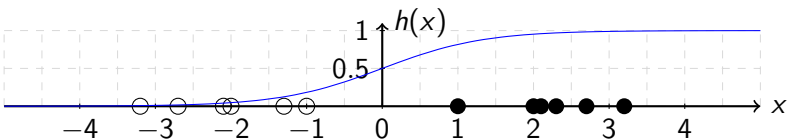


Regresja logistyczna - przykład uczenia

Rozważmy jednowymiarowy zbiór danych przedstawiony na ilustracji. Zgodnie z modelem regresji logistycznej

$$h(x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

gdzie $w = 1.5$ i $b = 0$. Czy to byłby ostateczny wynik uczenia zgodny z MLE?

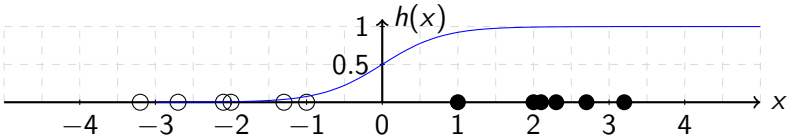


Regresja logistyczna - przykład uczenia

Rozważmy jednowymiarowy zbiór danych przedstawiony na ilustracji. Zgodnie z modelem regresji logistycznej

$$h(x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

gdzie $w = 2.5$ i $b = 0$. Czy to byłby ostateczny wynik uczenia zgodny z MLE?

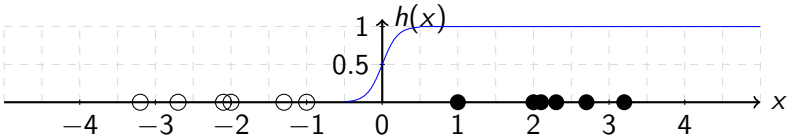


Regresja logistyczna - przykład uczenia

Rozważmy jednowymiarowy zbiór danych przedstawiony na ilustracji. Zgodnie z modelem regresji logistycznej

$$h(x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

gdzie $w = 10$ i $b = 0$. Czy to byłby ostateczny wynik uczenia zgodny z MLE?

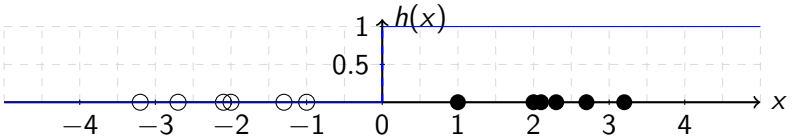


Regresja logistyczna - przykład uczenia

Rozważmy jednowymiarowy zbiór danych przedstawiony na ilustracji. Zgodnie z modelem regresji logistycznej

$$h(x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

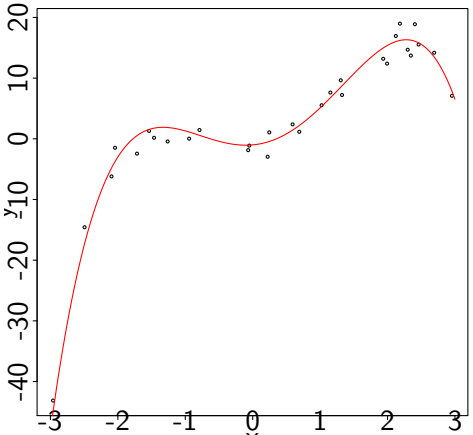
gdzie $w \to \infty$ i $b = 0$. Czy to byłby ostateczny wynik uczenia zgodny z MLE?



Przykład regresji liniowej z cechami wielomianowymi

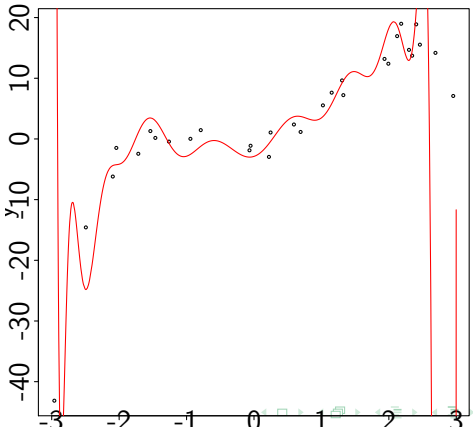
Regresja z cechami wielomianowymi 4-stopnia

$$w = [-1.1, 5, 0.95, 0.81]$$



Regresja z cechami wielomianowymi 20-stopnia

$$w = [61.30, -23.29, 22.69, -26.33, 2.03, ...]$$



Pomysł: ograniczenie przestrzeni hipotez

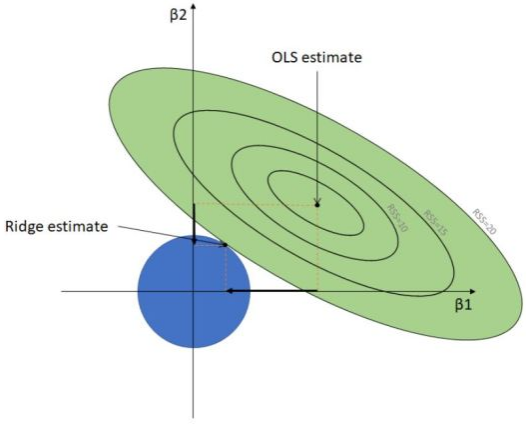
Możemy do problemu najmniejszych kwadratów dodać dodatkowe ograniczenie:

$$\min_{w} \sum_{i=1}^{n} (h(x) - y)^2$$

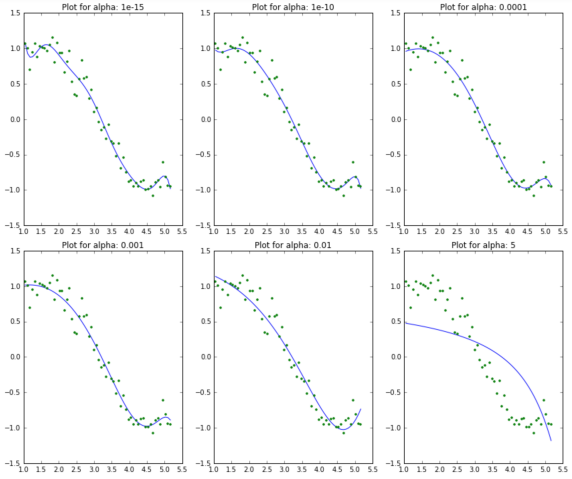
Przy ograniczeniach:

$$\sum_{i=1}^{d} w_i^2 \leq B$$

gdzie B to parametr metody - „budżet" na wartości wag.



Regresja grzbietowa



Po przekształceniach uzyskujemy:

$$\min_{w} \sum_{i=1}^{n} (h(x) - y)^2 + \lambda \sum_{i=1}^{d} w_i^2$$

gdzie $\lambda \propto \frac{1}{B}$ jest parametrem metody.

Regresja LASSO

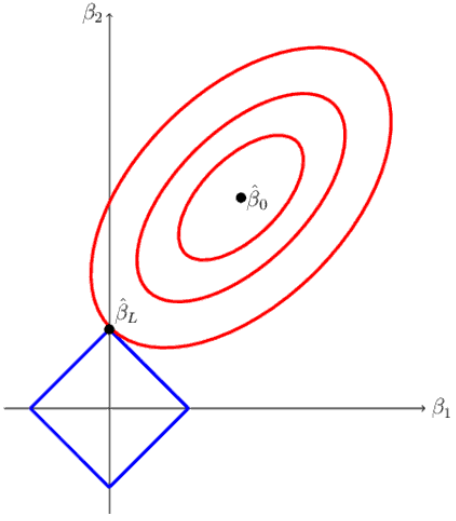
$$\min_w \sum_{i=1}^{n} (h(x) - y)^2$$

$$\sum_{i=1}^{d} |w_i| \leq B$$

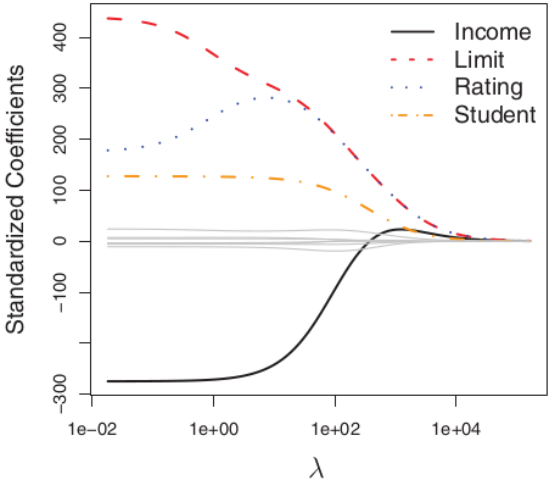
Po przekształceniach:

$$\min_w \sum_{i=1}^{n} (h(x) - y)^2 + \lambda \sum_{i=1}^{d} |w_i|$$

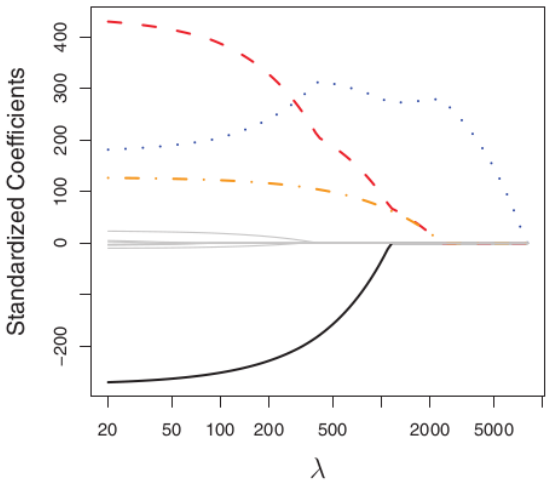
gdzie $\lambda \propto \frac{1}{B}$ jest parametrem metody.



Regresja LASSO - zaskakująca właściwość



Regresja grzbietowa



Regresja LASSO

ERM w praktyce

Zwykle modele uczące się zgodnie z ERM zawierają w funkcji celu dodatkowy term regularyzujący:

$$\arg\min_h L(h(x), y) + \lambda \cdot Complexity(h)$$

gdzie $L()$ to błąd na danych uczących a $Complexity()$ to np.

- regularyzator L2 (regresja grzbietowa) $\sum_{i=1}^{d} w_i^2$
- regularyzator L1 (regresja LASSO) $\sum_{i=1}^{d} |w_i|$
- rozmiar drzewa decyzyjnego (pruning)
- „gibkość" rozkładu prawdopodobieństwa kontrolowana poprzez rozmywanie estymat[1]
- ...

[1] Schematy rozmywania estymat można często zapisać jako MAP

Dziękuję za uwagę!

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego