Metody probabilistyczne w uczeniu maszynowym

Wykład 1: przykładowe problemy i zagadnienia uczenia maszynowego, podstawy estymacji

Katarzyna Grygiel

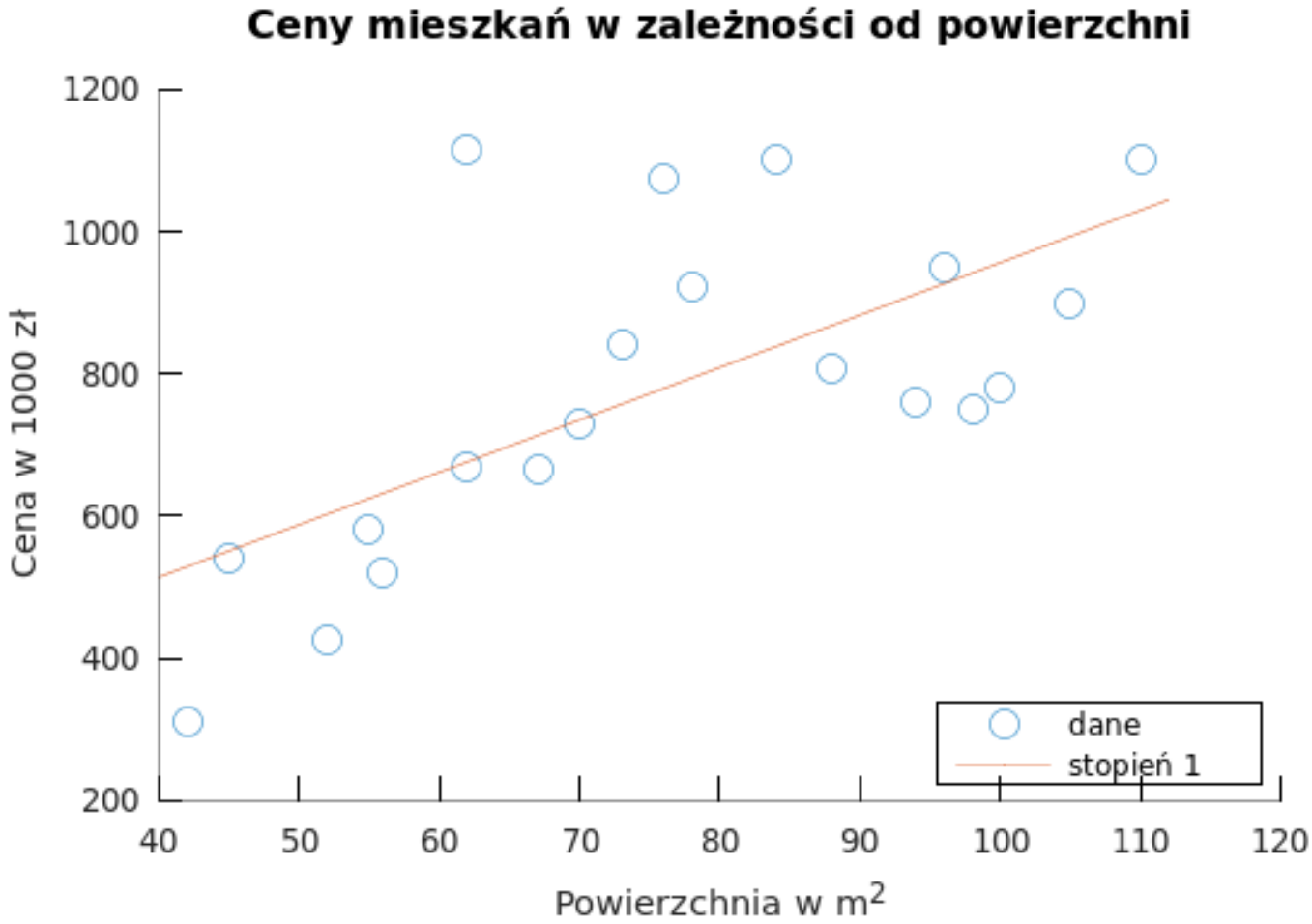
Semestr letni 2019/2020

Przykładowe problemy i rodzaje uczenia maszynowego

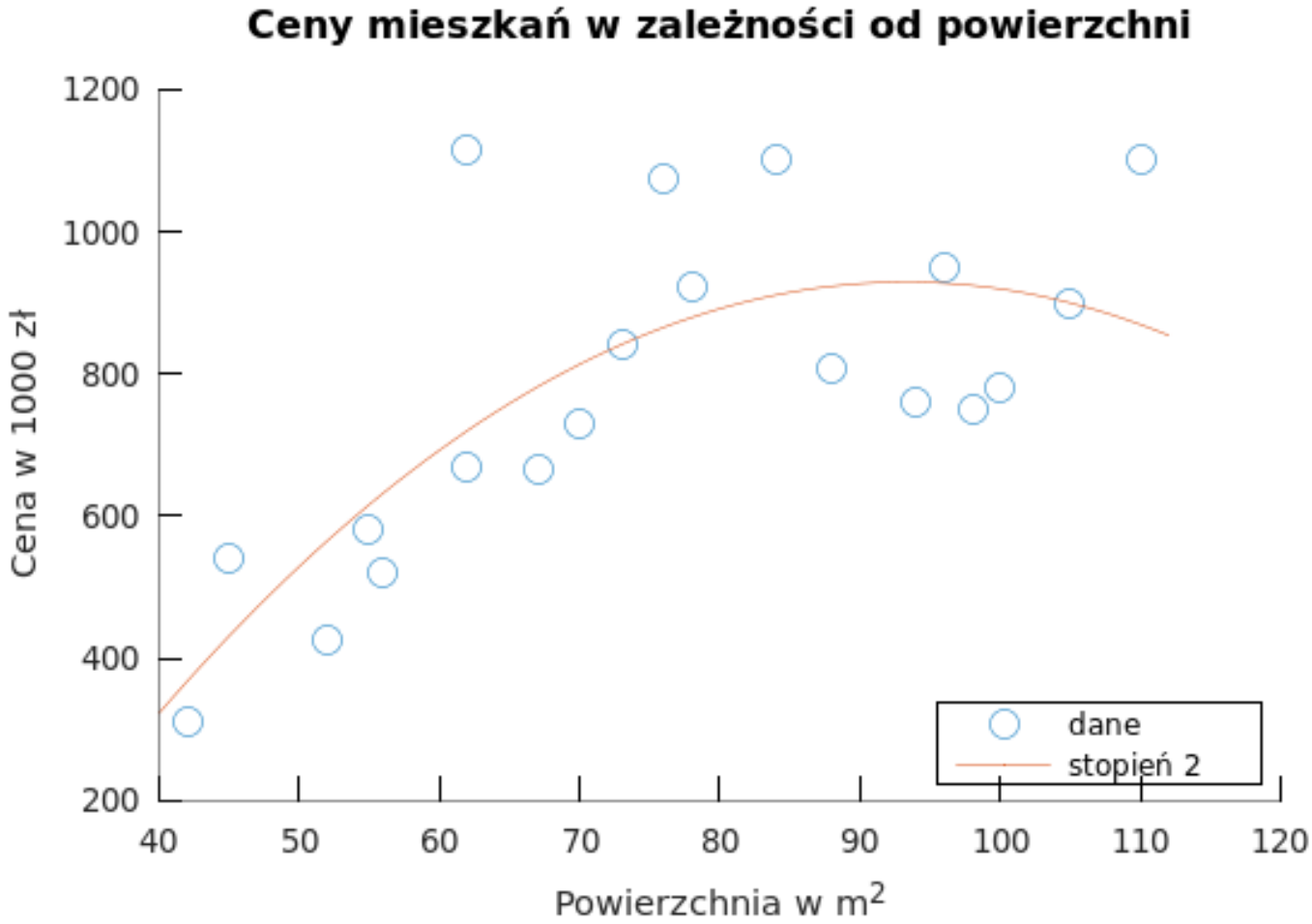
Podstawowe zagadnienia

Podejścia klasyczne i bayesowskie

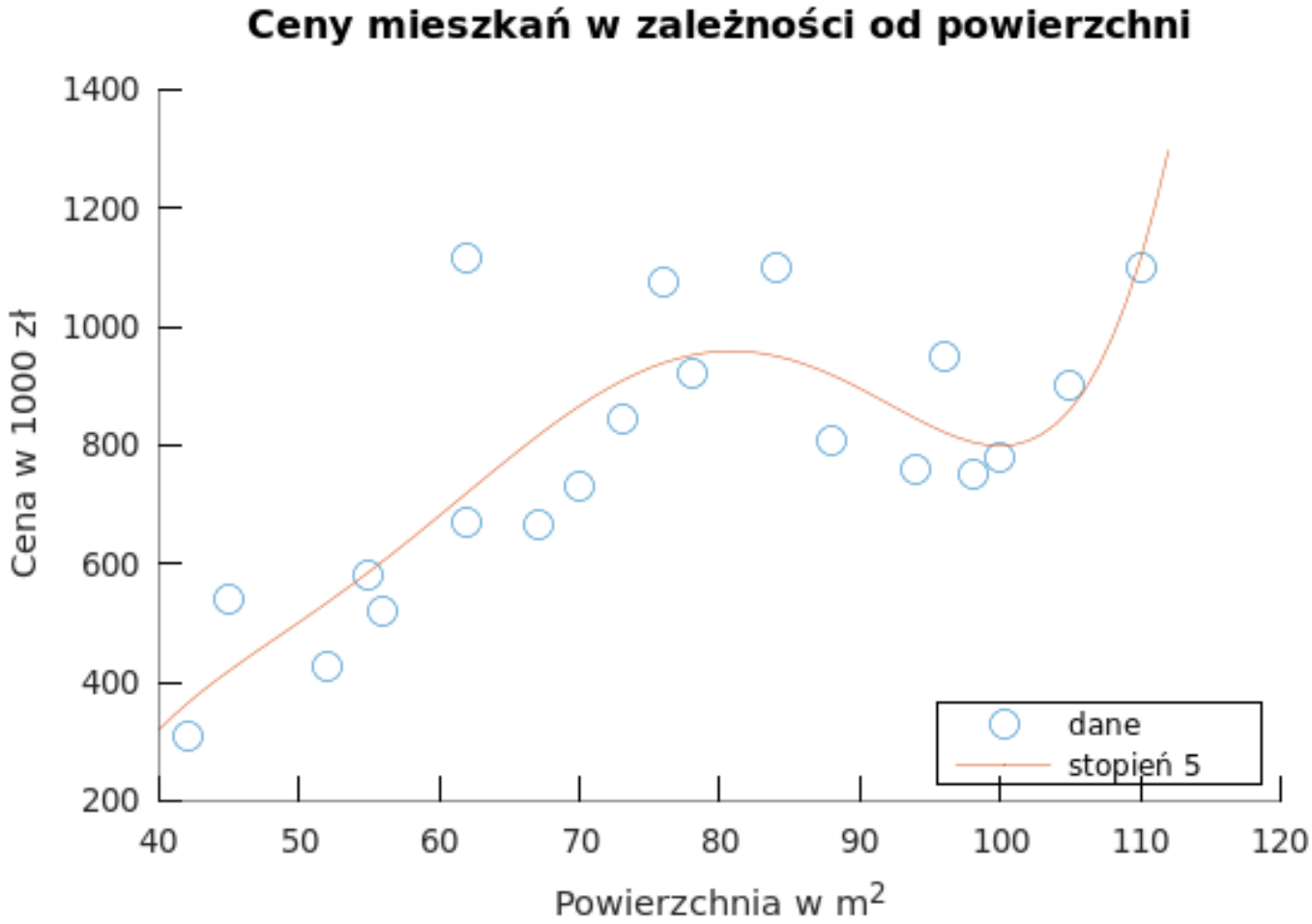
Przykład 1: prognozowanie cen mieszkań



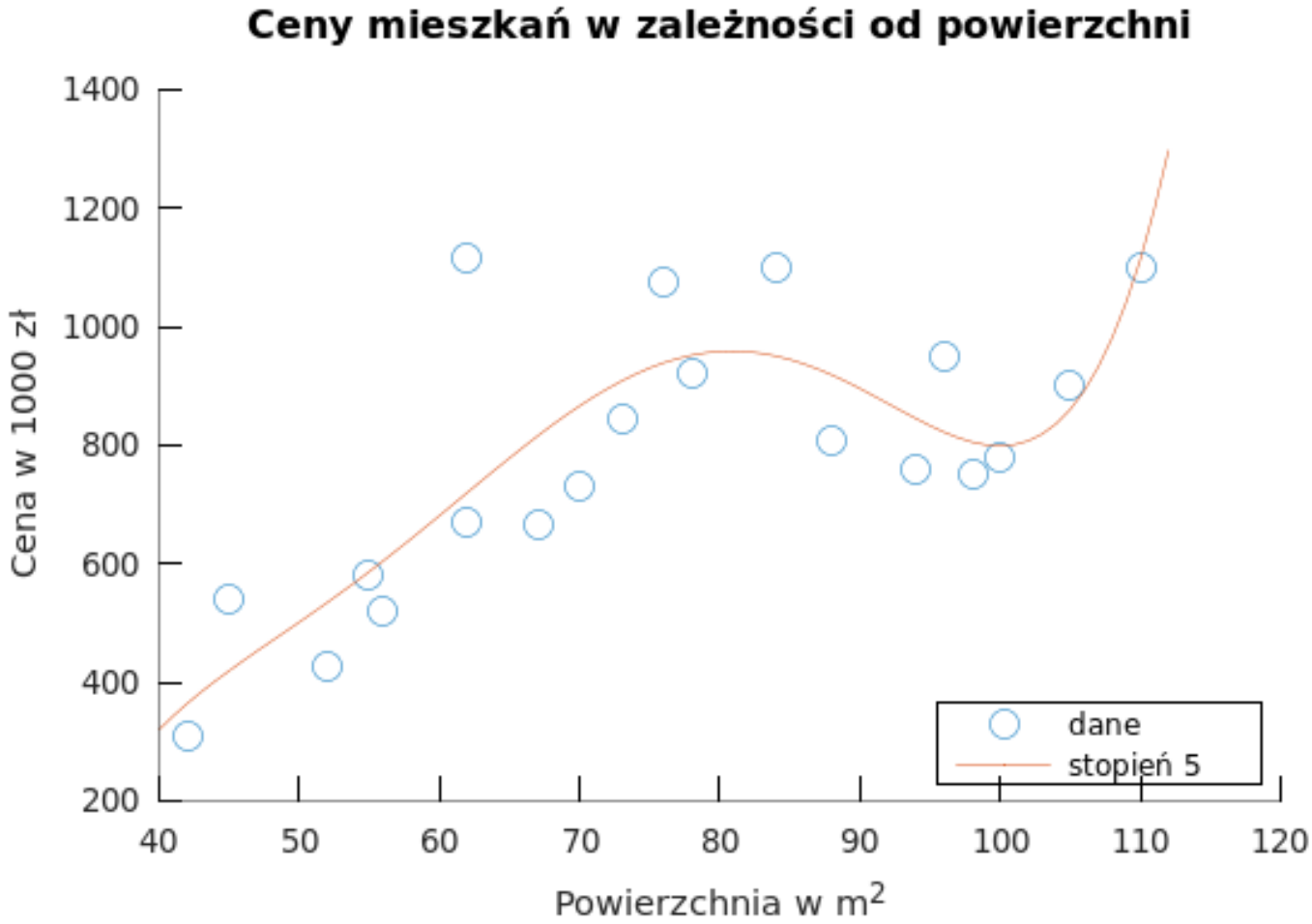
Przykład 1: prognozowanie cen mieszkań



Przykład 1: prognozowanie cen mieszkań



Przykład 1: prognozowanie cen mieszkań



Regresja

Przykład 2: spam czy nie spam?

PP

Piotr Pogoda <piotr.pogoda@im.uj.edu.pl>

Pn, 17.02.2020 12:08

pracownicy@ii.uj.edu.pl; pracownicy@im.uj.edu.pl; pracownicy@tcs.uj.edu.pl; doktoranci@ii.uj.edu.pl; doktoranci@im.uj.edu.pl +5 innych

Szanowni Państwo,

Dnia 4 marca br. Odbywać się będzie Dzień Wydziału. W związku z tym zwracam się z gorącą prośbą do Państwa o korzystanie tego dnia z garażu podziemnego.

Z poważaniem

Piotr Pogoda

A

agrarbiztositas@kekcegcsoport.hu

Nie, 16.02.2020 08:36

Recipients <agrarbiztositas@kekcegcsoport.hu>

Otrzymales darowizne w wysokosci 240000,00 EURO, wygralem loterie amerykanska o wartosci 40 milionów dolarów amerykanskich w Ameryce i postanowilem przekazac jej czesc pieciu szczesliwym ludziom i domom charytatywnym na pamiatke mojej zmarlej zony, która zmarla na raka. Skontaktuj sie ze mna, aby uzyskac wiecej informacji na: infotomcristt1@gmail.com

Przykład 2: spam czy nie spam?

PP

Piotr Pogoda <piotr.pogoda@im.uj.edu.pl>
Pn, 17.02.2020 12:08
pracownicy@ii.uj.edu.pl; pracownicy@im.uj.edu.pl; pracownicy@tcs.uj.edu.pl; doktoranci@ii.uj.edu.pl; doktoranci@im.uj.edu.pl +5 innych

Szanowni Państwo,

Dnia 4 marca br. Odbywać się będzie Dzień Wydziału. W związku z tym zwracam się z gorącą prośbą do Państwa o korzystanie tego dnia z garażu podziemnego.

Z poważaniem

Piotr Pogoda

A

agrarbiztositas@kekcegcsoport.hu
Nie, 16.02.2020 08:36
Recipients <agrarbiztositas@kekcegcsoport.hu>

Otrzymales darowizne w wysokosci 240000,00 EURO, wygralem loterie amerykanska o wartosci 40 milionów dolarów amerykanskich w Ameryce i postanowilem przekazac jej czesc pieciu szczesliwym ludziom i domom charytatywnym na pamiatke mojej zmarlej zony, która zmarla na raka. Skontaktuj sie ze mna, aby uzyskac wiecej informacji na: infotomcristt1@gmail.com

Klasyfikacja binarna

Przykład 3: grupowanie irysów

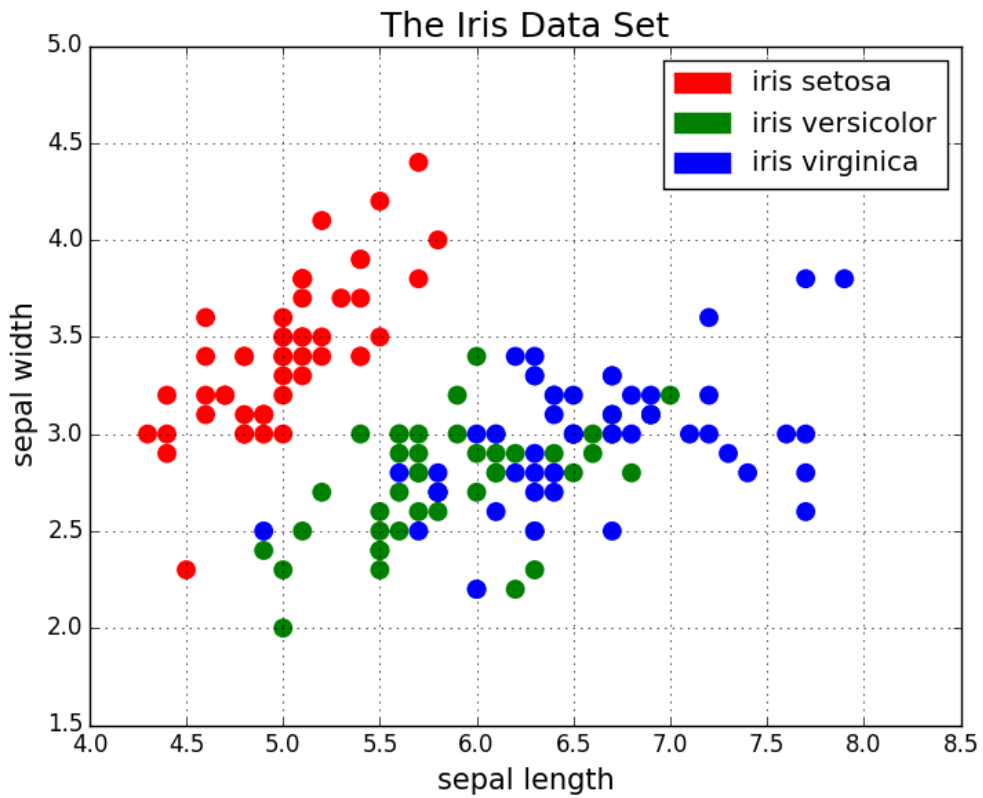


```
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
```

Przykład 3: grupowanie irysów



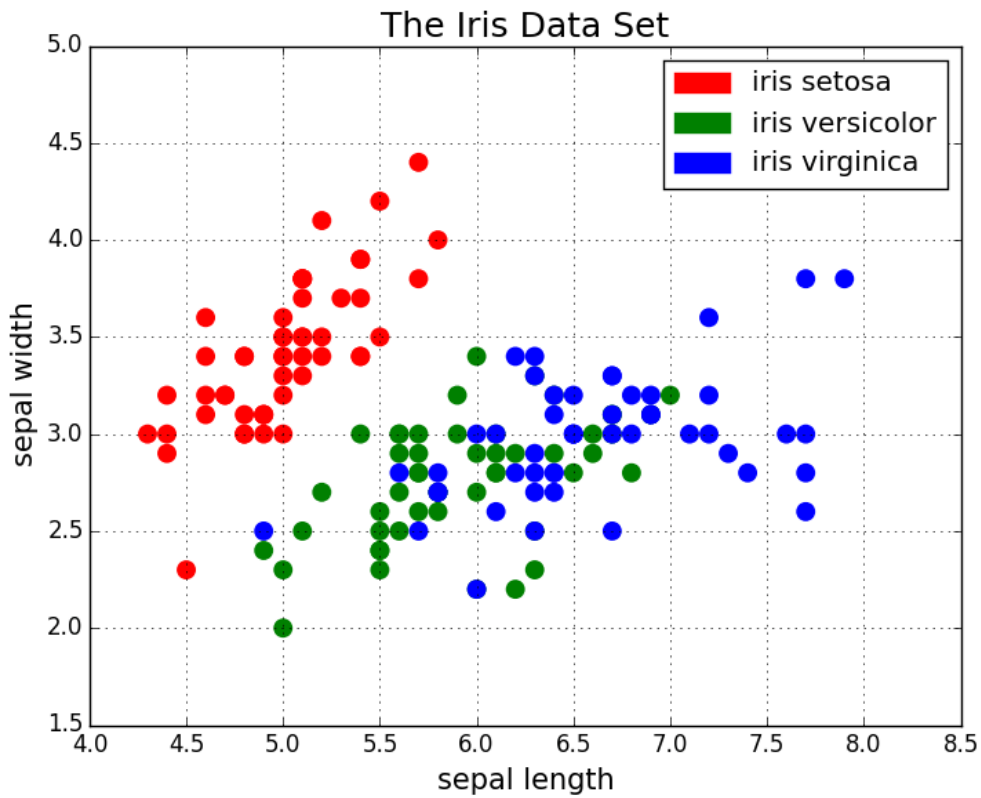
```
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
```



Przykład 3: grupowanie irysów

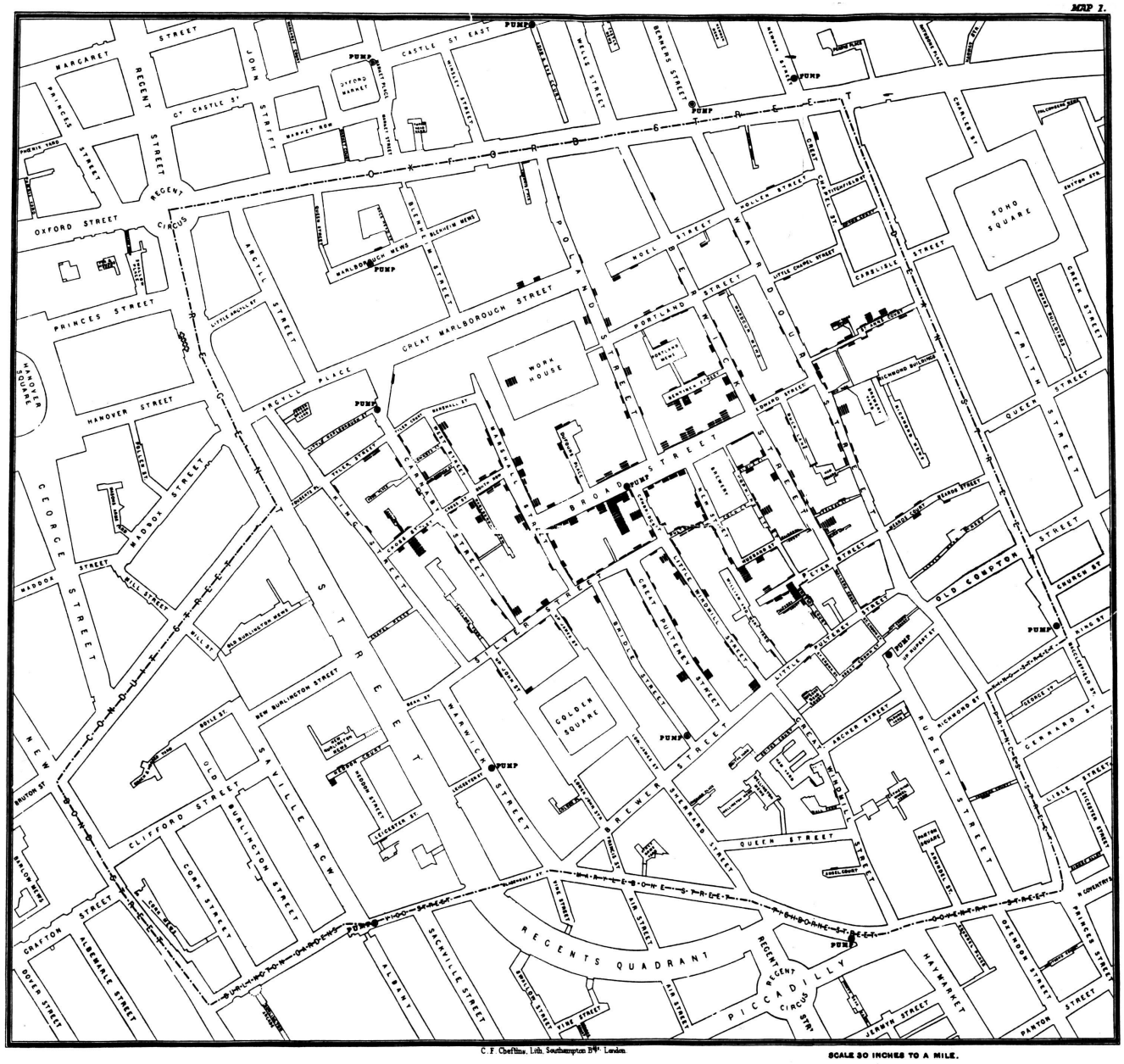


```
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
```



Klasyfikacja wieloklasowa

Przykład 4: identyfikacja źródła cholery



Przykład 4: identyfikacja źródła cholery

Przykład 4: identyfikacja źródła cholery

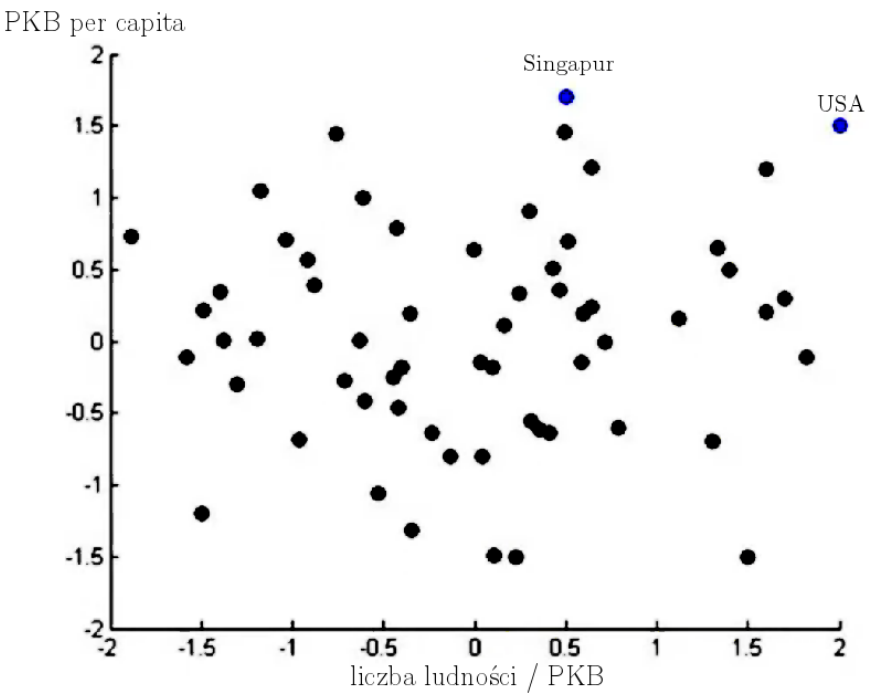
Klasteryzacja

Przykład 5: wizualizacja danych

Kraj	PKB	PKB per capita	Wskaźnik rozwoju	Średnia dł. życia	Wskaźnik ubóstwa	Średni dochód	...
Kanada	1.577	39.17	0.908	80.7	32.6	67.293	...
Chiny	5.878	7.54	0.687	73	46.9	10.22	...
Indie	1.632	3.41	0.547	64.7	36.8	0.735	...
Rosja	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapur	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...

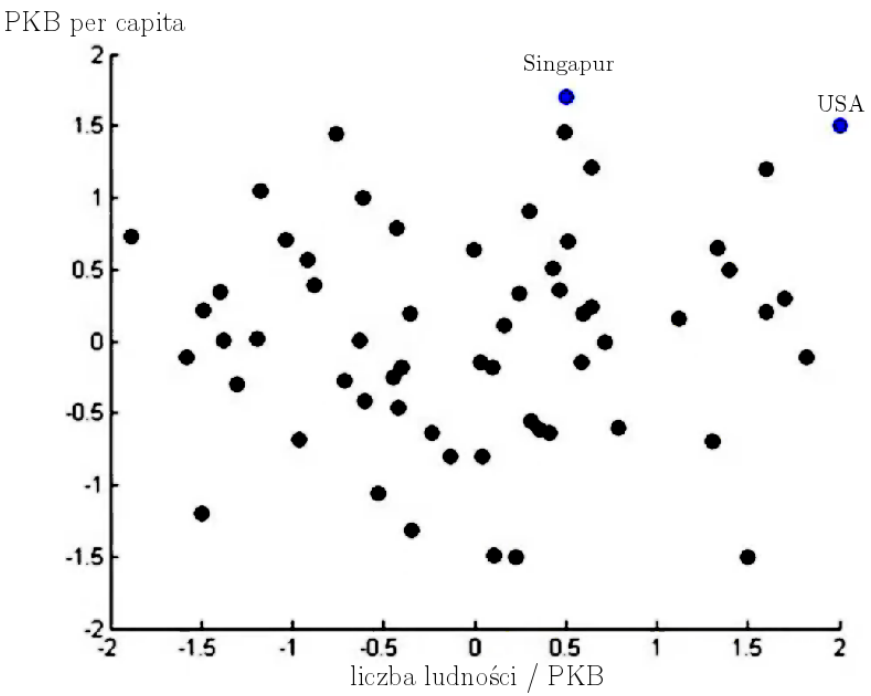
Przykład 5: wizualizacja danych

Kraj	PKB	PKB per capita	Wskaźnik rozwoju	Średnia dł. życia	Wskaźnik ubóstwa	Średni dochód	...
Kanada	1.577	39.17	0.908	80.7	32.6	67.293	...
Chiny	5.878	7.54	0.687	73	46.9	10.22	...
Indie	1.632	3.41	0.547	64.7	36.8	0.735	...
Rosja	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapur	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...



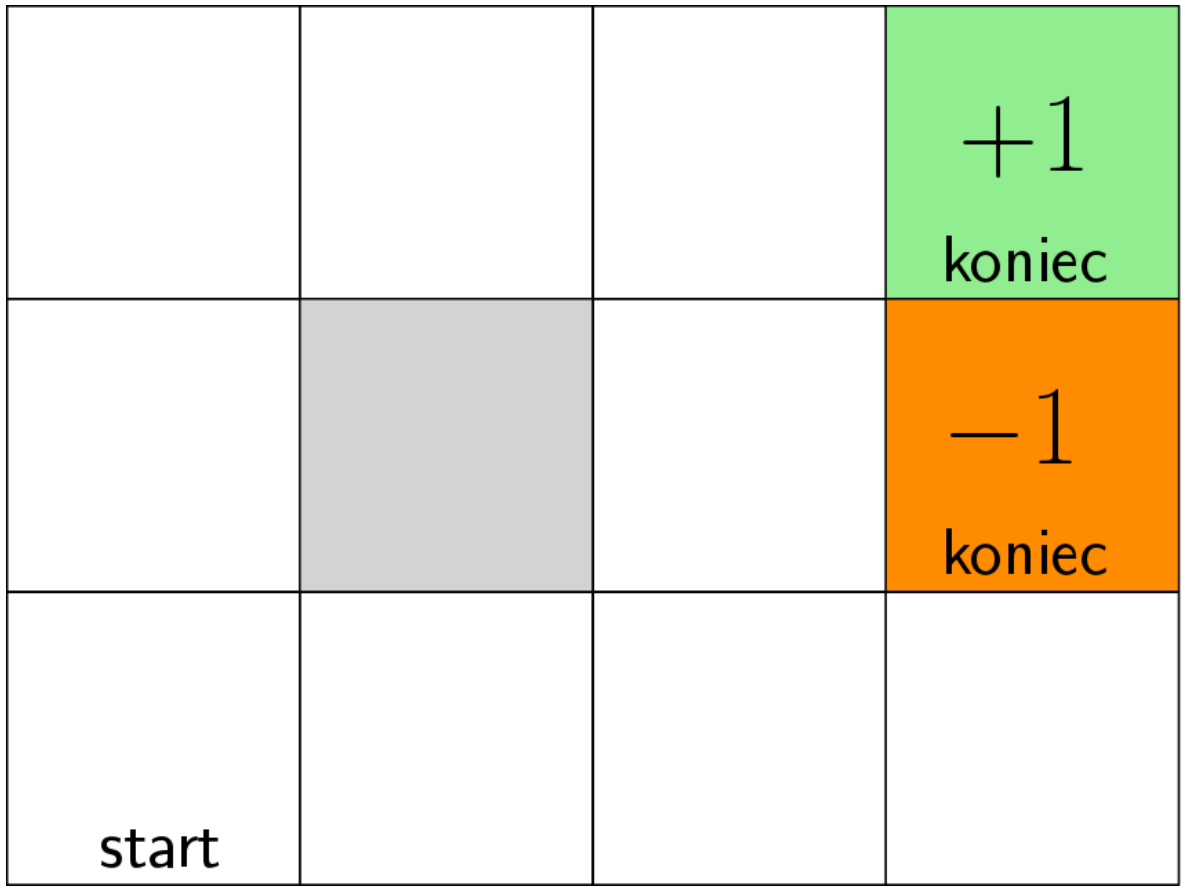
Przykład 5: wizualizacja danych

Kraj	PKB	PKB per capita	Wskaźnik rozwoju	Średnia dł. życia	Wskaźnik ubóstwa	Średni dochód	...
Kanada	1.577	39.17	0.908	80.7	32.6	67.293	...
Chiny	5.878	7.54	0.687	73	46.9	10.22	...
Indie	1.632	3.41	0.547	64.7	36.8	0.735	...
Rosja	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapur	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...

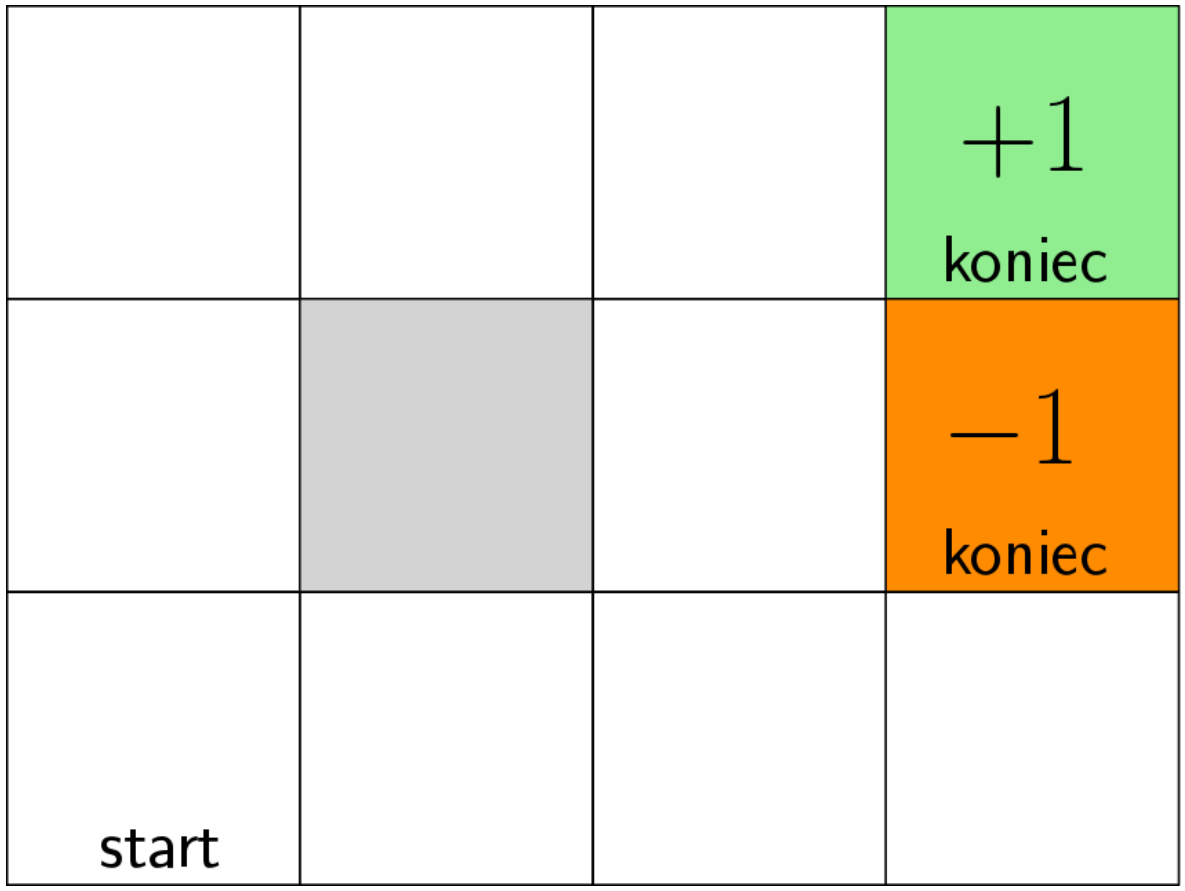


Redukcja wymiarów

Przykład 6.



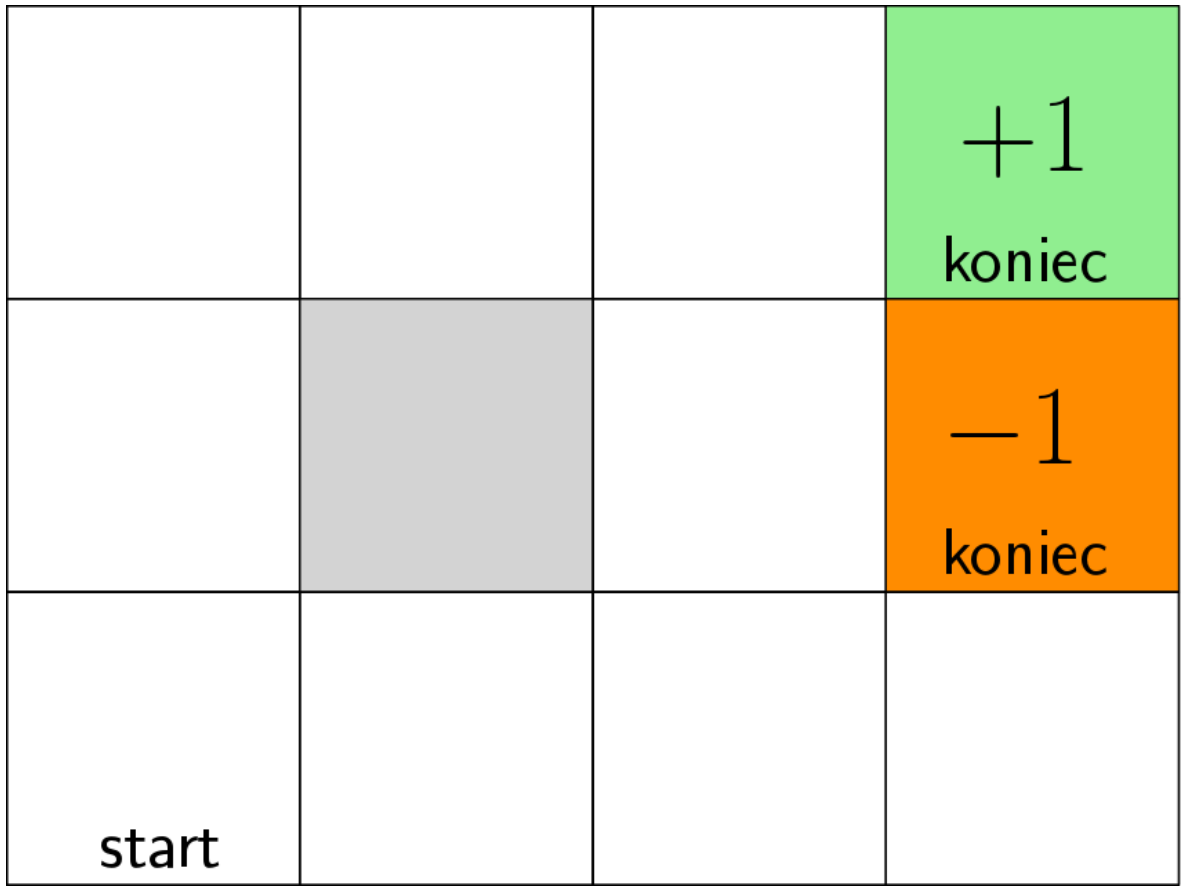
Przykład 6.



Akcje:
$\uparrow, \leftarrow, \downarrow, \rightarrow$

Koszt akcji:
$-0,04$

Przykład 6.



Akcje:
$\uparrow, \leftarrow, \downarrow, \rightarrow$

Koszt akcji:
$-0,04$

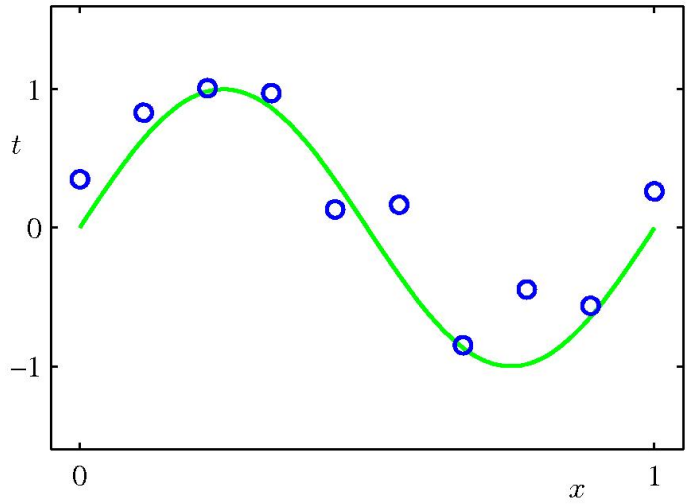
Uczenie ze wzmacnianiem

Rodzaje uczenia maszynowego

- Uczenie nadzorowane (*supervised learning*)
  - regresja
    cena mieszkania w zależności od metrażu
  - klasyfikacja binarna i wieloklasowa
    spam/nie spam, rozpoznawanie cyfr
  - tworzenie rankingów
    lista preferencji towarów dla danej grupy klientów
- Uczenie nienadzorowane (*unsupervised learning*)
  - klasteryzacja
    segmentacja obrazu, grupowanie danych
  - redukcja wymiarów
    wizualizacja danych, uproszczenie modelu
- Uczenie ze wzmacnianiem (*reinforcement learning*)
  - nawigacja, gaming

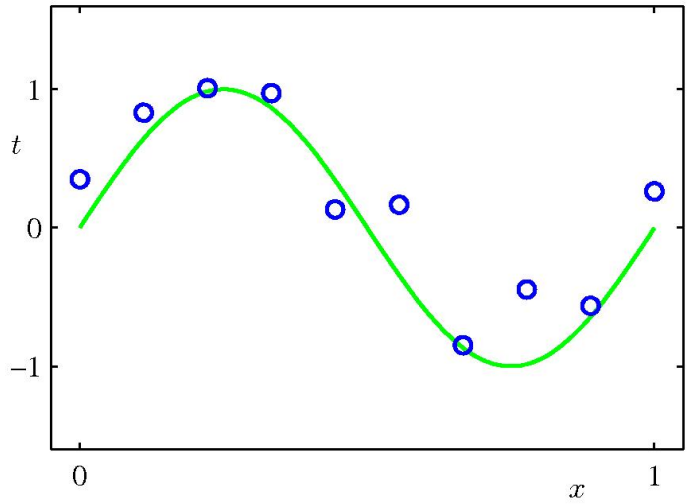
Wielomianowe dopasowanie krzywej

Mamy zbiór treningowy $\{(x^{(i)}, y^{(i)}) : i = 1, \ldots, m\}$.
Na jego podstawie chcemy przewidzieć, jaka będzie wartość y dla nowego argumentu x.



Wielomianowe dopasowanie krzywej

Mamy zbiór treningowy $\{(x^{(i)}, y^{(i)}) : i = 1, \ldots, m\}$.
Na jego podstawie chcemy przewidzieć, jaka będzie wartość y dla nowego argumentu x.

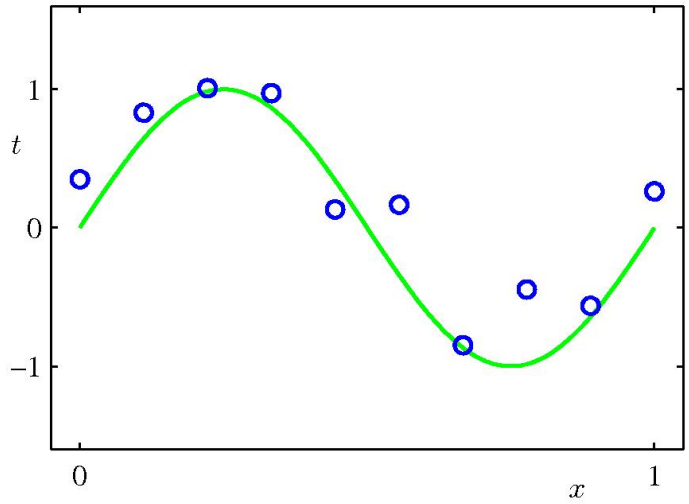


Rozważmy następujący model:

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_M x^M = \sum_{i=0}^{M} \theta_i x^i.$$

Wielomianowe dopasowanie krzywej

Mamy zbiór treningowy $\{(x^{(i)}, y^{(i)}) : i = 1, \ldots, m\}$.
Na jego podstawie chcemy przewidzieć, jaka będzie wartość y dla nowego argumentu x.



Rozważmy następujący model:

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_M x^M = \sum_{i=0}^{M} \theta_i x^i.$$

- Jaki stopień wielomianu M wybrać?
- Jak wyznaczyć parametry θ_i?
- Jak ocenić trafność modelu?

Wyznaczanie parametrów modelu

Dla modelu

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_M x^M$$

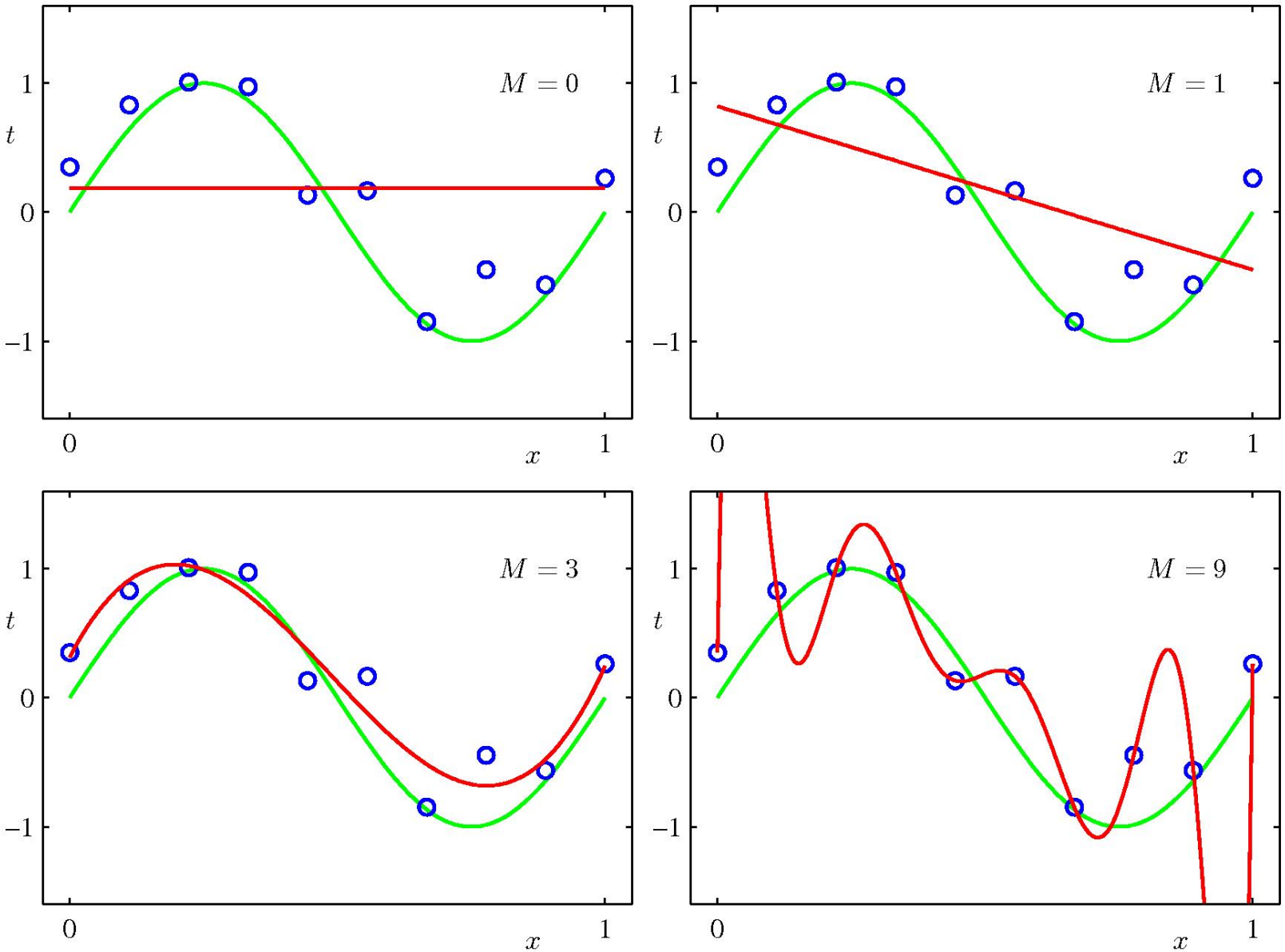
definiujemy funkcję kosztu

$$\mathcal{J}(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2.$$

Możliwe rozwiązanie polega na wyznaczeniu parametru $\theta = [\theta_0, \ldots, \theta_M]^T$ minimalizującego funkcję kosztu, czyli

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^{M+1}} \mathcal{J}(\theta).$$

Wyznaczanie stopnia wielomianu



Wyznaczanie stopnia wielomianu

Gdy stopień wielomianu jest zbyt niski, to mamy do czynienia ze zjawiskiem *niedouczenia* (*underfitting*). Mówimy wtedy, że model ma duże obciążenie.
Gdy stopień wielomianu jest zbyt wysoki, to mamy do czynienia ze zjawiskiem *przeuczenia* (*overfitting*). Mówimy wtedy, że model ma dużą wariancję.

Wyznaczanie stopnia wielomianu

Gdy stopień wielomianu jest zbyt niski, to mamy do czynienia ze zjawiskiem *niedouczenia* (*underfitting*). Mówimy wtedy, że model ma duże obciążenie.
Gdy stopień wielomianu jest zbyt wysoki, to mamy do czynienia ze zjawiskiem *przeuczenia* (*overfitting*). Mówimy wtedy, że model ma dużą wariancję.

No free lunch theorem

Nie istnieje najlepszy uniwersalny algorytm optymalizacyjny (dla wszystkich zadań). Niezależnie od miary jakości algorytmu optymalizacyjnego, dowolne dwa różne algorytmy optymalizacyjne zachowują się „średnio” tak samo dla wszystkich zadań optymalizacyjnych.

Wyznaczanie stopnia wielomianu

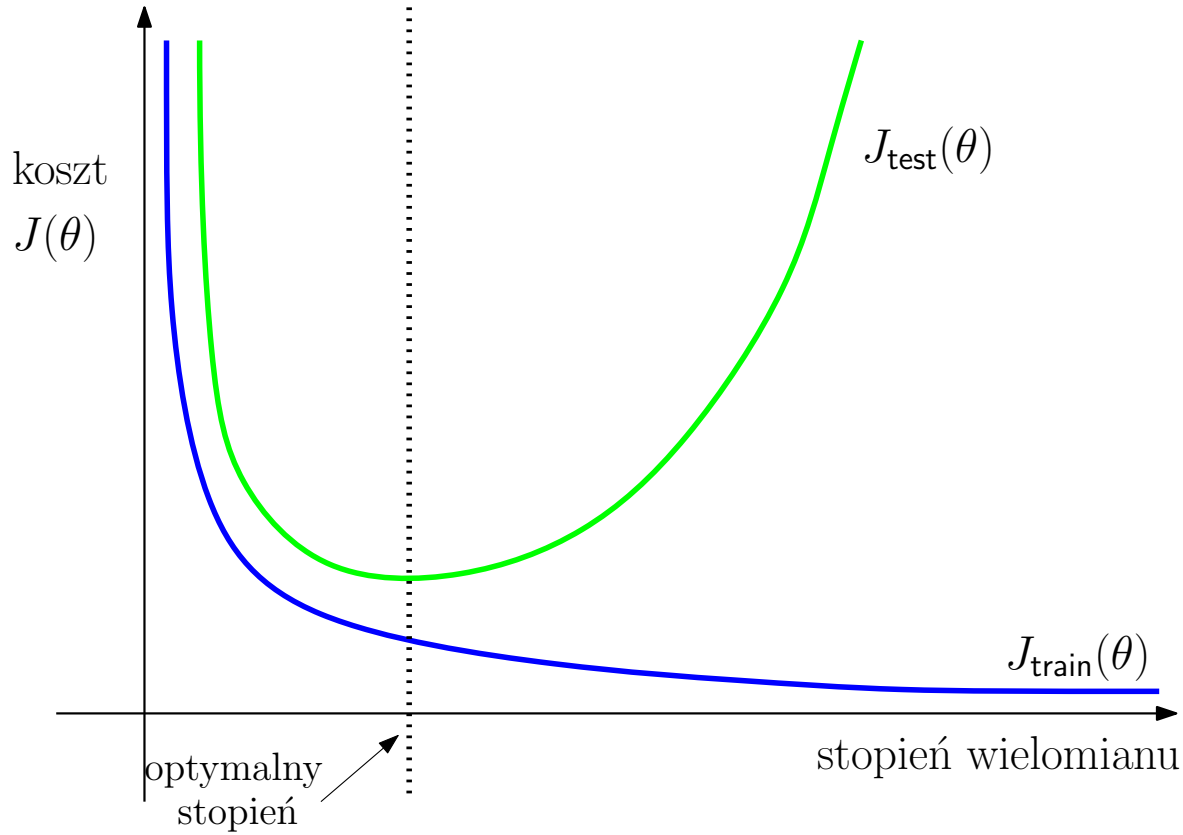
Gdy stopień wielomianu jest zbyt niski, to mamy do czynienia ze zjawiskiem *niedouczenia* (*underfitting*). Mówimy wtedy, że model ma duże obciążenie.
Gdy stopień wielomianu jest zbyt wysoki, to mamy do czynienia ze zjawiskiem *przeuczenia* (*overfitting*). Mówimy wtedy, że model ma dużą wariancję.

No free lunch theorem

Nie istnieje najlepszy uniwersalny algorytm optymalizacyjny (dla wszystkich zadań). Niezależnie od miary jakości algorytmu optymalizacyjnego, dowolne dwa różne algorytmy optymalizacyjne zachowują się „średnio” tak samo dla wszystkich zadań optymalizacyjnych.

All models are wrong, but some are useful. (George Box)

Kompromis między wariancją a obciążeniem



Zasada brzytwy Ockhama

Spośród hipotez, które jednakowo dobrze wyjaśniają dane zjawisko, wybierz najprostszą – tę, która wymaga dokonania najmniejszej liczby założeń.

Regularyzacja

Aby uniknąć zjawiska przeuczenia możemy zmienić funkcję kosztu uwzględniając regularyzację:

$$\mathcal{J}(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + \frac{\lambda}{2} \|\theta\|^2.$$

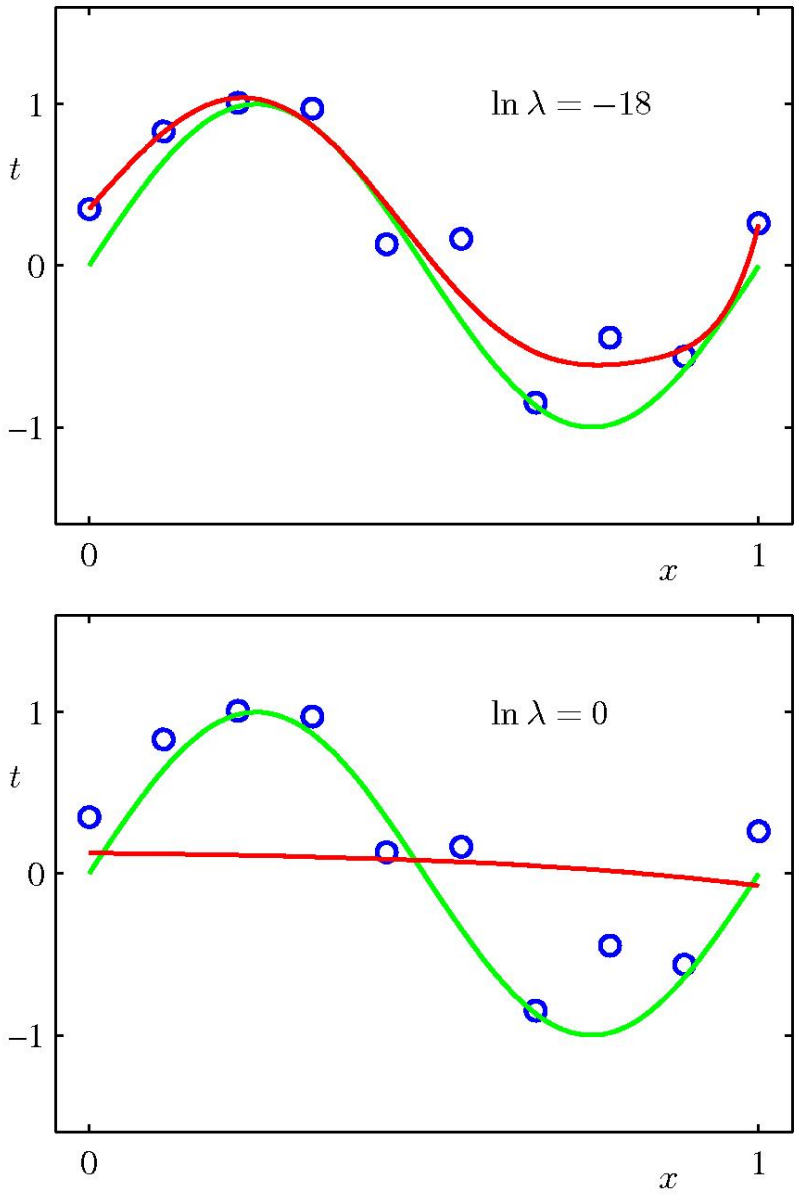
Regularyzacja

Aby uniknąć zjawiska przeuczenia możemy zmienić funkcję kosztu uwzględniając regularyzację:

$$\mathcal{J}(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2}\|\theta\|^2.$$

	$\lambda = 0$	$\lambda = \exp(-18)$	$\lambda = 1$
θ_0	0.35	0.35	0.13
θ_1	232.37	4.74	−0.05
θ_2	−5321.83	−0.77	−0.06
θ_3	48568.31	−31.97	−0.05
θ_4	−231639.30	−3.89	−0.03
θ_5	640042.26	55.28	−0.02
θ_6	−1061800.52	41.32	−0.01
θ_7	1042400.18	−45.95	−0.00
θ_8	−557682.99	−91.53	0.00
θ_9	125201.43	72.68	0.01

Regularyzacja



Skalowanie cech

Często dane zawierają cechy o różnej wielkości, jednostkach i zakresie. Wtedy jedna cecha, która wyraża się w bardzo dużej wielkości, może wpłynąć na przewidywanie o wiele bardziej niż pozostałe.
Przykładowe skalowania dla danych $x = \{x^{(i)} : i = 1, \ldots, m\}$:

- min-max 1

$$x^{(i)} \mapsto \frac{x^{(i)} - \min(x)}{\max(x) - \min(x)},$$

- min-max 2

$$x^{(i)} \mapsto \frac{x^{(i)} - m(x)}{\max(x) - \min(x)},$$

- standaryzacja

$$x^{(i)} \mapsto \frac{x^{(i)} - m(x)}{s(x)},$$

gdzie $m(x)$ to średnia, a $s(x)$ to odchylenie standardowe dla danych.

Co zrobić, gdy jest źle?

Błędom w przewidywaniach można zaradzić na kilka sposobów:

- zwiększając zbiór treningowy,
- zmniejszając zbiór cech,
- dodając nowe cechy,
- zmieniając charakter cech (np. stosując cechy wielomianowe zamiast liniowych),
- zwiększając lub zmniejszając parametr regularyzacji…

Nie należy jednak próbować takich rozwiązać losowo!

Ocena hipotezy

Nasza hipoteza może mieć mały błąd na danych treningowych, ale być zupełnie nietrafiona (np. z powodu przeuczenia).

Zbiór danych, na których algorytm się uczy, możemy podzielić na dwa podzbiory: zbiór treningowy i zbiór testowy, a następnie:

- wyznaczyć θ minimalizując błąd na zbiorze treningowym $\mathcal{J}_{train}(\theta)$,
- obliczyć błąd na zbiorze testowym $\mathcal{J}_{test}(\theta)$.

Wybór modelu i zbiory treningowe/walidacyjne/testowe

To, że algorytm uczący dobrze działa na danych treningowych, nie oznacza wcale, że nasza hipoteza jest dobra.
Błąd mierzony na danych, na podstawie których wyznaczyliśmy parametry modelu, będzie niemal zawsze mniejszy niż na jakichkolwiek innych danych.
Bez zbioru walidacyjnego (zła praktyka!):

- korzystając ze zbioru treningowego zoptymalizuj parametry θ dla wszystkich możliwych stopni wielomianu;
- znajdź stopień wielomianu d o najmniejszym błędzie na zbiorze testowym;
- oszacuj błąd uogólniony $\mathcal{J}_{test}(\theta^{(d)})$ na zbiorze testowym.

Do wyznaczenia stopnia d został użyty zbiór testowy.
Błąd będzie większy dla innych danych.

Zbiory treningowe/walidacyjne/testowe

Użyj zbioru walidacyjnego (etap pośredni), aby wyznaczyć stopień wielomianu. Dzięki temu zbiór testowy da nam odpowiedni i nie nazbyt mały błąd.
Przykładowy podział danych:

zbiór treningowy	60%
zbiór walidacyjny	20%
zbiór testowy	20%

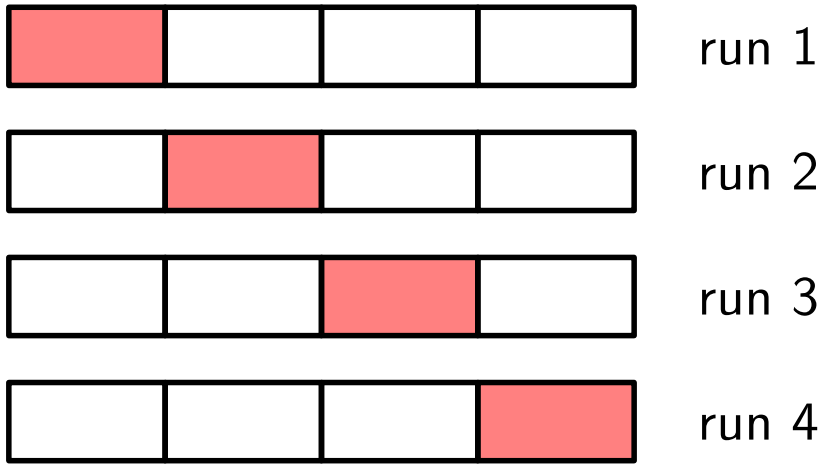
Następnie wyznaczamy błąd na każdym ze zbiorów.
Ze zbiorem walidacyjnym:

- korzystając ze zbioru treningowego zoptymalizuj parametry θ dla wszystkich możliwych stopni wielomianu;
- znajdź stopień wielomianu d o najmniejszym błędzie na zbiorze walidacyjnym;
- oszacuj błąd uogólniony $\mathcal{J}_{test}(\theta^{(d)})$ na zbiorze testowym.

Zbiór testowy nie został użyty do wyznaczenia parametru d.

Co zrobić, gdy mamy mało danych?

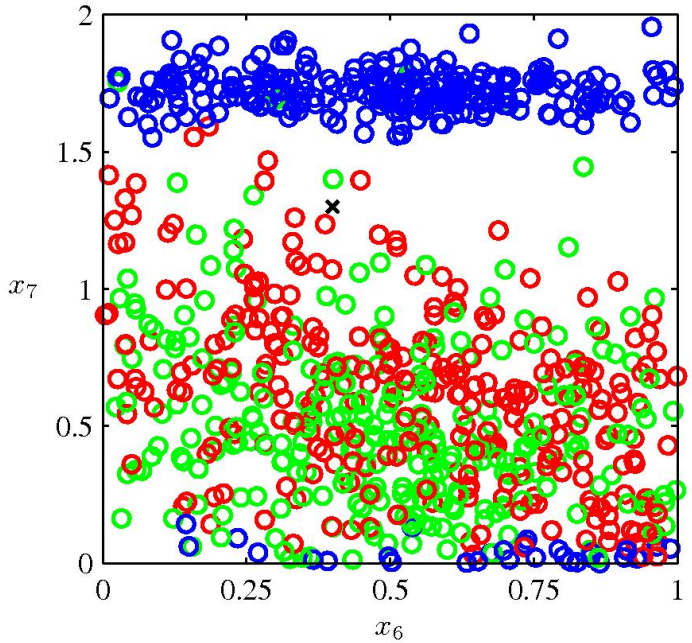
Oryginalną próbę dzielimy na K podzbiorów. Następnie kolejno każdy z nich traktujemy jako zbiór testowy, a pozostałe $K-1$ jako zbiór treningowy. Analizę wykonujemy zatem K razy. Ostatecznie K wyników uśredniamy. Taką metodę nazywamy *K-krotną walidacją krzyżową*.



Jeśli $K = m$, gdzie m jest liczebnością próby (czyli każdy podzbiór jest jednoelementowy), to taką walidację nazywamy walidacją *leave-one-out*. Stosuje się ją dla bardzo małych zbiorów danych.

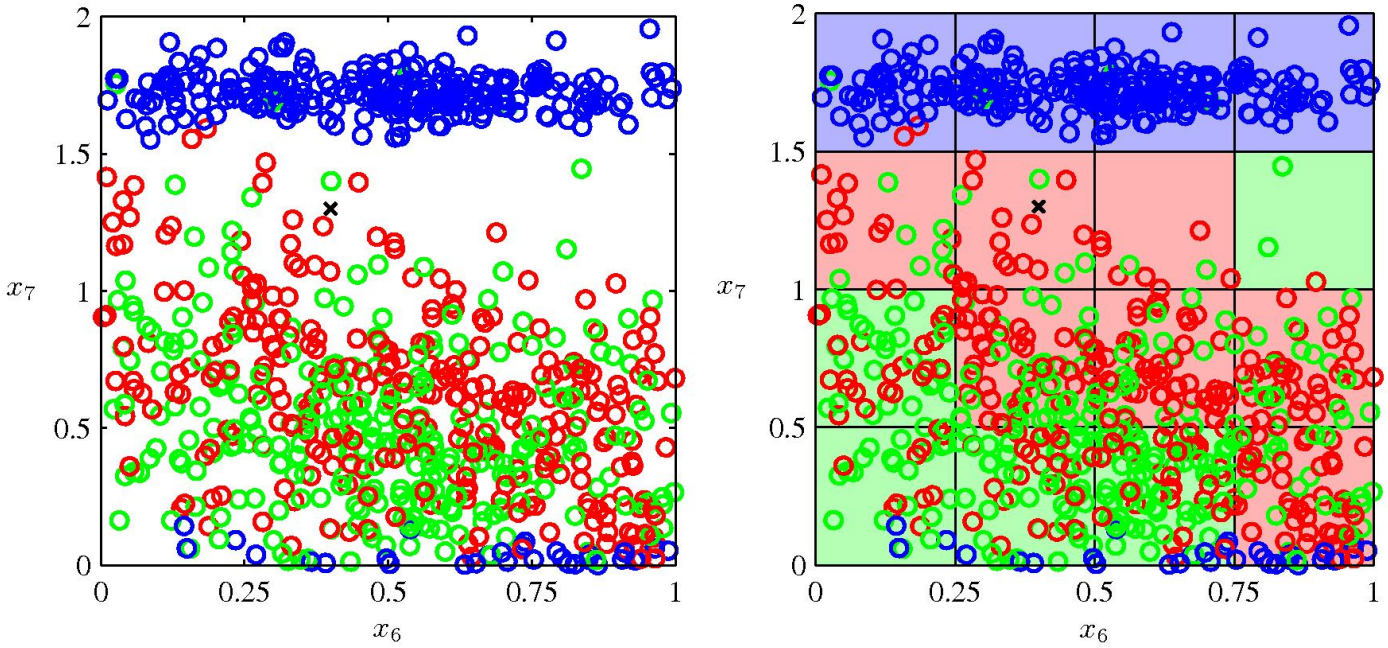
Przekleństwo wielowymiarowości

Klasyfikacja na podstawie dwóch z dwunastu cech



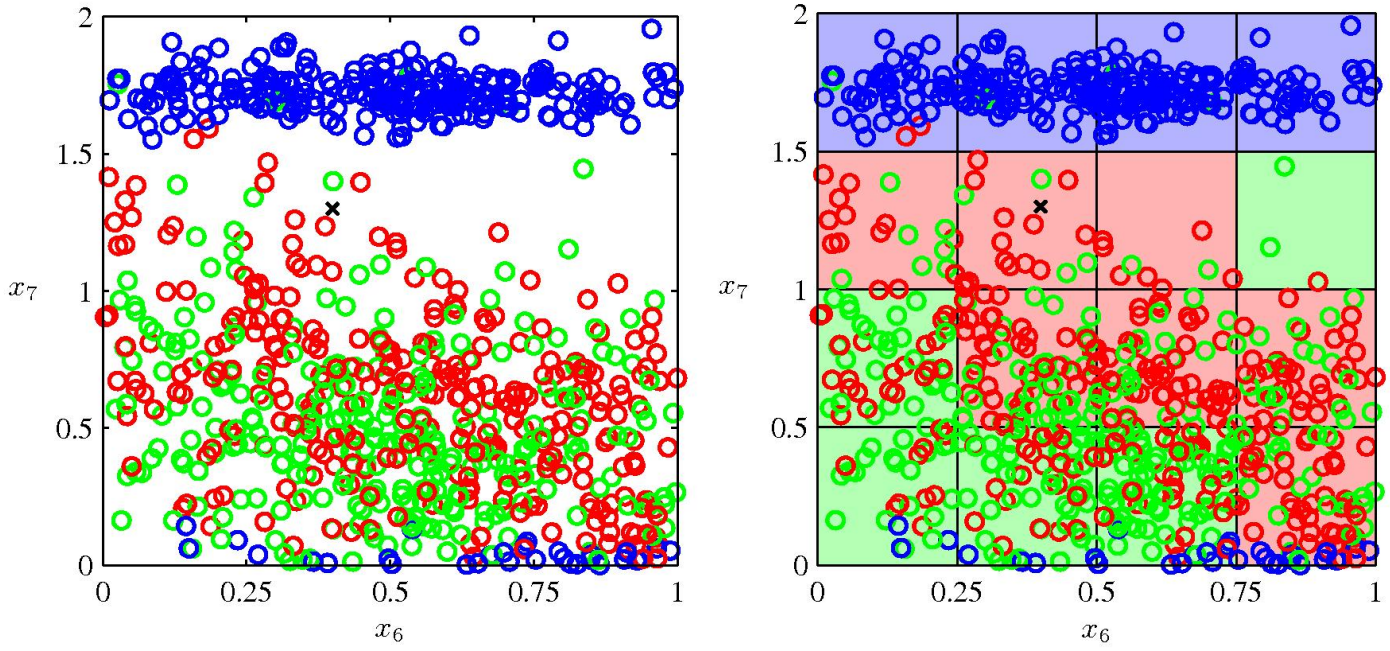
Przekleństwo wielowymiarowości

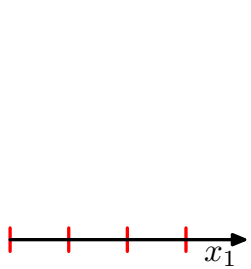
Klasyfikacja na podstawie dwóch z dwunastu cech



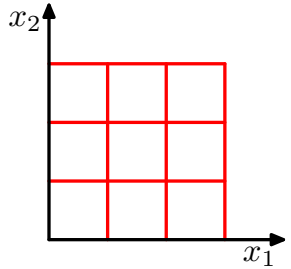
Przekleństwo wielowymiarowości

Klasyfikacja na podstawie dwóch z dwunastu cech

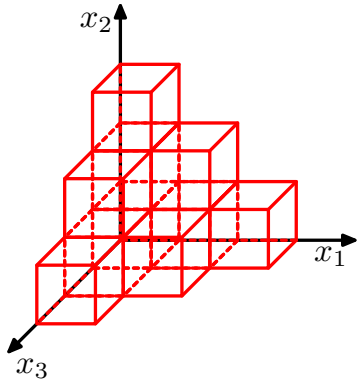




$D = 1$



$D = 2$



$D = 3$

Modele parametryczne

Model parametryczny to zbiór rozkładów prawdopodobieństwa

$$\{p(y|\theta) : \theta \in \Theta\}$$

indeksowanych *parametrem* θ z *przestrzeni parametrów* Θ.

rozkład	parametry
$\mathsf{Bern}(p)$	$\theta = p$
$\mathsf{Poi}(\lambda)$	$\theta = \lambda$
$\mathsf{Uni}(a, b)$	$\theta = (a, b)$
$\mathcal{N}(\mu, \sigma^2)$	$\theta = (\mu, \sigma^2)$
$Y = mX + b$	$\theta = (m, b)$

Uwaga dotycząca notacji

Dla rozkładów dyskretnych $p(\cdot)$ oznaczamy prawdopodobieństwo, natomiast dla rozkładów ciągłych $p(\cdot)$ jest funkcją gęstości prawdopodobieństwa.

Estymatory

Estymatorem nazywamy parametr obliczony dla próby, na podstawie którego szacujemy prawdziwą wartość parametru w populacji.

Estymator nazywamy *nieobciążonym*, gdy wartość oczekiwana rozkładu estymatora jest równa wartości szacowanego parametru:

$$\mathbb{E}(\widehat{\theta}) = \theta.$$

Jeśli różnica pomiędzy wartością oczekiwaną rozkładu estymatora a wartością szacowanego parametru jest zależna od estymatora:

$$\mathbb{E}(\widehat{\theta}) - \theta = \text{bias}(\widehat{\theta}),$$

to estymator nazywamy *obciążonym*, natomiast różnicę nazywamy *obciążeniem* estymatora.

Twierdzenie Bayesa

Mamy zbiór obserwacji $D = \{x_1, \ldots, x_M\}$ oraz model zależny od parametru θ opisujący rozkład, z które pochodzą dane. Wtedy

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

- $p(\theta)$ to *prawdopodobieństwo a priori* (przekonanie o parametrze θ przed obserwacją danych),
- $p(D|\theta)$ to *wiarygodność* lub *rozkład próby* (prawdopodobieństwo warunkowe uzyskania danych D z modelu o parametrze θ),
- $p(\theta|D)$ to *prawdopodobieństwo a posteriori* (niepewność dotycząca parametru θ po obserwacjach D),
- $p(D)$ to *wiarygodność brzegowa* (rozkład danych w gęstości brzegowej względem parametrów, tj. $\int_\Theta p(D|\theta)p(\theta)d\theta$).

a posteriori $\propto$ wiarygodność $\times$ a priori

Rozkłady sprzężone

Rodzinę rozkładów a priori nazywamy *sprzężoną* do rodziny rozkładów próby, jeżeli każdy rozkład a posteriori należy również do rodziny rozkładów a priori.

rozkład próby	a priori	a posteriori
dwumianowy	beta	beta
Poissona	gamma	gamma
geometryczny	beta	beta
wielomianowy	Dirichleta	Dirichleta
wykładniczy	gamma	gamma
normalny	normalny	normalny

Aby uprościć obliczenia przy wyznaczaniu prawdopodobieństwa a posteriori, często wybieramy rozkład a priori w ten sposób, aby był on sprzężony z rozkładem próby.

Rozkład beta

Rozkład beta to ciągły rozkład prawdopodobieństwa dany funkcją gęstości zdefiniowaną na przedziale $[0, 1]$ wzorem

$$\mathrm{Beta}(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1},$$

gdzie $\alpha, \beta > 0$ są parametrami rozkładu, a Γ to funkcja gamma Eulera:

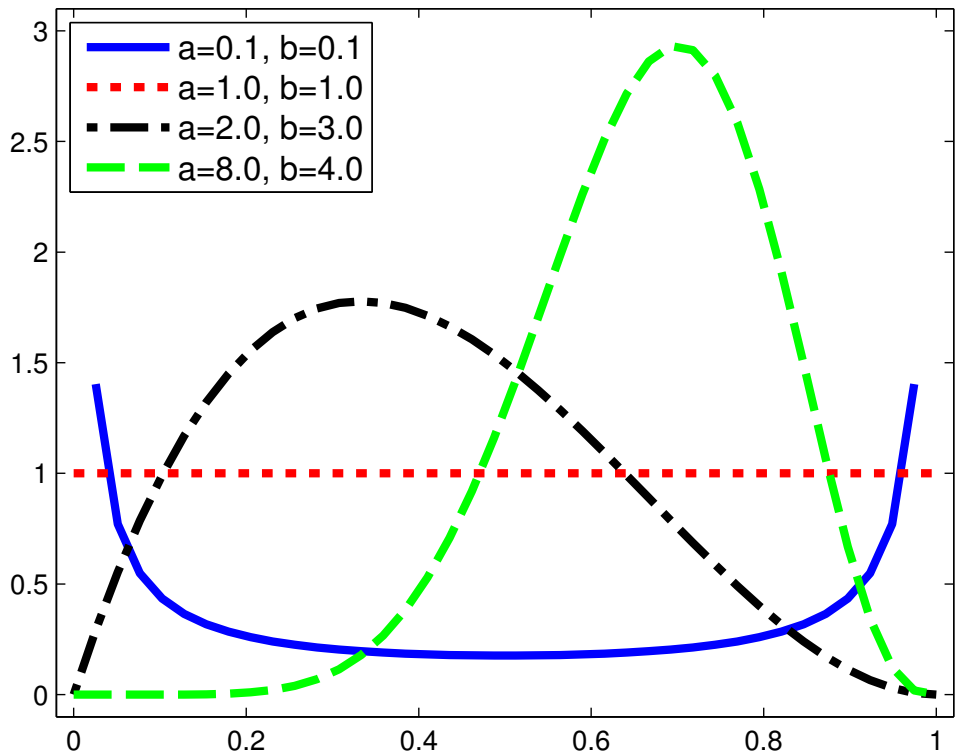
$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

Funkcja gamma rozszerza pojęcie silni na zbiór liczb zespolonych. W szczególności

$$\Gamma(z+1) = z\Gamma(z) \quad \text{oraz} \quad \Gamma(n+1) = n! \text{ dla } n \in \mathbb{N}.$$

Rozkład beta

W specjalnym przypadku, kiedy $\alpha = \beta = 1$, rozkład beta przyjmuje postać standardowego rozkładu jednostajnego. Gdy $\alpha, \beta < 1$, to rozkład jest bimodalny, dla $\alpha, \beta > 1$ – unimodalny. Gdy $\alpha = \beta$, to rozkład jest symetryczny względem $1/2$.



Przykładowy problem: czy student zda egzamin?

Przeprowadzamy egzamin z pewnego ciekawego przedmiotu wśród studentów TCS. W tej sesji przystąpiło do niego na razie 38 studentów i 30 spośród z nich zdało. Za chwilę wejdzie trzydziesty dziewiąty student. Jakie są jego szanse?

Przykładowy problem: czy student zda egzamin?

Przeprowadzamy egzamin z pewnego ciekawego przedmiotu wśród studentów TCS. W tej sesji przystąpiło do niego na razie 38 studentów i 30 spośród z nich zdało. Za chwilę wejdzie trzydziesty dziewiąty student. Jakie są jego szanse?

1. Na podstawie danych stwierdzamy prosto: $\theta = 30/38 \approx 0.79$.

Przykładowy problem: czy student zda egzamin?

Przeprowadzamy egzamin z pewnego ciekawego przedmiotu wśród studentów TCS. W tej sesji przystąpiło do niego na razie 38 studentów i 30 spośród z nich zdało. Za chwilę wejdzie trzydziesty dziewiąty student. Jakie są jego szanse?

1. Na podstawie danych stwierdzamy prosto: $\theta = 30/38 \approx 0.79$.
2. Wiemy jednak, że w poprzednich latach współczynnik zdawalności wynosił ok. 50%. Na tej podstawie skłaniamy się do hipotezy, że wchodzący student zda z prawdopodobieństwem $\theta \in (0.5, 0.79)$.

Przykładowy problem: czy student zda egzamin?

Przeprowadzamy egzamin z pewnego ciekawego przedmiotu wśród studentów TCS. W tej sesji przystąpiło do niego na razie 38 studentów i 30 spośród z nich zdało. Za chwilę wejdzie trzydziesty dziewiąty student. Jakie są jego szanse?

1. Na podstawie danych stwierdzamy prosto: $\theta = 30/38 \approx 0.79$.
2. Wiemy jednak, że w poprzednich latach współczynnik zdawalności wynosił ok. 50%. Na tej podstawie skłaniamy się do hipotezy, że wchodzący student zda z prawdopodobieństwem $\theta \in (0.5, 0.79)$.
3. Nie lubimy orzekać niczego tak definitywnie. Prawdopodobieństwo zdania przez studenta egzaminu to tak naprawdę zmienna losowa uwzględniająca nasze wcześniejsze doświadczenia.

Metoda największej wiarygodności (MLE)

Mamy model $\{p(y|\theta) : \theta \in \Theta\}$ i próbkę $D = \{x_1, \ldots, x_M\}$.
Funkcja wiarygodności względem parametru θ dla próbki D zadana jest wzorem

$$L(\theta) = p(D|\theta).$$

Estymator największej wiarygodności $\widehat{\theta}_{MLE}$ parametru θ to argument maksymalizujący wartość funkcji wiarygodności $L(\theta)$:

$$\widehat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} L(\theta).$$

Równoważnie

$$\widehat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \ell(\theta),$$

gdzie $\ell(\theta) := \log L(\theta)$.

Estymacja MLE dla przykładu ze studentami

Funkcja wiarygodności ma postać

$$L(\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Chcemy zmaksymalizować wartość funkcji

$$\ell(\theta) = \log \binom{n}{k} + k \log \theta + (n-k) \log(1-\theta),$$

skąd

$$\widehat{\theta}_{MLE} = \frac{k}{n}.$$

Gdy $n = 38$ i $k = 30$, to $\widehat{\theta}_{MLE} \approx 0.79$.

Estymacja MLE dla przykładu ze studentami

Funkcja wiarygodności ma postać

$$L(\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Chcemy zmaksymalizować wartość funkcji

$$\ell(\theta) = \log \binom{n}{k} + k \log \theta + (n-k) \log(1-\theta),$$

skąd

$$\widehat{\theta}_{MLE} = \frac{k}{n}.$$

Gdy $n = 38$ i $k = 30$, to $\widehat{\theta}_{MLE} \approx 0.79$.
Jednak gdy $n = 2$ i $k = 0$, to $\widehat{\theta}_{MLE} = 0$. A my przecież mamy większą wiarę w studentów.

Metoda estymacji maksymalnego *a posteriori* (MAP)

Mamy model $\{p(y|\theta) : \theta \in \Theta\}$ i próbkę $D = \{x_1, \ldots, x_M\}$. Dodatkowo mamy pewne przekonania dotyczące parametru θ, czyli prawdopodobieństwo *a priori* $p(\theta)$.
Chcemy wyznaczyć prawdopodobieństwo *a posteriori* $p(\theta|D)$ uwzględniające nasze wcześniejsze przekonania oraz uzyskane dane.
Chcemy zatem znaleźć taki estymator $\widehat{\theta}_{MAP}$ parametru θ, który maksymalizuje wartość funkcji $p(\theta|D)$:

$$\widehat{\theta}_{MAP} := \arg\max_{\theta \in \Theta} p(\theta|D).$$

Korzystając z twierdzenia Bayesa otrzymujemy

$$\widehat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} \frac{p(D|\theta)p(\theta)}{p(D)} = \arg\max_{\theta \in \Theta} p(D|\theta)p(\theta).$$

Estymacja MAP dla przykładu ze studentami

Chcemy wyznaczyć estymator $\widehat{\theta}_{MAP}$ parametru θ taki, że

$$\widehat{\theta}_{MAP} = \arg\max_{\theta \in (0,1)} p(n, k|\theta) p(\theta).$$

Wiemy, że

$$p(n, k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Za $p(\theta)$ przyjmujemy rozkład sprzężony z rokładem Bernoullego. Ponadto uwzględniając wcześniejszą zdawalność (ok. 50%), niech $p(\theta) = \text{Beta}(\theta|\alpha, \beta)$ dla pewnych $\alpha, \beta > 1$ takich, że $\alpha = \beta$. Otrzymujemy

$$\widehat{\theta}_{MAP} = \frac{k + \alpha - 1}{n + \alpha + \beta - 2}.$$

Gdy $n = 38$ i $k = 30$ oraz $\alpha = \beta = 10$, to $\widehat{\theta}_{MAP} \approx 0.696$.
Gdy $n = 2$ i $k = 0$ oraz $\alpha = \beta = 10$, to $\widehat{\theta}_{MAP} = 0.45$.

Modelowanie bayesowskie

Przykład Savage'a (1961) ilustrujący znaczenie prawdopodobieństwa subiektywnego

1. Ekspert z dziedziny muzyki twierdzi, że jest zdolny odróżnić muzykę Haydna od Mozarta na podstawie dowolnej strony z zapisem nutowym tych kompozytorów. W dziesięciu próbach wykonuje to zadanie poprawnie za każdym razem.
2. Kobieta, która lubi dodawać mleko do herbaty, uważa że jest w stanie rozpoznać, czy do kubka wlano najpierw herbatę czy mleko. W dziesięciu próbach, rozpoznaje to prawidłowo w każdym przypadku.
3. Twój nietrzeźwy znajomy stwierdza, że jest w stanie przewidzieć wynik rzutu monetą. W dziesięciu próbach przeprowadzonych w celu sprawdzenia jego słów, właściwie przewiduje wszystkich dziesięć rzutów.

Wnioskowanie bayesowskie

Wnioskowanie bayesowskie polega na wyznaczeniu rozkładu prawdopodobieństwa a posteriori. W odróżnieniu od estymatorów MLE i MAP wyznaczamy nie pojedynczą wartość, a rozkład gęstości (dla ciągłej zmiennej θ) lub prawdopodobieństwo (dla dyskretnej zmiennej θ):

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

Wnioskowanie bayesowskie dla przykładu ze studentami

Zakładamy a priori $p(\theta) = \text{Beta}(\theta|\alpha, \beta)$. Po przeegzaminowaniu n studentów, spośród których k zdało egzamin, otrzymujemy

$$
\begin{aligned}
p(\theta|n, k) &= \frac{p(n, k|\theta)p(\theta)}{p(n, k)} \\
&= \frac{\text{Bern}(n, k|\theta)\,\text{Beta}(\theta|\alpha, \beta)}{\text{Bern}(n, k)} \\
&= \frac{\Gamma(n + \alpha + \beta)}{\Gamma(k + \alpha)\Gamma(n - k + \beta)}\theta^{k+\alpha-1}(1 - \theta)^{n-k+\beta-1} \\
&= \text{Beta}(\theta|k + \alpha, n - k + \beta).
\end{aligned}
$$

Wnioskowanie bayesowskie dla przykładu ze studentami

Zakładamy a priori $p(\theta) = \text{Beta}(\theta|\alpha, \beta)$. Po przeegzaminowaniu n studentów, spośród których k zdało egzamin, otrzymujemy

$$\begin{aligned} p(\theta|n,k) &= \frac{p(n,k|\theta)p(\theta)}{p(n,k)} \\ &= \frac{\text{Bern}(n,k|\theta)\,\text{Beta}(\theta|\alpha,\beta)}{\text{Bern}(n,k)} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(k+\alpha)\Gamma(n-k+\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \\ &= \text{Beta}(\theta|k+\alpha, n-k+\beta). \end{aligned}$$

Dla $\alpha = \beta = 10$, $n = 38$ i $k = 30$

$$p(\theta|38,30) = \text{Beta}(\theta|48,18).$$