Wstęp do uczenia maszynowego

Uczenie statystyczne

Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski

18 marca 2024

UNIWERSYTET
WARSZAWSKI

Czym jest statystyczne uczenie?

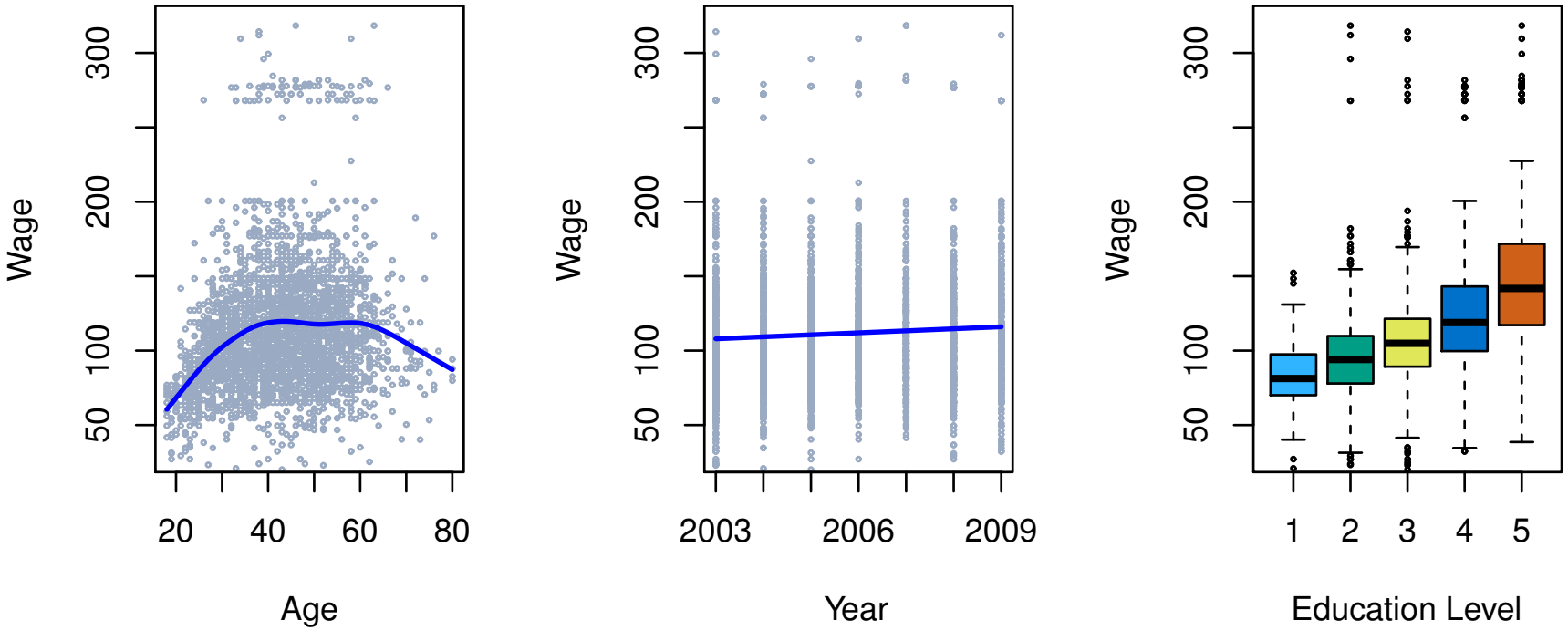
Wstęp do metod
statystycznego uczenia

- Machine learning
- Data science

Dwa rodzaje zagadnień statystycznego uczenia

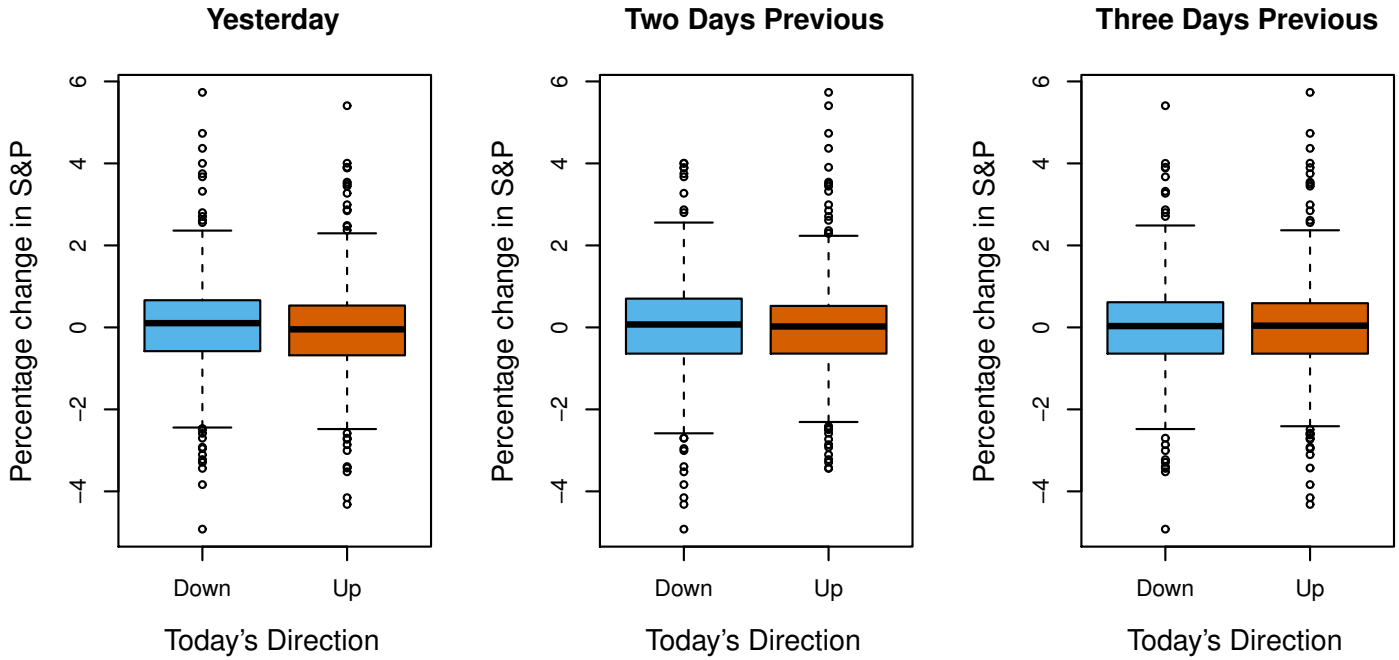
- **Pod nadzorem: (Ang. supervised)**
  - Budowanie modelu statystycznego dla przewidywania wyniku badanego zjawiska z danych wejściowych
  - Model jest trenowany na przykładach: dla zmiennych wejściowych podane są ich wyniki
  - **Klasyfikacja**: przewidywanie klasy, kategorii, z ustalonego skończonego zbioru (np. rodzaj raka, marka samochodu), przyjmowanych przez zmienne **jakościowe**
  - **Regresja**: przewidywanie ciągłych wartości liczbowych przyjmowanych przez zmienne **ilościowe**
- **Bez nadzoru: (Ang. unsupervised)**
  - Analiza danych wejściowych bez wiedzy o wynikach badanego zjawiska (lub bez ich uwzględniania)
  - **Klastrowanie** (analiza skupień)

Jakie zjawiska bywają badane? Zarobki mężczyzn w USA



- cel: przewidzieć zarobki w zależności od wieku, roku, edukacji
- wyniki ciągłe (lub ilościowe) – problem regresji
- przykładowy model: regresja liniowa
- są widoczne nieliniowe zależności

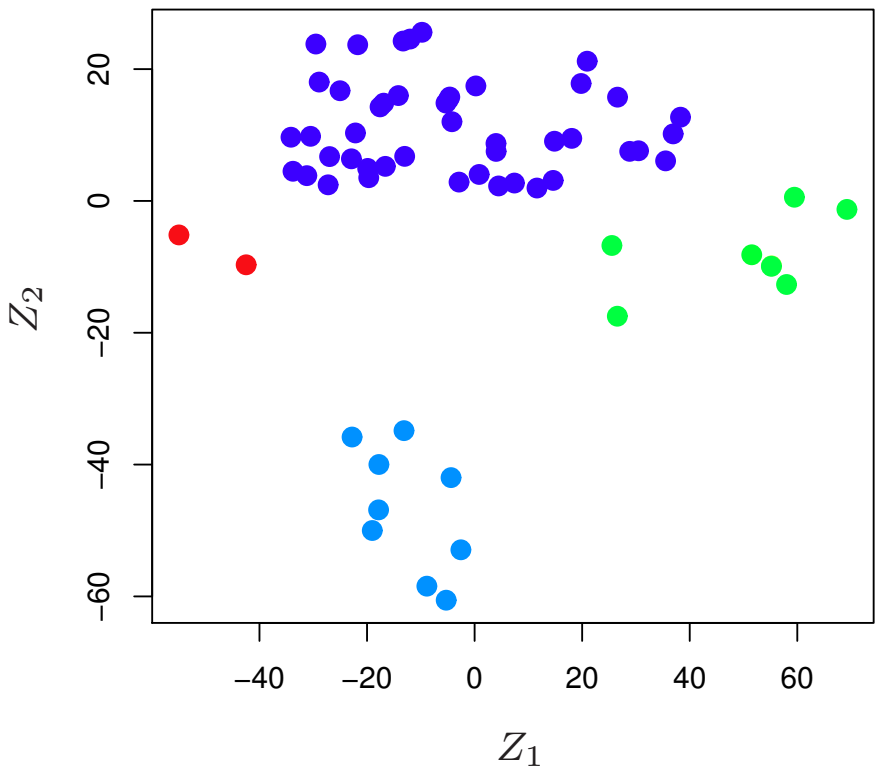
Jakie zjawiska bywają badane? Papiery wartościowe

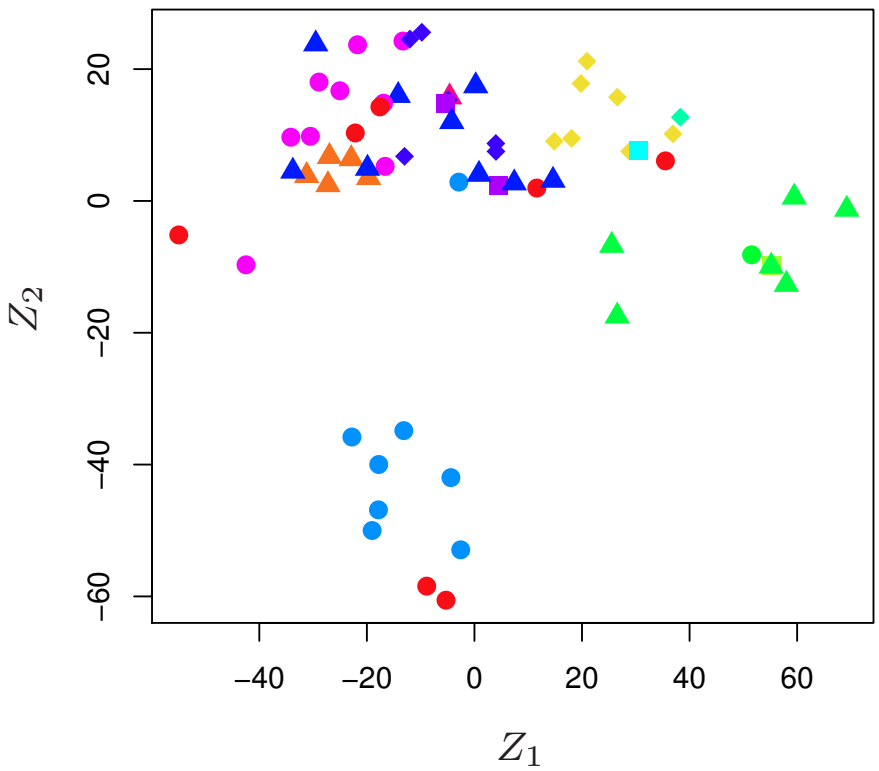


- cel: przewidzieć kierunek indeksu giełdowego w oparciu o procentowe zmiany indeksu z: 1 dnia poprzedniego, 2 poprzednich dni, bądź 3 poprzednich dni
- wyniki jakościowe ('Up', 'Down') – problem klasyfikacji
- wykresy nie wskazują na zależność kierunku indeksu od zmian z poprzednich dni

Jakie zjawiska bywają badane? Ekspresja genów

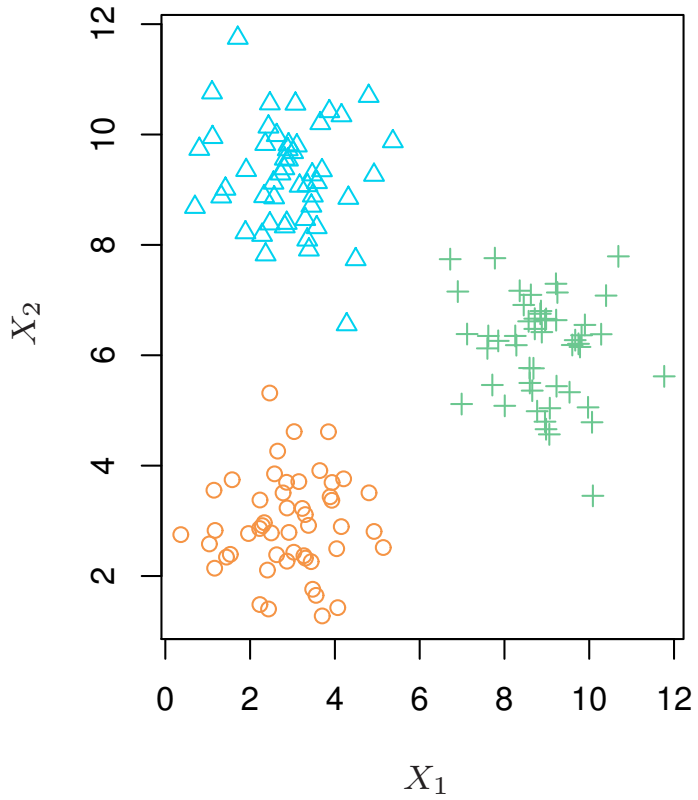
- 64 rakowe linie komórkowe. Ekspresja 6830 genów.



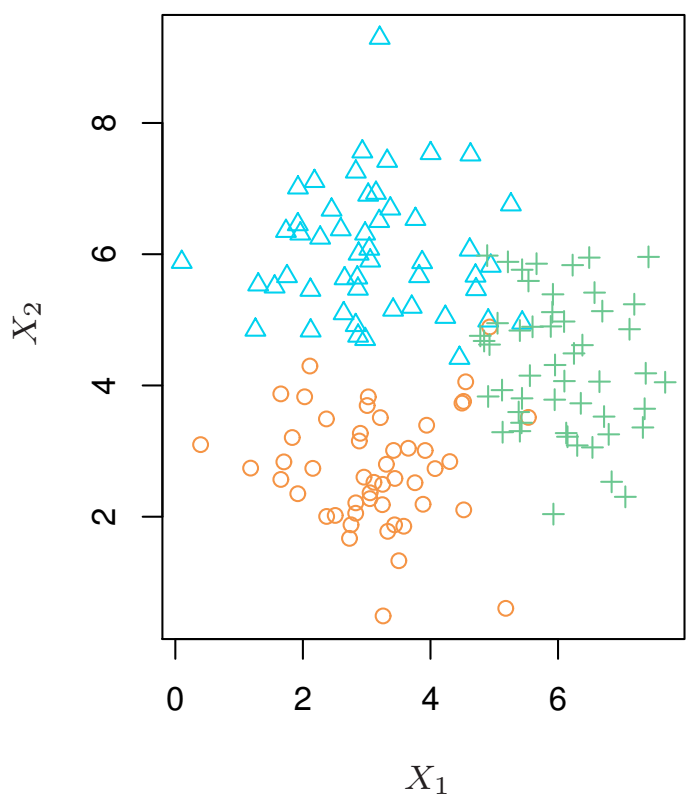


- Dwie pierwsze główne składowe; problem klasteryzacji (uczenie bez nadzoru)
- Naniesiona informacja o 14 typach raka

Przykład uczenia bez nadzoru: analiza skupień (klasteryzacja). 150 obserwacji i dwa predyktory



Łatwe zadanie klasteryzacji.



Trudne zadanie klasteryzacji

Problem uczenia statystycznego pod nadzorem

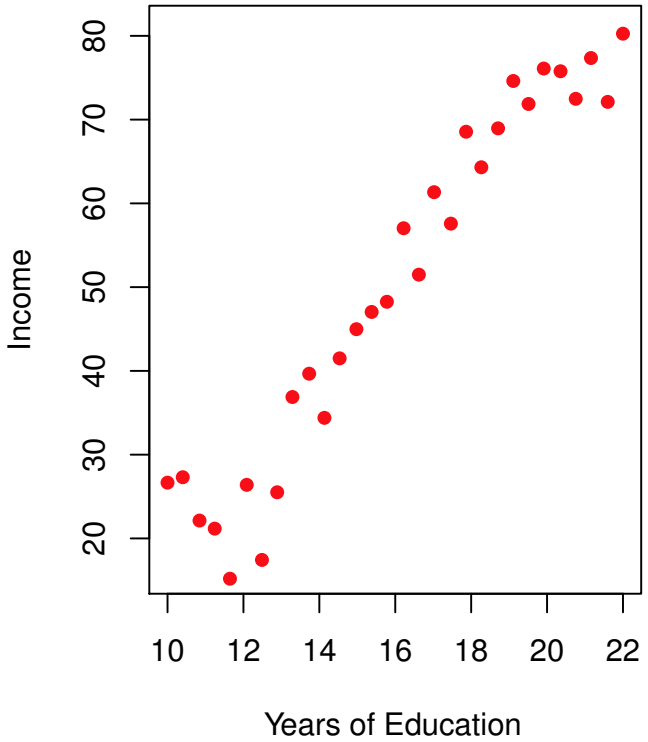
- Mamy *p zmiennych objaśniających* (predyktorów, cech) $X_1, \ldots, X_p$ (na przykład, dla danych o zarobkach: wiek, rok, edukacja)
- *Zmienna objaśniana* Y (zmienna odpowiedzi, prognozowana) reprezentująca wartość dla obserwowanej próbki
- Nieznana zależność wejście–wyjście:

$$Y = f(X_1, \ldots, X_p) + \varepsilon$$

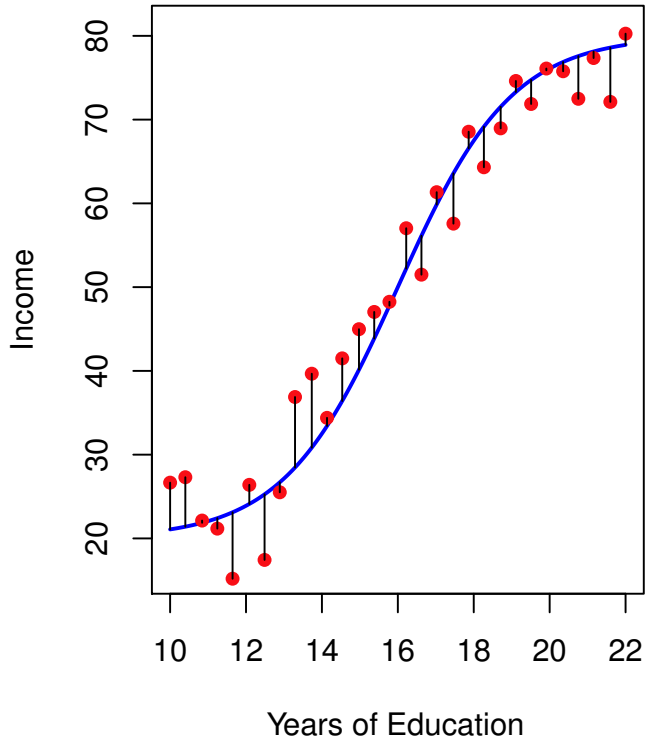
gdzie

- f jest nieznaną funkcją, którą chcemy estymować,
- ε jest błędem losowym o wartości oczekiwanej 0.

Przykład: zależność wysokości zarobków od lat edukacji

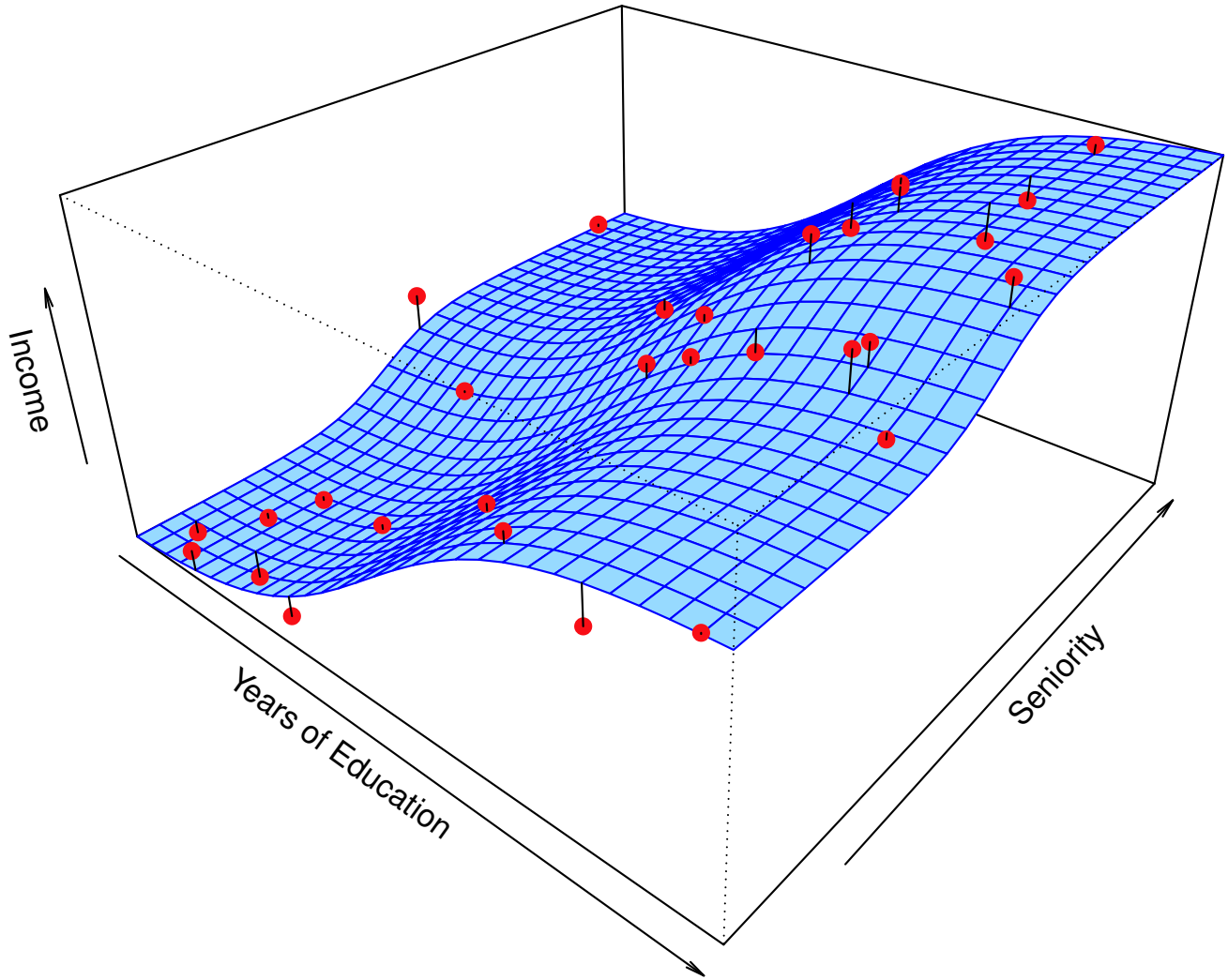


- Obserwowane wartości dla 30 osób
- Widoczna jakaś zależność, ale nie wiemy jaka ona jest - należy wyestymować f



- Prawdziwa funkcja zależności (z niej generowano dane)
- Czarne kreski – błąd losowy. niektóre błędy dodatnie, niektóre ujemne

Przykład: zależność zarobków od dwóch predyktorów



Problem uczenia statystycznego pod nadzorem: zadania

Estymujemy f, otrzymując funkcję $\hat{f}$.

1. **Predykcja**: z danych predyktorów X chcemy dobrze przewidzieć zmienną objaśnianą Y
2. **Wnioskowanie**: chcemy zrozumieć charakter zależności zmiennej objaśnianej Y od predyktorów X

Predykcja

- Predykcja zadana jest przez $\hat{Y} = \hat{f}(X) = \hat{f}(X_1, \ldots, X_p)$
- Mniej interesuje nas samo $\hat{f}$
- Bardziej interesuje nas $\hat{Y}$ (chcemy by było możliwie bliskie Y)
- Model $\hat{f}$ musi być *dokładny* a nie musi być *interpretowalny*
- Przykład: na podstawie badania krwi dobrze przewidzieć odpowiedź na lek

Predykcja: błąd redukowalny a błąd nieredukowalny

- Estymujemy f, otrzymując funkcję $\hat{f}$
- Wówczas wynikowa predykcja to $\hat{Y} = \hat{f}(X_1, \ldots, X_p)$
- Nawet jeśli $\hat{f}$ idealnie przybliża f, predykcja $\hat{Y}$ nadal będzie mieć błąd

Predykcja: błąd redukowalny a błąd nieredukowalny

- Estymujemy f, otrzymując funkcję $\hat{f}$
- Wówczas wynikowa predykcja to $\hat{Y} = \hat{f}(X_1, \ldots, X_p)$
- Nawet jeśli $\hat{f}$ idealnie przybliża f, predykcja $\hat{Y}$ nadal będzie mieć błąd
- Przyjmy, że $f, \hat{f}$ oraz $X = (X_1, \ldots, X_p)$ są stałe
- Wtedy oczekiwana wartość błędu predykcji:

$$\begin{aligned}
\mathbb{E}[(Y - \hat{Y})^2] &= \mathbb{E}[(f(X) + \varepsilon - \hat{f}(X))^2] \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2 + 2\epsilon(f(X) - \hat{f}(X)) + \epsilon^2] \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2] + 2\,\mathbb{E}[\epsilon]\,\mathbb{E}[f(X) - \hat{f}(X)] + \mathbb{E}[\epsilon^2] \\
&= \left(f(X) - \hat{f}(X)\right)^2 + Var(\varepsilon)
\end{aligned}$$

Pierwszy składnik to **błąd redukowalny**, a drugi to **błąd nieredukowalny**.

Wnioskowanie

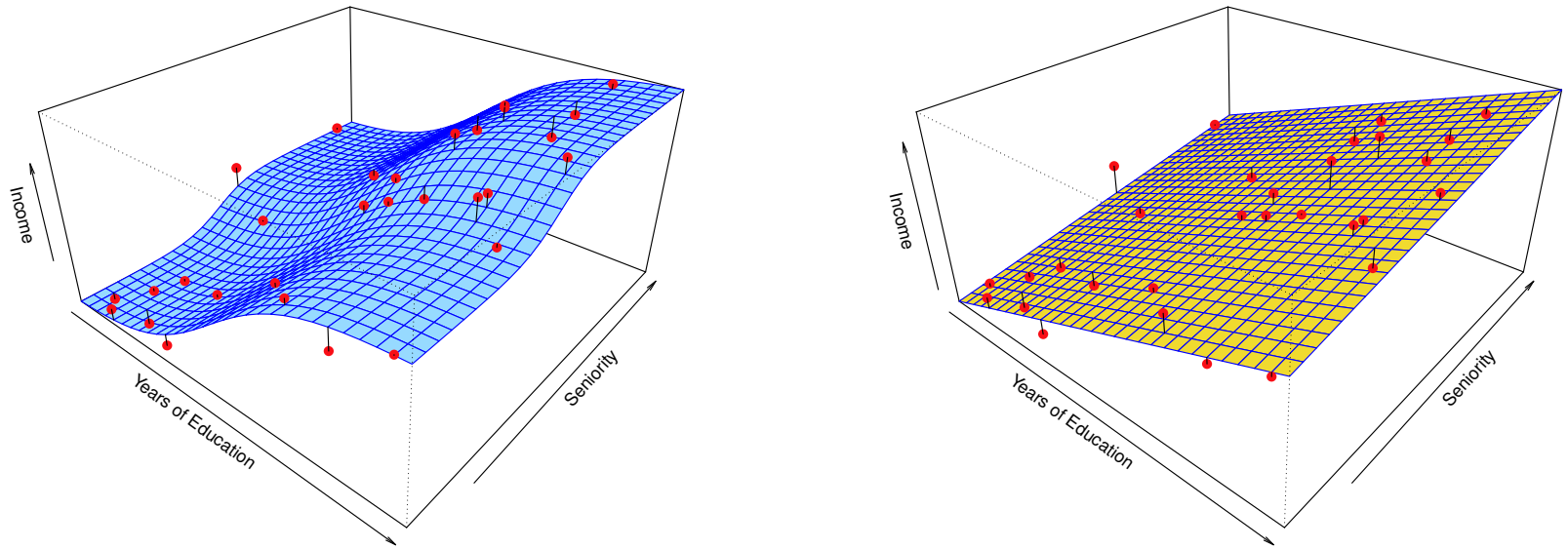
- Bardziej interesuje nas samo $\hat{f}$
- Model $\hat{f}$ nie musi być bardzo *dokładny* ale dobrze, gdy jest *interpretowalny*
- Które zmienne objaśniające silniej wpływają na Y?
- Jaki jest związek pomiędzy zmienną Y a każdą poszczególną zmienną objaśniającą?
- Czy taki związek można dobrze opisać używając równania liniowego?
- Przykład: dowiedzieć się, jak dwukrotny wzrost ciśnienia wpływa na odpowiedź na lek

Parametryczne metody estymacji f

Przykład: model liniowy

$$Y \approx \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

parametry $\beta_0, \ldots, \beta_p$ dobieramy, np. *metodą najmniejszych kwadratów*.



Przybliżenie wysokości zarobków funkcją liniową.

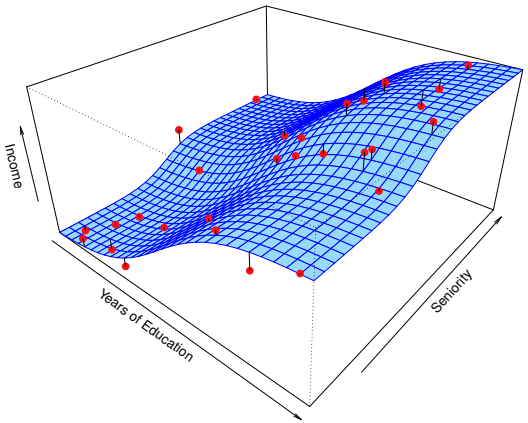
Elastyczne modele a przeuczenie (overfitting)

Elastyczność osiąga się poprzez branie pod uwagę większą klasę funkcji używanych do estymacji.

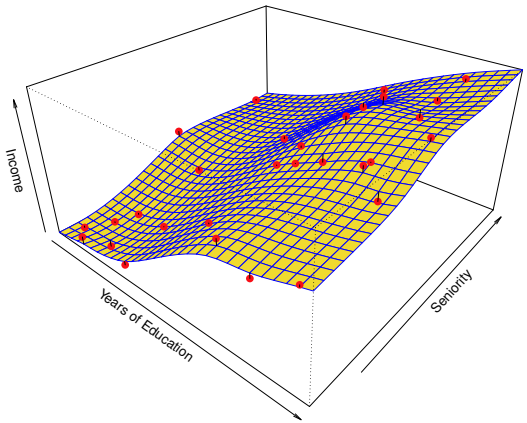
- Zbyt duża liczba parametrów może prowadzić do **przeuczenia**
- Przeuczenie polega na zbyt dokładnym dopasowaniu modelu do obserwowanych wartości zmiennej Y w danych treningowych
- Prowadzi do "modelowania"błędów losowych, czyli szumu.
- Modele mało elastyczne są mniej podatne na over-fitting niż złożone modele o dużej liczbie parametrów

Nieparametryczne metody estymacji f

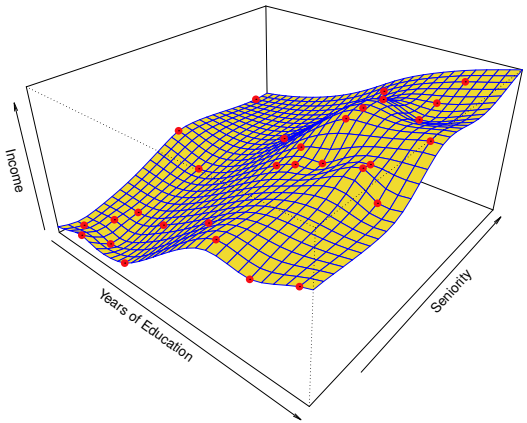
- Wybieramy bardzo szeroką klasę funkcji do estymacji f.
- Np. dopasowujemy funkcje sklejane (Ang. splines) złożoność modelu kontrolujemy "gładkością" dopasowanej funkcji, czyli całką z kwadratu drugiej pochodnej
- Bardziej gładkie funkcje sklejane są "prostszymi" modelami



Prawdziwa funkcja

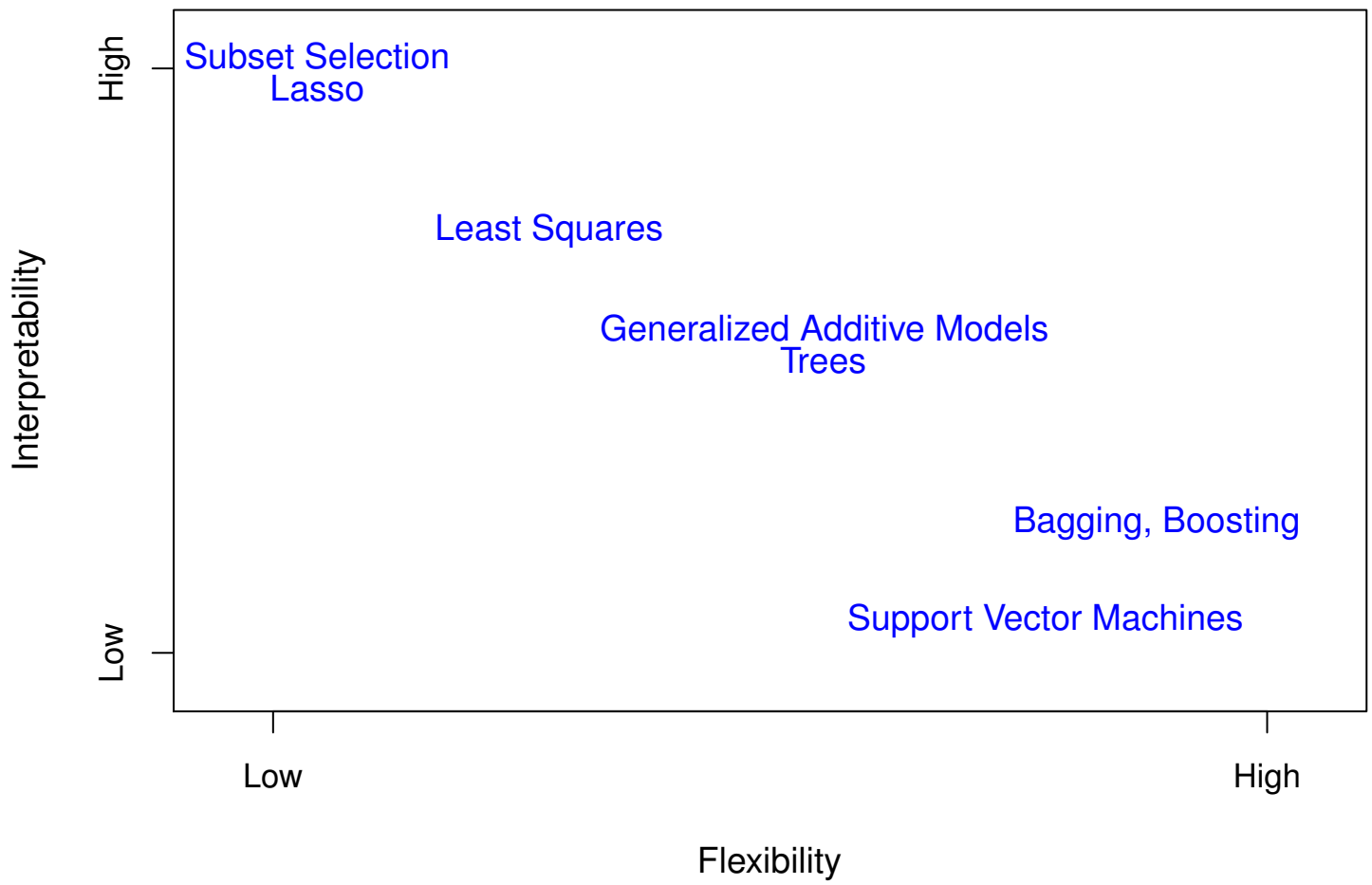


Bardzo gładkie dopasowanie



Mniej gładkie dopasowanie - overfitting.

Kompromis pomiędzy elastycznością a interpretowalnością



Ocena dokładości modelu dla regresji

Podział danych

- **Dane treningowe**
  - $(x_{\text{Train}}, y_{\text{Train}})$ użyte do estymacji $\hat{f}$
- **Dane testowe**
  - $(x_{\text{Test}}, y_{\text{Test}})$ użyte do ewaluacji jak dobrze estymowaliśmy $\hat{f}$
- **Zastosowanie**
  - x dla których nie znamy y i chcemy je przewidzieć

Ocena dokładności modelu; przypadek regresji

- $x = \left(x^{(1)}, \ldots, x^{(n)}\right)$ obserwacje dla predyktorów, gdzie $x^{(j)} = \left(x_1^{(j)}, \ldots, x_p^{(j)}\right)$
- $y = \left(y^{(1)}, \ldots, y^{(n)}\right)$ wartości zmiennej objaśnianej
- dane są postaci $(x, y) = (x^{(j)}, y^{(j)})_{j \in \{1,\ldots,n\}}$

Średni błąd kwadratowy (*mean squared error*) predykcji:

$$MSE = (1/n) \sum_{j=1}^{n} (y^{(j)} - \hat{f}(x^{(j)}))^2,$$

gdzie $\hat{f}(x^{(j)})$ jest predykcją otrzymaną z $\hat{f}$ dla j-tej obserwacji, a $y^{(j)}$ jest poprawną wartością zmiennej objaśnianej Y dla tej obserwacji.

MSE treningowy a MSE testowy

- **Dane treningowe**: $(x_{\text{Train}}, y_{\text{Train}}) = \left((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\right)$
- Na danych treningowych uczymy (trenujemy) model reprezentowany przez funkcję $\hat{f}$
- **Treningowy MSE** według poprzedniego wzoru: $MSE_{\text{Train}} = (1/n) \sum_{j=1}^{n} (y^{(j)} - \hat{f}(x^{(j)}))^2$

MSE treningowy a MSE testowy

- **Dane treningowe**: $(x_{\text{Train}}, y_{\text{Train}}) = \left((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\right)$
- Na danych treningowych uczymy (trenujemy) model reprezentowany przez funkcję $\hat{f}$
- **Treningowy MSE** według poprzedniego wzoru: $MSE_{\text{Train}} = (1/n)\sum_{j=1}^{n}(y^{(j)} - \hat{f}(x^{(j)}))^2$
- **Dane testowe**: $(x_{\text{Test}}, y_{\text{Test}}) = \left((x^{(n+1)}, y^{(n+1)}), \ldots, (x^{(n+m)}, y^{(n+m)})\right)$
- Danych testowych **nie używamy do uczenia modelu**
- Danych testowych używamy do oszacowania jak dobrze estymowaliśmy $\hat{f}(x)$
- **Testowy MSE**: $MSE_{\text{Test}} = (1/m)\sum_{j=n+1}^{n+m}(y^{(j)} - \hat{f}(x^{(j)}))^2$
- Model oceniamy przede wszystkim w oparciu o testowy MSE.

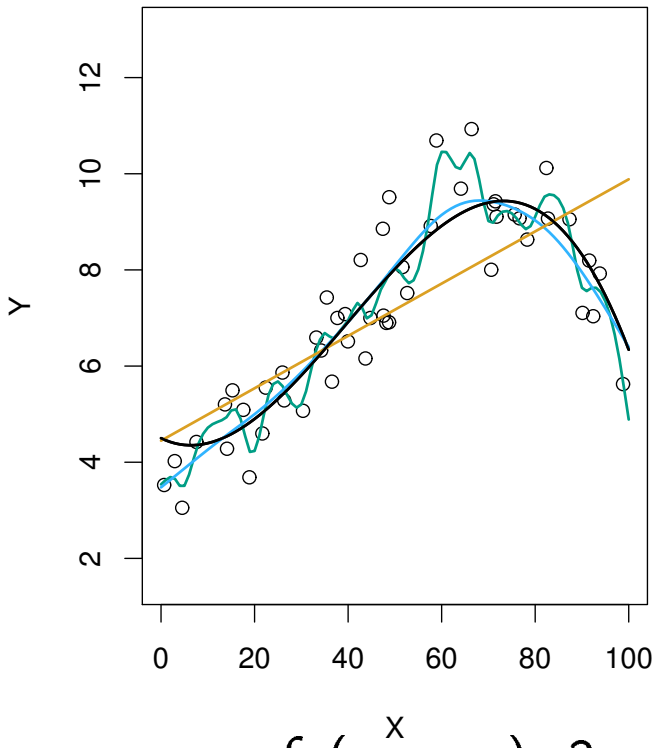
MSE treningowy a MSE testowy

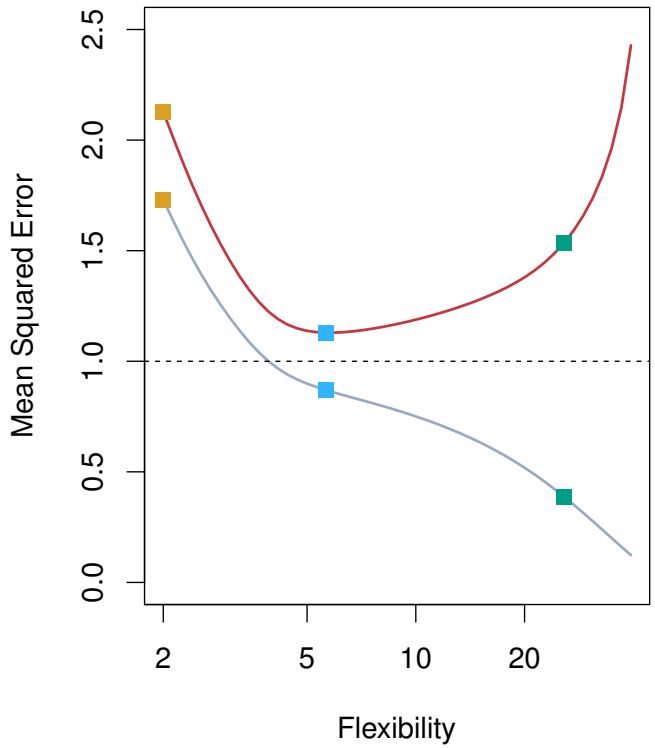
- **Dane treningowe**: $(x_{\text{Train}}, y_{\text{Train}}) = \left((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\right)$
- Na danych treningowych uczymy (trenujemy) model reprezentowany przez funkcję $\hat{f}$
- **Treningowy MSE** według poprzedniego wzoru: $MSE_{\text{Train}} = (1/n)\sum_{j=1}^{n}(y^{(j)} - \hat{f}(x^{(j)}))^2$
- **Dane testowe**: $(x_{\text{Test}}, y_{\text{Test}}) = \left((x^{(n+1)}, y^{(n+1)}), \ldots, (x^{(n+m)}, y^{(n+m)})\right)$
- Danych testowych **nie używamy do uczenia modelu**
- Danych testowych używamy do oszacowania jak dobrze estymowaliśmy $\hat{f}(x)$
- **Testowy MSE**: $MSE_{\text{Test}} = (1/m)\sum_{j=n+1}^{n+m}(y^{(j)} - \hat{f}(x^{(j)}))^2$
- Model oceniamy przede wszystkim w oparciu o testowy MSE.

Modele zbyt elastyczne będą miały duży testowy MSE. Nazywamy to przeuczeniem modelu (Ang. overfitting)

Elastyczność modelu a MSE

Nieliniowa zależność $X \to Y$.



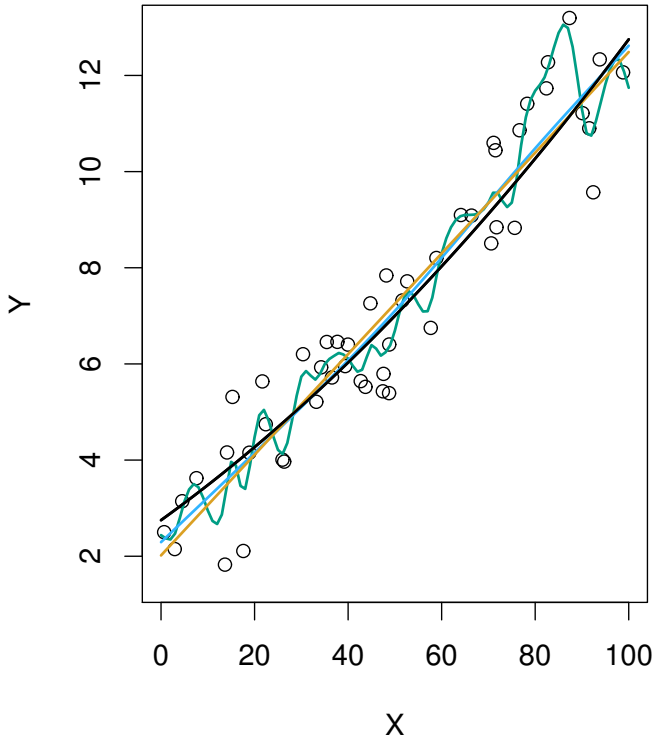


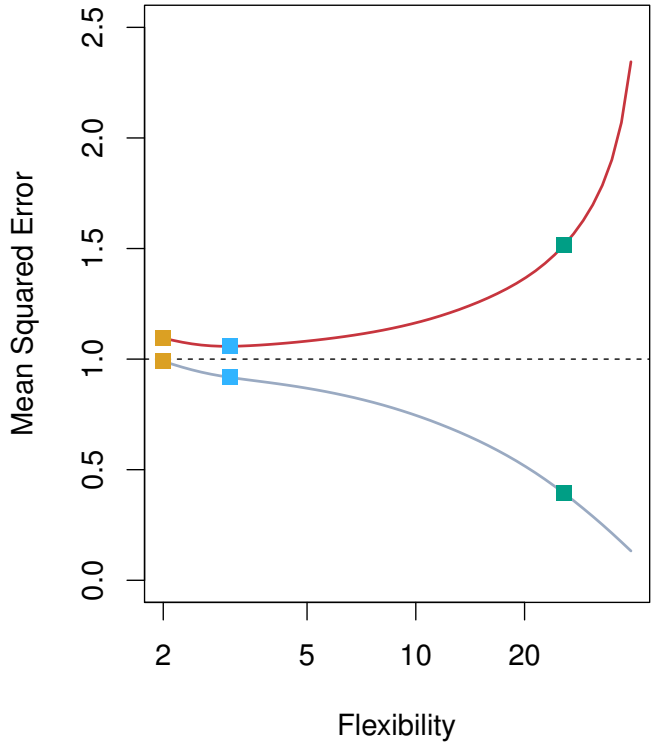
Dane generowane z f (czarny); 3 estymacje: liniowa regresja (pomarańczowy); splajny (duża gładkość – niebieski; mniejsza gładkość – zielony).

MSE treningowy (szary); MSE testowy (czerwony).

Elastyczność modelu a MSE

Prawie liniowa zależność $X \to Y$.



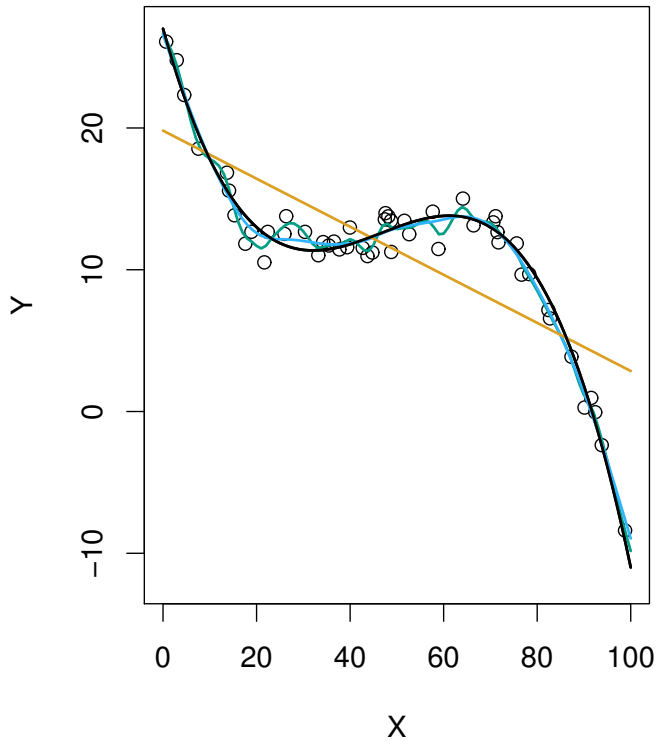


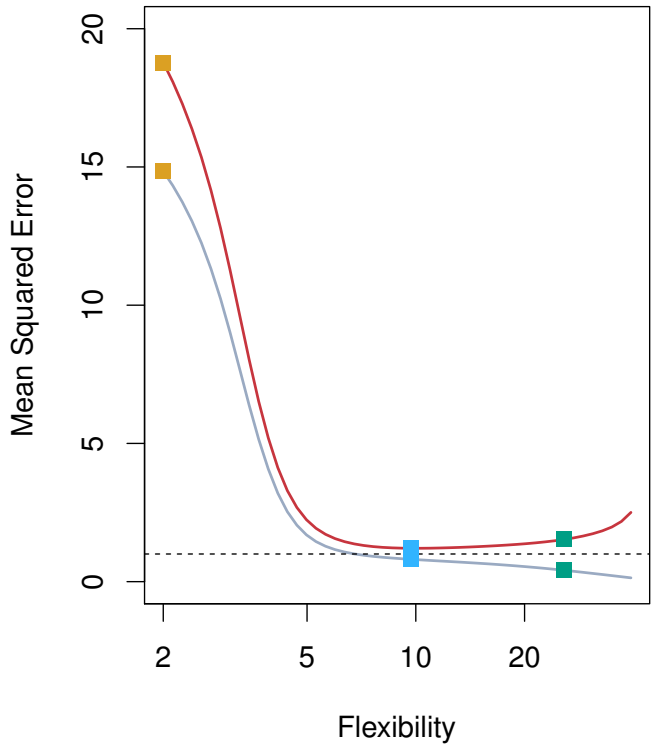
Dane generowane z f (czarny); 3 estymacje: liniowa regresja (pomarańczowy); splajny (duża gładkość – niebieski; mniejsza gładkość – zielony).

MSE treningowy (szary); MSE testowy (czerwony).

Elastyczność modelu a MSE

Mocno nieliniowa zależność $X \to Y$.





Dane generowane z f (czarny); 3 estymacje: liniowa regresja (pomarańczowy); splajny (duża gładkość – niebieski; mniejsza gładkość – zielony).

MSE treningowy (szary); MSE testowy (czerwony).

Przypomnienie: obciążenie i wariancja estymatora

- Niech $X_1, \ldots, X_n$ - próba losowa
- θ_n estymator
- Obciążenie estymatora: $b(\theta_n) = \mathbb{E}[\theta_n] - \theta$.
- Wariancja estymatora: $\mathrm{Var}(\theta_n)$

Dekompozycja obciążenie-wariancja dla błędu średniokwadratowego

- Średni błąd kwadratowy estymatora wyraża się wzorem

$$\mathbb{E}[(\theta_n - \theta)^2] = b(\theta_n)^2 + \mathrm{Var}(\theta_n)$$

- Co widać ze wzorów:

$$\begin{aligned} \mathbb{E}[(\theta_n - \theta)^2] &= \mathbb{E}[\theta_n^2] + \theta^2 - 2\theta\,\mathbb{E}[\theta_n] \\ b(\theta_n)^2 &= (\mathbb{E}[\theta_n] - \theta)^2 \\ &= \mathbb{E}[\theta_n]^2 + \theta^2 - 2\theta\,\mathbb{E}[\theta_n] \\ \mathrm{Var}(\theta_n) &= \mathbb{E}[\theta_n^2] - \mathbb{E}[\theta_n]^2 \end{aligned}$$

Kompromis pomiędzy wariancją a obciążeniem predyktora

- Rozważmy zadaną wartość obserwacji testowej x_0, o jakiejś prawdziwej wartości zmiennej $Y = y_0$
- Model $\hat{f}$ będzie różny dla różnych zbiorów treningowych - $\hat{f}$ zależy od próby i jest estymatorem
- Jak będzie wyglądać wartość oczekiwana błędu dla x_0 po tych różnych próbach?

$$\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [b(\hat{f}(x_0))]^2 + \mathrm{Var}(\varepsilon)$$

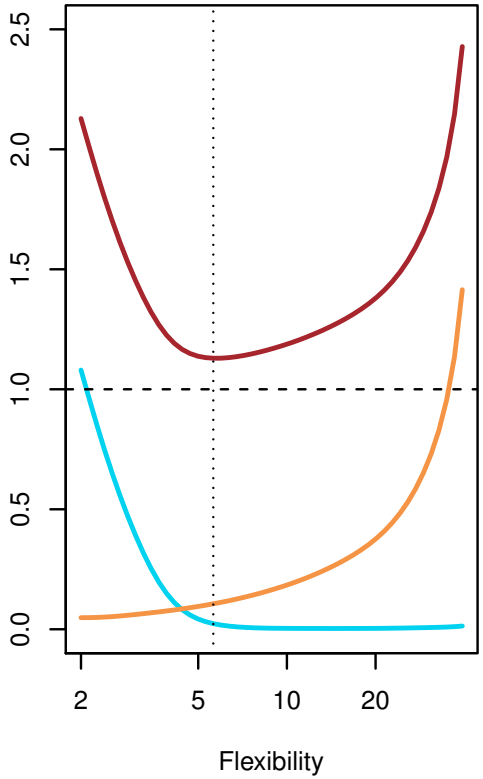
Kompromis pomiędzy wariancją a obciążeniem predyktora

- Rozważmy zadaną wartość obserwacji testowej x_0, o jakiejś prawdziwej wartości zmiennej $Y = y_0$
- Model $\hat{f}$ będzie różny dla różnych zbiorów treningowych - $\hat{f}$ zależy od próby i jest estymatorem
- Jak będzie wyglądać wartość oczekiwana błędu dla x_0 po tych różnych próbach?

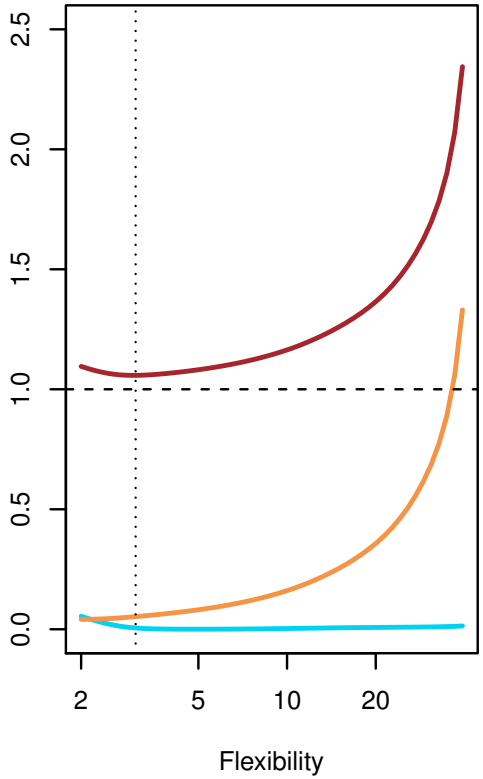
$$\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + [b(\hat{f}(x_0))]^2 + \mathrm{Var}(\varepsilon)$$

- Najmniejszy oczekiwany błąd testowy to nieredukowalny błąd $\mathrm{Var}(\varepsilon)$
- Chcemy jednocześnie minimalizować wariancję i obciążenie
- Elastyczne modele będą miały dużą wariancję i małe obciążenie
- Zbyt proste modele będą miały małą wariancję i duże obciążenie
- Szukamy złotego środka

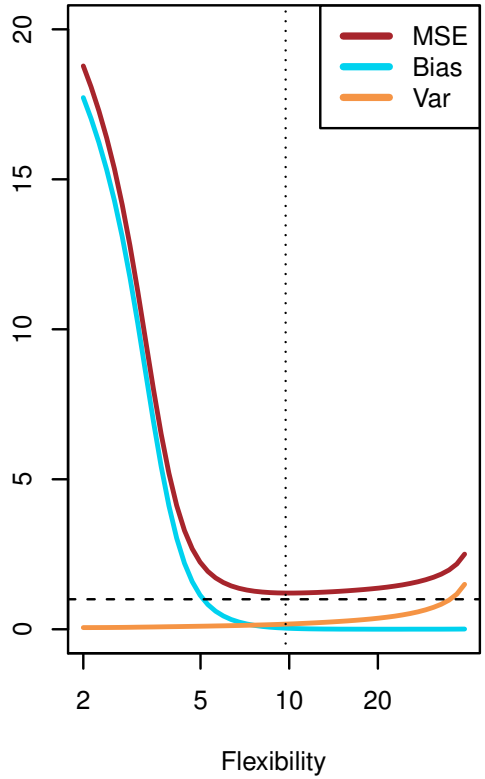
Związek pomiędzy obciążeniem^2, wariancją i testowym MSE



średnio nieliniowa zależność w danych



prawie liniowa zależność



mocno nieliniowa zależność.

Linia przerywana: wartość $Var(\varepsilon)$.

Ocena dokładności modelu dla klasyfikacji

Ocena dokładności modelu; przypadek klasyfikacji

Szukamy estymatora $\hat{f}$ dla danych treningowych $(x_{\text{Train}}, y_{\text{Train}}) = \left((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\right)$.

Średni błąd treningowy:

$$(1/n) \sum_{j=1}^{n} I(y^{(j)} \neq \hat{f}(x^{(j)})).$$

$I(x \neq y)$ to *zmienna indykatorowa* przyjmująca wartość 1, gdy zachodzi wypisana własność, oraz 0 w przeciwnym przypadku.

Średni błąd testowy:

$$(1/m) \sum_{j=n+1}^{n+m} I(y^{(j)} \neq \hat{f}(x^{(j)})).$$

Ocena modelu dla dwóch klas

$\hat{Y}/Y$	$Y=1$	$Y=0$	
$\hat{Y}=1$	TP	FP	TP+FP
$\hat{Y}=0$	FN	TN	FN+TN
	P	N	

- **Czułość** (ang. sensitivity, recall, true positive rate (TPR))

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

- **Swoistość** (ang. specificity, selectivity, true negative rate (TNR))

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

Ocena modelu dla dwóch klas

$\hat{Y}/Y$	$Y=1$	$Y=0$	
$\hat{Y}=1$	TP	FP	TP+FP
$\hat{Y}=0$	FN	TN	FN+TN
	P	N	

- **Precyzja** (ang. precision, positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = 1 - \mathrm{FDR}$$

- **False discovery rate (FDR)**

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}} = 1 - \mathrm{PPV}$$

- **Dokładność** (accuracy (ACC))

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

Klasyfikator bayesowski jako optymalny klasyfikator

X – zmienna losowa przedstawiająca wynik obserwacji testowej; Y – zmienna losowa przedstawiająca odpowiedź klasyfikatora.

Klasyfikator bayesowski (M klas): przypisujemy obserwowaną wartość x_0 do tej klasy j ($1 \leq j \leq M$), dla której prawdopodobieństwo warunkowe

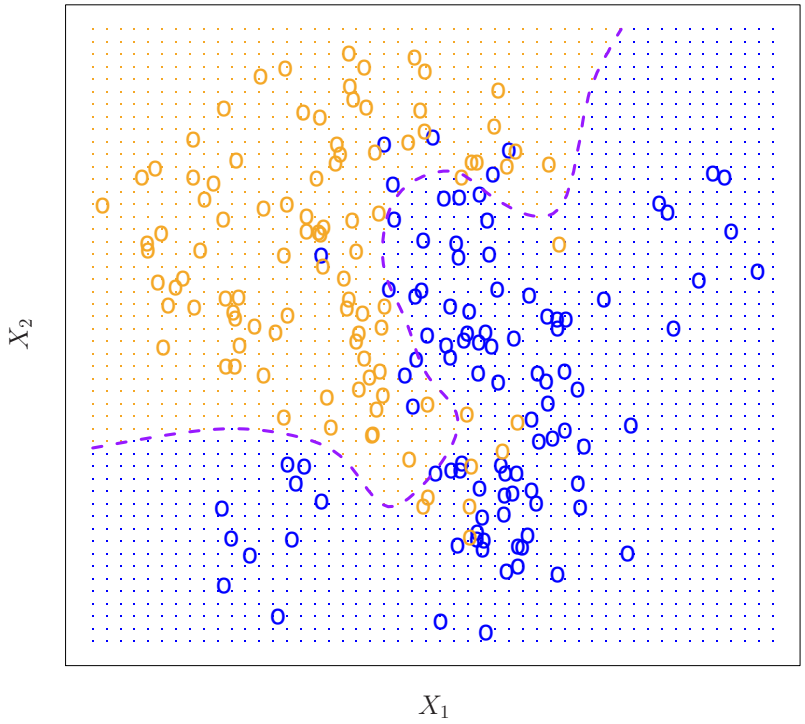
$$Pr(Y = j \mid X = x_0)$$

jest największe.

Fakt: Klasyfikator bayesowski minimalizuje błąd testowy.

Problem: zwykle przestrzeń probabilistyczna użyta powyżej jest nieznana.

Symulowane dane dla dwuwymiarowej przestrzeni cech



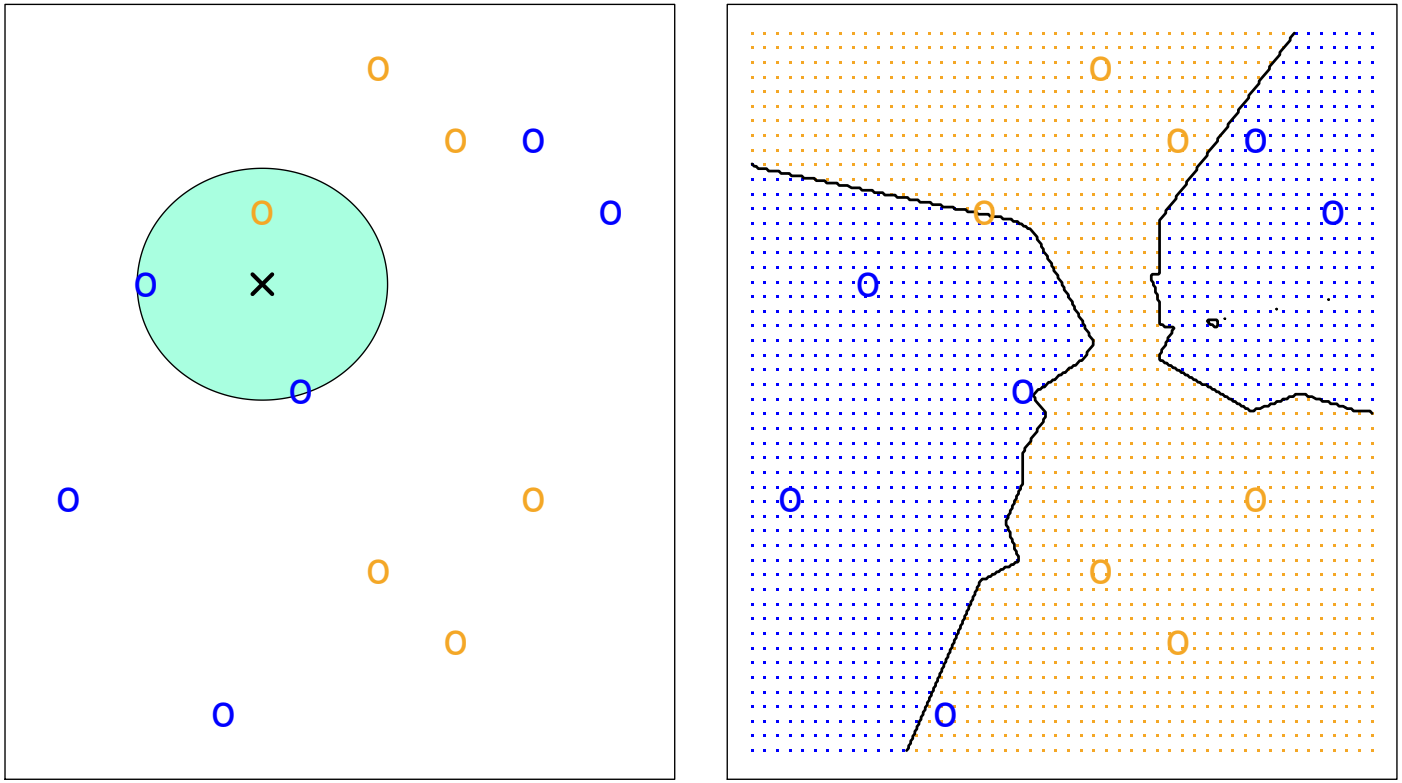
Dwie klasy - pomarańczowa i niebieska. $X = (X_1, X_2)$. Obszar pomarańczowych kropek- zbiór punktów X, dla których $Pr(Y = \text{'pomarańczowy'} \mid X) > 0.5$. Różowa przerywana linia - punkty X dla których $Pr(Y = \text{'pomarańczowy'} \mid X) = 0.5$ (**bayesowska granica decyzyjna**).

Metoda K najbliższych sąsiadów (KNN) jako przybliżenie bayesowskiego klasyfikatora

- Dana liczba naturalna K oraz obserwacja x_0.
- Klasyfikator KNN znajduje K punktów w danych treningowych, które są położone najbliżej x_0.
- Niech T będzie zbiorem tych punktów.
- Dla każdej klasy j, niech t_j będzie liczbą wszystkich punktów z T należących do klasy j.
- Wówczas przyjmujemy *aproksymację klasyfikatora bayesowskiego* przez

$$Pr(Y = j \mid X = x_0) = t_j/K.$$

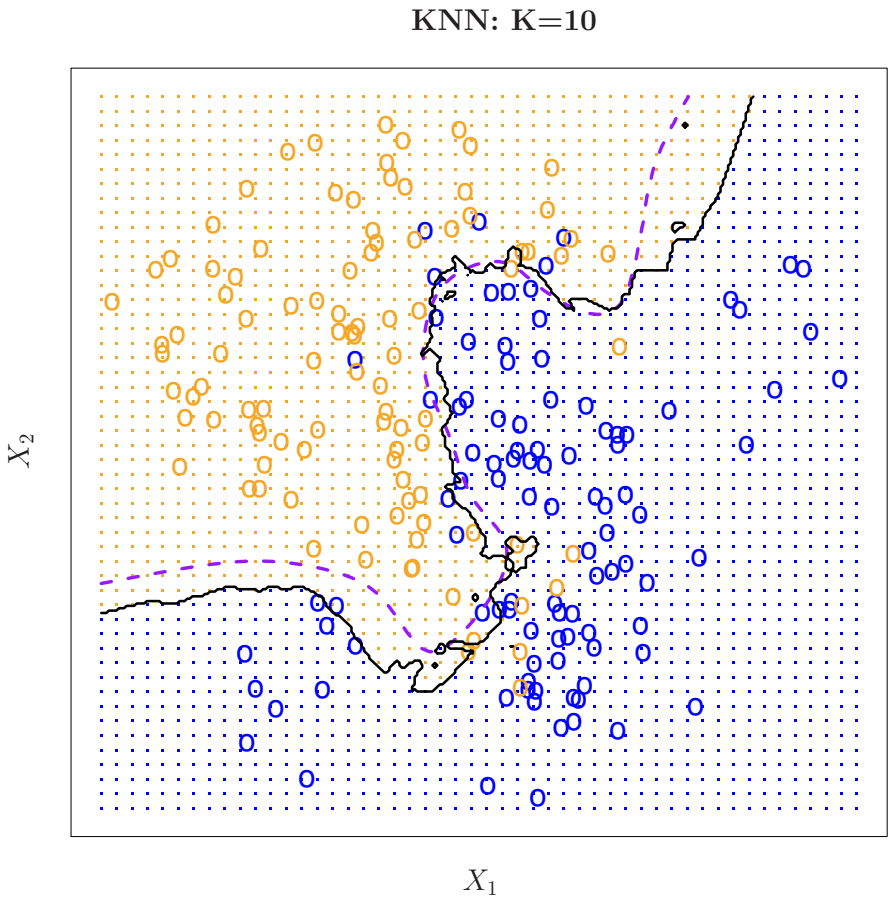
Ilustracja metody KNN dla $K = 3$



Lewy panel: $\times$ - obserwacja x_0, $K = 3$. Klasyfikator KNN przypisuje klasę niebieską.

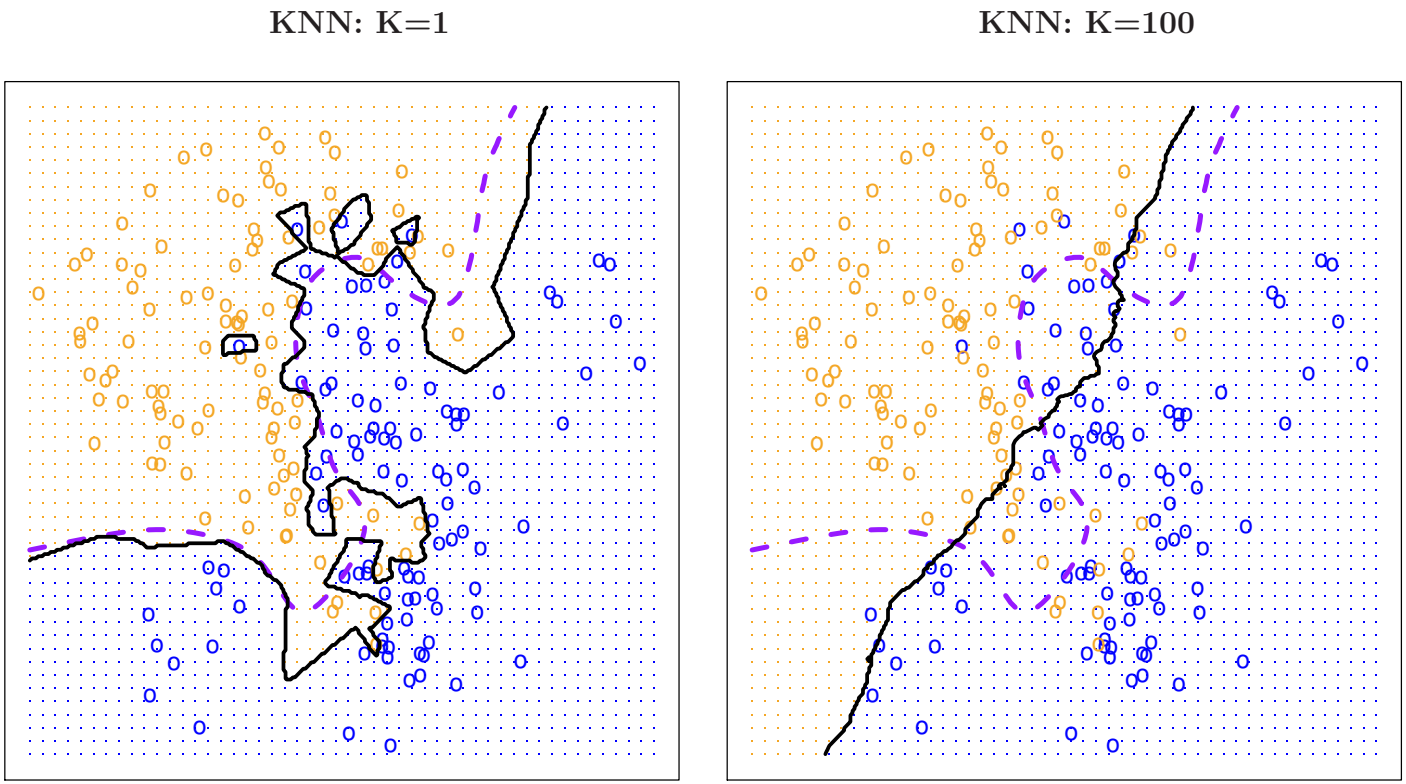
Prawy panel: czarna linia - granica decyzyjna metody KNN (dwie klasy, $K = 3$). Punkty niebieskie zostaną przypisane do klasy niebieskiej przez KNN.

Aproksymacja bayesowskiego klasyfikatora przez KNN przy $K = 10$



Różowa przerywana linia - bayesowska granica decyzyjna. Czarna linia - granica decyzyjna KNN.

Parametr K ma dramatyczny wpływ na jakość klasyfikacji



Małe K: model bardzo elastyczny – przeuczenie (małe obciążenie, duża wariancja). **Duże** K: model staje się mało elastyczny (duże obciążenie, mała wariancja) i zbliża się do liniowej aproksymacji.

Podsumowanie

- Uczenie z nadzorem i bez nadzoru
- Predykcja, wnioskowanie
- Błąd redukowalny i nieredukowalny
- Średni błąd kwadratowy dla regresji
- Kompromis między wariancją a obciążeniem
- Średni błąd kwadratowy dla klasyfikacji
- Czułość, swoistość, precyzja, dokładność
- Klasyfikator bayesowski
- Klasyfikator KNN