Analiza danych

prof. UAM dr hab. Tomasz Górecki

tomasz.gorecki@amu.edu.pl
http://drizzt.home.amu.edu.pl

Zakład Statystyki Matematycznej i Analizy Danych
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Regresja, a współliniowość zmiennych

Zgodnie z założeniami Klasycznego Modelu Regresji Liniowej zmienne objaśniające w modelu powinny być skorelowane ze zmienną objaśnianą i nieskorelowane między sobą. Ale rzeczywiste dane zawsze są w pewnym stopniu skorelowane, więc regresory są współliniowe (*ang. multicollinearity*). Weźmy pod uwagę model regresji wielokrotnej dla jedynie dwóch zmiennych objaśniających:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \;\; i = 1, 2, \ldots, n,$$

gdzie $\mathbb{E}(\varepsilon) = 0$ i $\mathbb{V}\text{ar}(\varepsilon) = \sigma^2$. Wariancję estymatorów modelu można zapisać jako:

$$\mathbb{V}\text{ar}(b_j) = \frac{\sigma^2}{(1 - r_{12}^2) \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2},$$

gdzie r_{12} jest korelacją pomiędzy zmiennymi x_1 i x_2.

Regresja, a współliniowość zmiennych

Jeśli zmienne objaśniające w modelu są silnie skorelowane to wariancja estymatora dąży do nieskończoności. W takim przypadku występują następujące problemy:

1. Niewielkie zmiany w zbiorze danych powodują duże zmiany w otrzymywanych estymatorach.
2. Współczynniki równania regresji mają duże błędy standardowe, oraz mogą być nieistotne statystycznie, nawet gdy łącznie są istotne, a współczynnik R^2 modelu jest wysoki.
3. Współczynniki równania regresji mają „złe", czyli niezgodne z teorią znaki, albo są zbyt małe lub zbyt duże.

Regresja, a współliniowość zmiennych

Powyższe równanie można uogólnić na przypadek wielu zmiennych objaśniających otrzymując:

$$\mathbb{V}\text{ar}(b_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \frac{1}{1 - R_j^2},$$

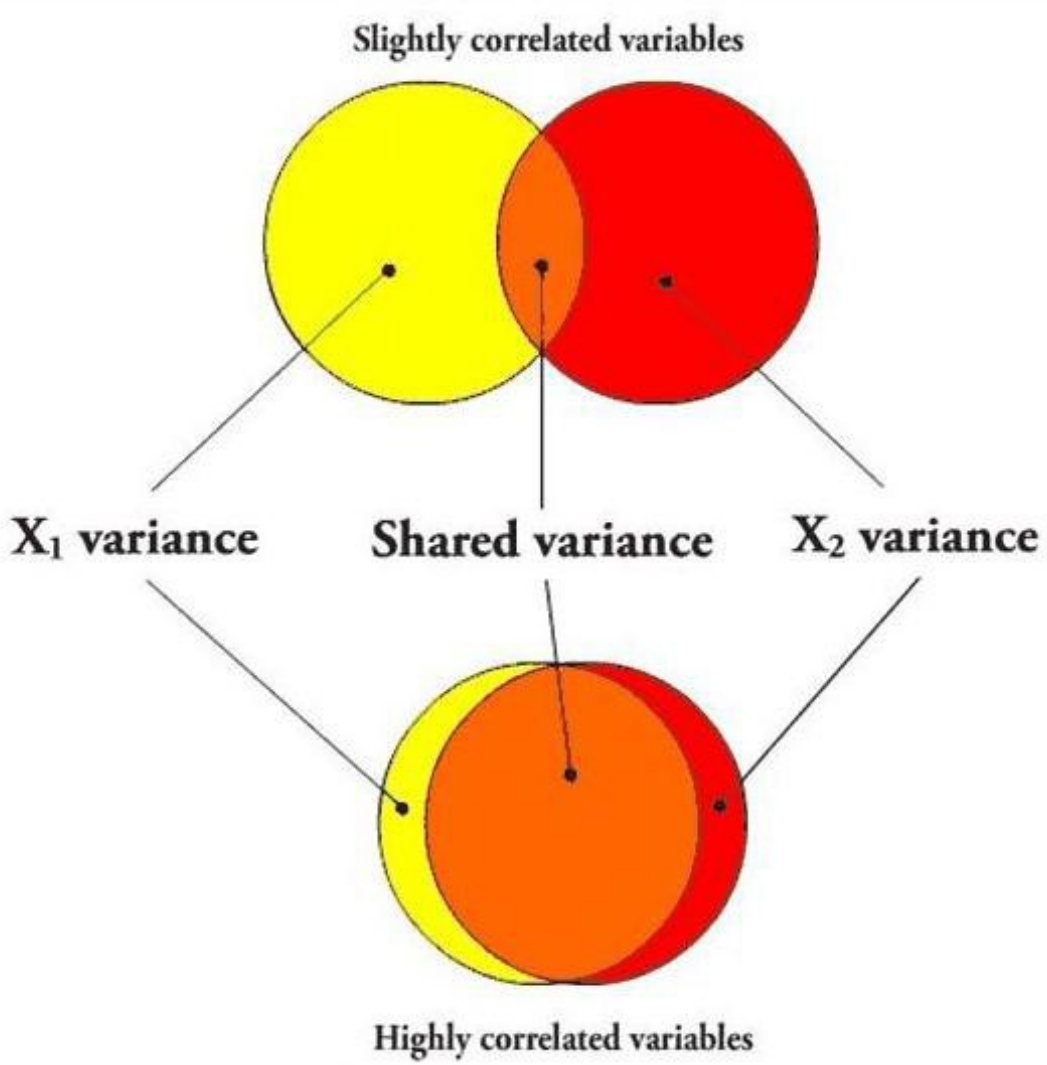
gdzie R_j^2 jest współczynnikiem determinacji w modelu regresji wielokrotnej, w którym zmienna x_j jest wyjaśniana za pomocą pozostałych zmiennych objaśniających. Z tego równania wynika, że wariancja estymatora parametru b_j rośnie wraz ze skorelowaniem j-tego regresora z pozostałymi, a maleje z wariancją j-tej zmiennej. Jeśli teraz wyznaczymy stosunek wariancji estymatora jeśli zmienne są nieskorelowane ($R_j^2 = 0$) i jeśli są skorelowane ($R_j^2 \neq 0$), to otrzymamy

$$\text{VIF}_j = \frac{1}{1 - R_j^2}.$$

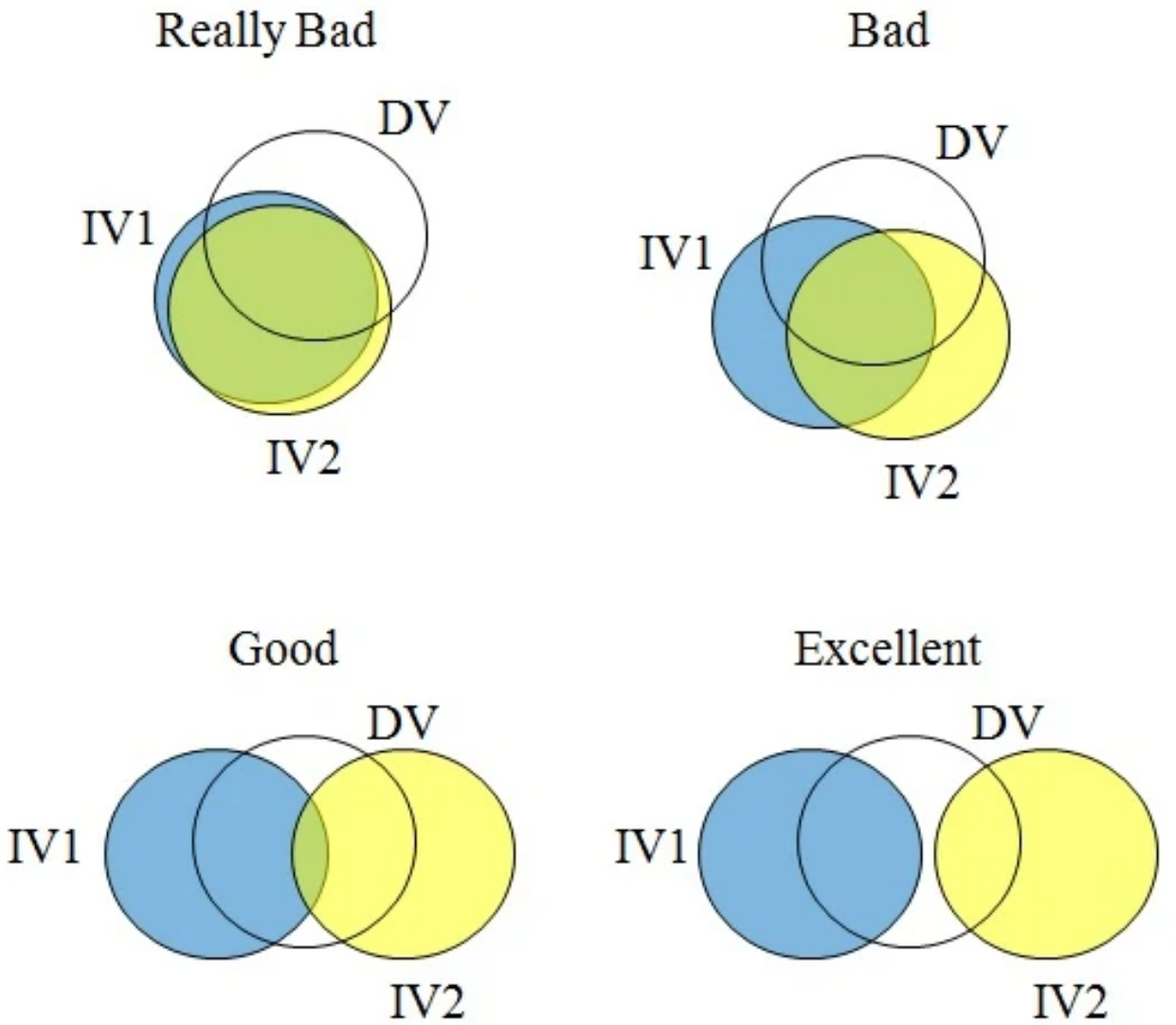
Regresja, a współliniowość zmiennych

Miara ta nazywana jest czynnikiem rozdęcia wariancji (*ang. variance inflation factor*). Mierzy on jaka część wariancji estymatora jest powodowana przez to, że zmienna x_j nie jest ortogonalna (nieskorelowana) względem pozostałych zmiennych objaśniających w modelu regresji. Jeśli VIF dla danej zmiennej przekracza 10 ($R^2 = 0.9$) uznaje się, że mamy problem ze współliniowością dla tej zmiennej. Jeśli VIF jest między 5 ($R^2 = 0.8$), a 10 to uznaje się, że możemy mieć problem ze współliniowością.

Regresja, a współliniowość zmiennych



Regresja, a współliniowość zmiennych



Regresja składowych głównych

Próbą uniknięcia problemu zależności zmiennych objaśniających jest regresja składowych głównych (*ang. principal components regression* – PCR). Zamiast oryginalnych zmiennych objaśniających używamy składowych głównych, które są nieskorelowane. W praktyce używamy jedynie kilku pierwszych składowych, które w zadowalający sposób odzwierciedlają zmienność oryginalnych danych. Pojawia się jednak pewien problem. Ponieważ usuwamy część składowych nigdy nie mamy pewności, że nie usunęliśmy ważnej informacji a zostawiliśmy zaburzenie (wybrane składowe niekoniecznie są maksymalnie skorelowane ze zmienną objaśnianą).

Regresja częściowych najmniejszych kwadratów

Próbą rozwiązania tego ostatniego problemu jest regresja częściowych najmniejszych kwadratów (*ang. partial least squares regression* – PLSR). W przypadku tej metody nowe zmienne objaśniające poszukiwane są w taki sposób, aby oprócz dobrego wyjaśniania zmienności oryginalnych danych, były maksymalnie skorelowane ze zmiennymi objaśnianymi. Metody tej używamy w przypadku gdy chcemy dokonać analizy zależności zbioru zmiennych objaśnianych od bardzo wielu zmiennych objaśniających. Szczególnie użyteczna bywa gdy liczba zmiennych jest większa od liczby obserwacji. Z tych względów szczególnie często bywa używana w chemometrii.

Regresja grzbietowa i LASSO

Inną próbą uniknięcia problemów ze zmiennymi skorelowanymi (lub ich dużą liczbą) jest regresja grzbietowa (*ang. ridge regression* – RR). Ponieważ problemy pojawiają się w związku z niemożnością odwrócenia macierzy $\boldsymbol{X}^\top \boldsymbol{X}$, to do jej przekątnej dodaje się pewną stałą $\lambda \geq 0$. Dla takiego zagadnienia otrzymuje się następujące rozwiązanie:

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Tego samego typu metodami są LASSO (*ang. least absolute shrinkage and selection operator*) oraz elastic net. Wszystkie te metody redukują wariancję estymatorów, aczkolwiek kosztem obciążenia.

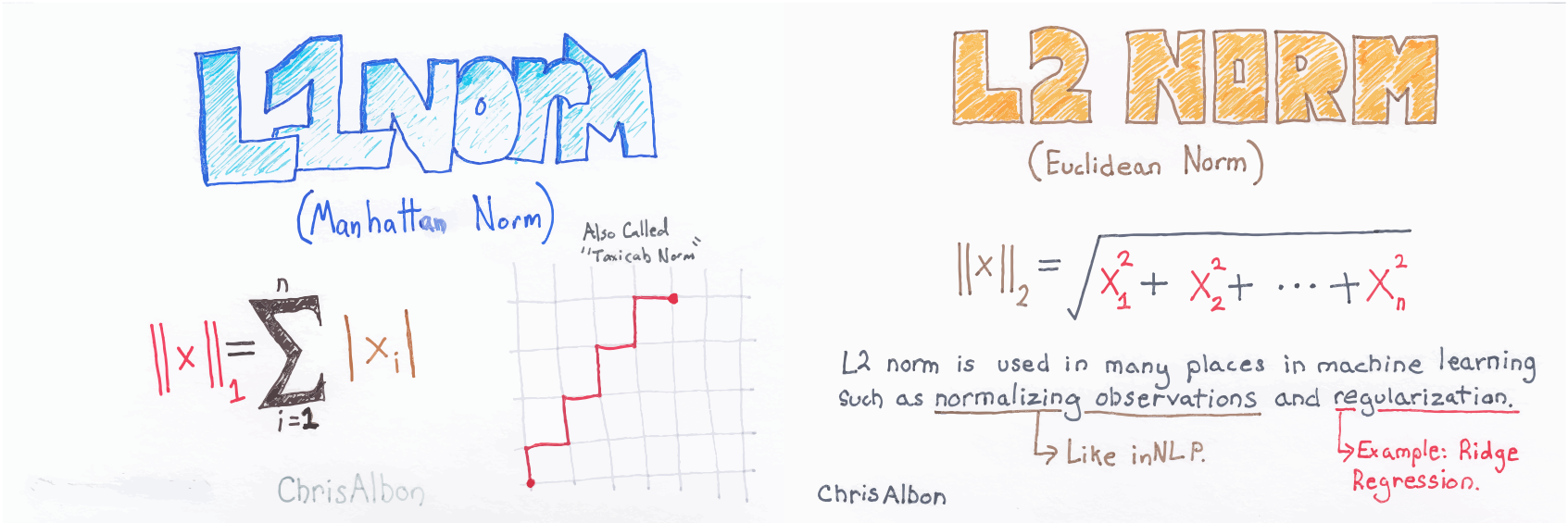
Regresja grzbietowa i LASSO

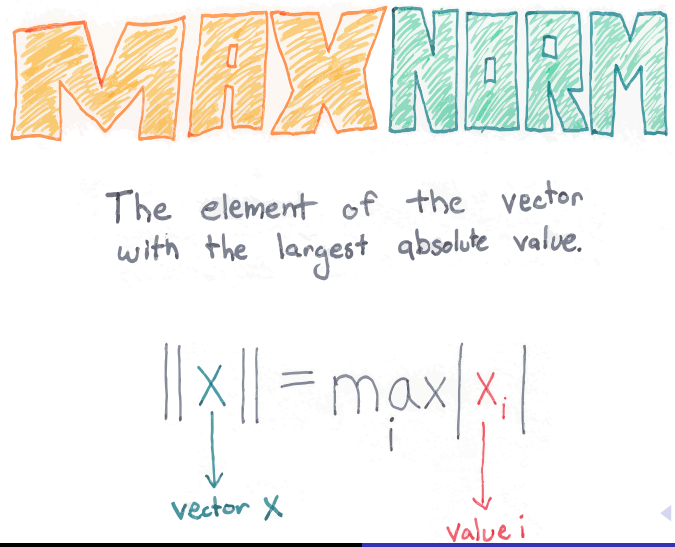
$$\hat{\beta} = \begin{cases} \arg\min_{\beta} ||\boldsymbol{y} - \boldsymbol{X}\beta||^2 & \text{– OLS,} \\ \arg\min_{\beta} ||\boldsymbol{y} - \boldsymbol{X}\beta||^2 + \lambda_2 ||\beta||_2^2 & \text{– RR,} \\ \arg\min_{\beta} ||\boldsymbol{y} - \boldsymbol{X}\beta||^2 + \lambda_1 ||\beta||_1 & \text{– LASSO,} \\ \arg\min_{\beta} ||\boldsymbol{y} - \boldsymbol{X}\beta||^2 + \lambda_2 ||\beta||_2^2 + \lambda_1 ||\beta||_1 & \text{– Elastic net.} \end{cases}$$

Gdzie $||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2$ i $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$.

Ponieważ estymatory w RR są wrażliwe na skalowanie danych, to powinniśmy raczej skalować dane przed jej przeprowadzeniem.

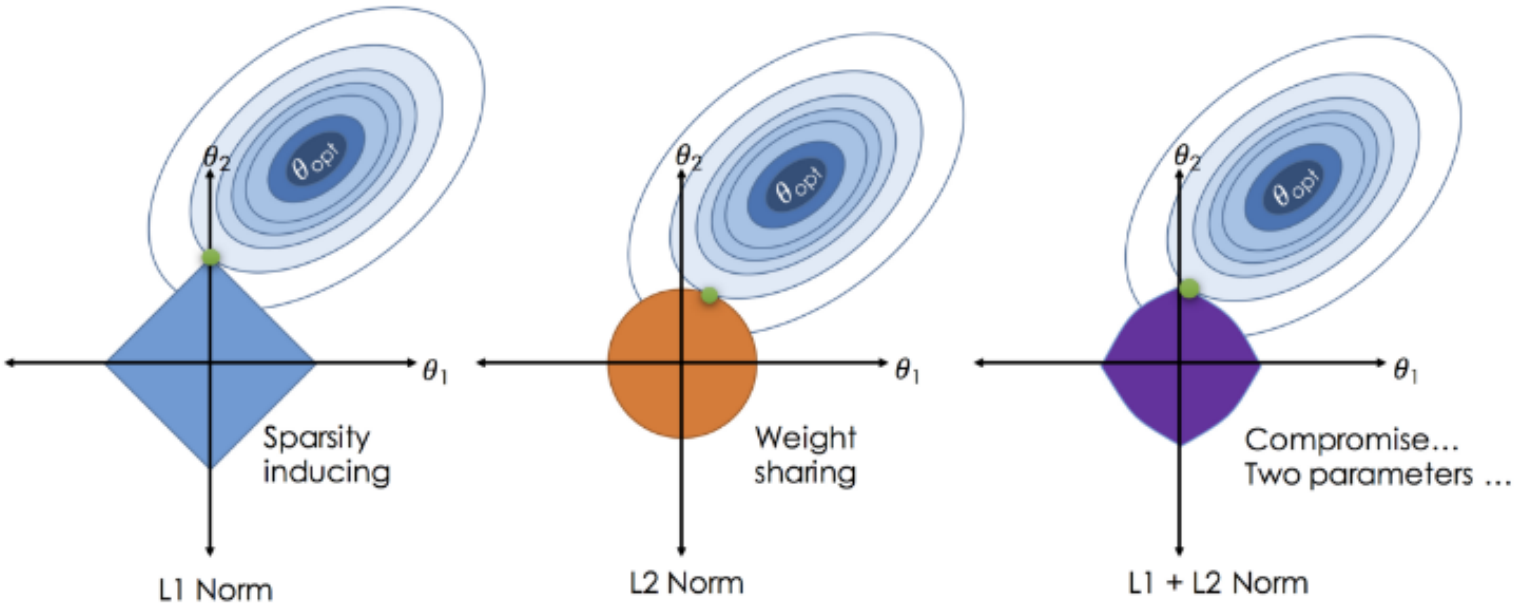
Regresja grzbietowa i LASSO





Regresja grzbietowa i LASSO

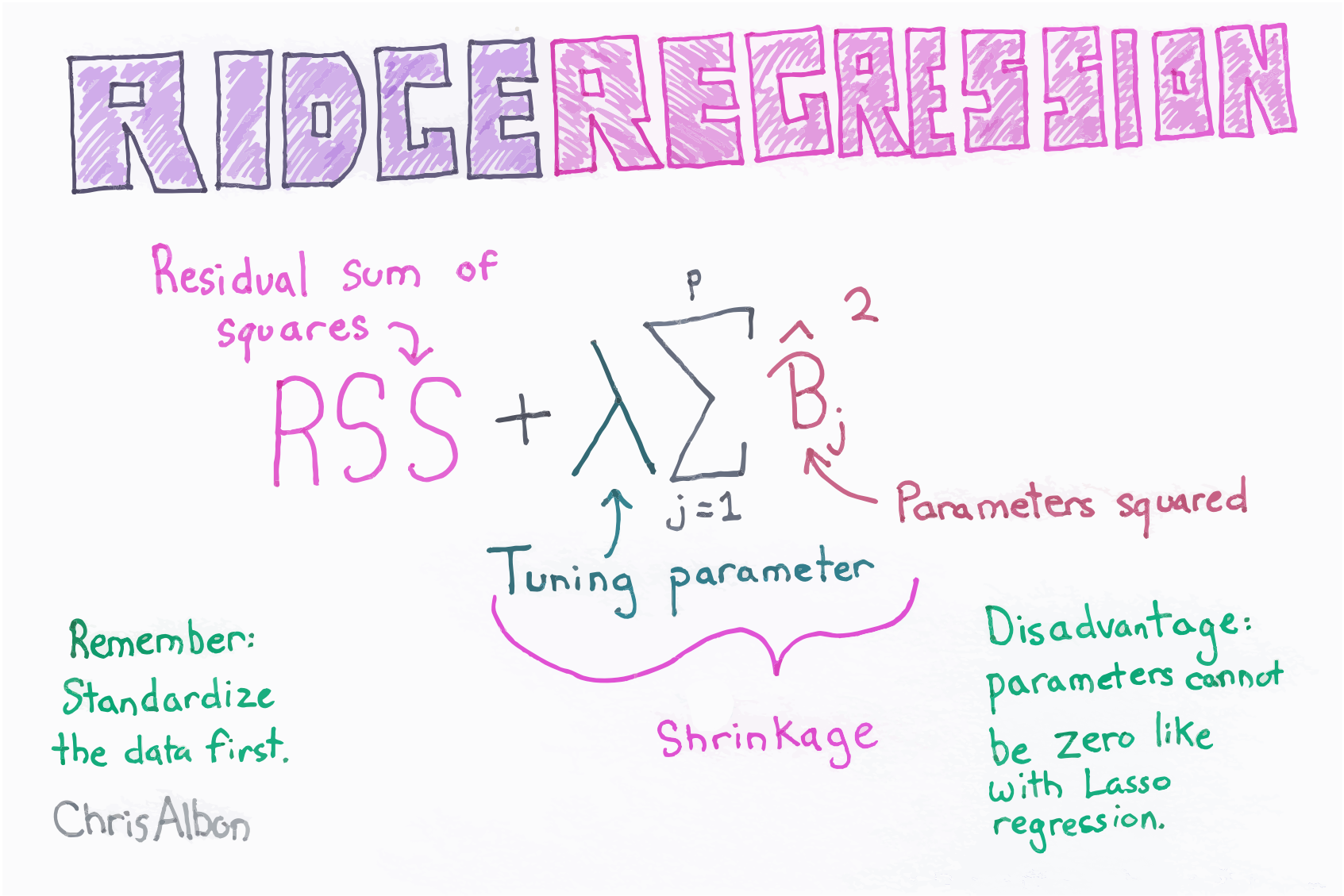
Współczynniki uzyskane za pomocą metody LASSO dają najmniejsze MSE z wszystkich punktów wewnątrz hiperkostki $\sum_{i=1}^{p} |\beta_i| \leq s$. Z drugiej strony, współczynniki uzyskane za pomocą metody RR dają najmniejsze MSE z wszystkich punktów wewnątrz hiperkuli $\sum_{i=1}^{p} \beta_i^2 \leq s$.



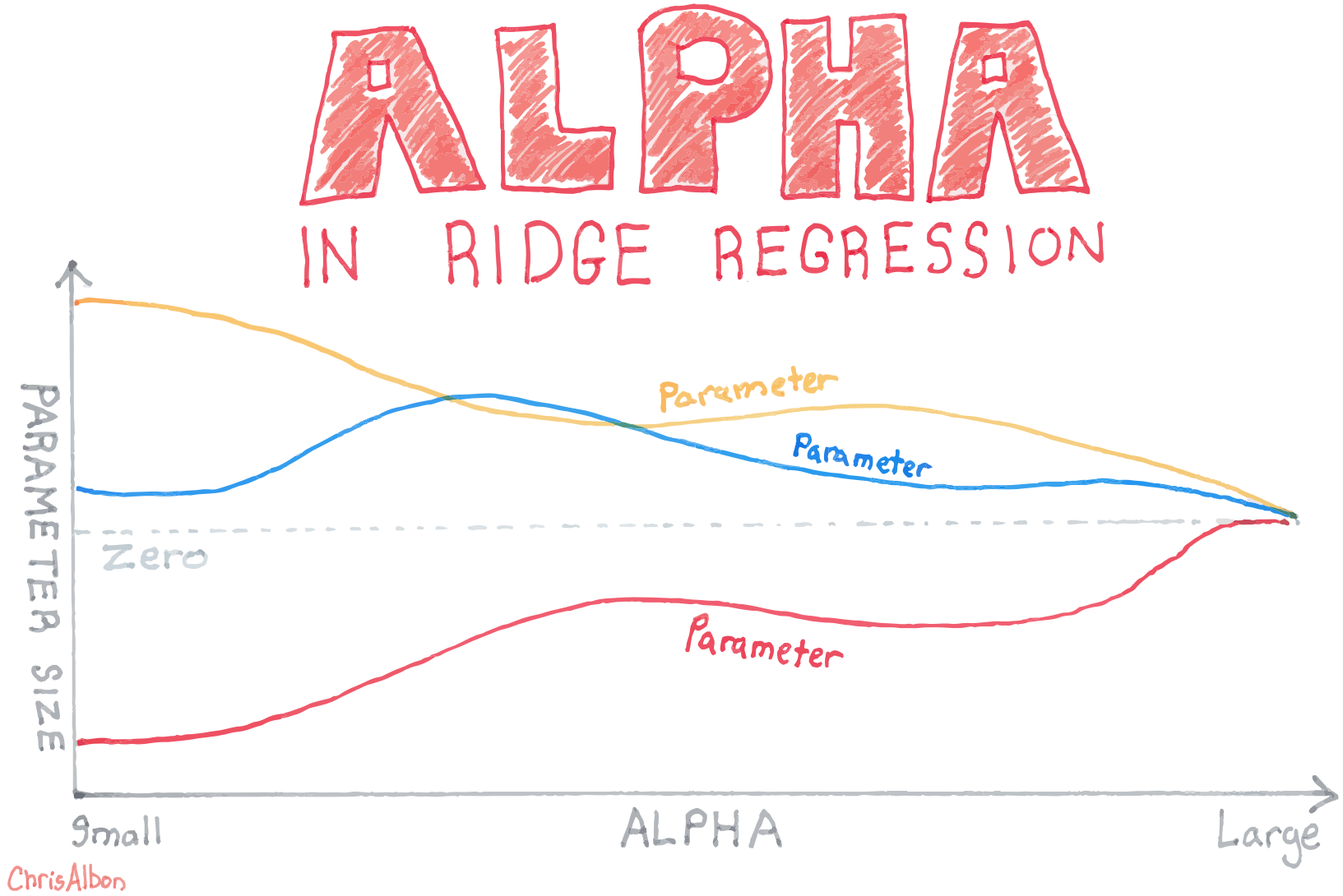
Regresja grzbietowa i LASSO

S.No	L1 Regularization	L2 Regularization
1	Panelizes the sum of absolute value of weights.	penalizes the sum of square weights.
2	It has a sparse solution.	It has a non-sparse solution.
3	It gives multiple solutions.	It has only one solution.
4	Constructed in feature selection.	No feature selection.
5	Robust to outliers.	Not robust to outliers.
6	It generates simple and interpretable models.	It gives more accurate predictions when the output variable is the function of whole input variables.
7	Unable to learn complex data patterns.	Able to learn complex data patterns.
8	Computationally inefficient over non-sparse conditions.	Computationally efficient because of having analytical solutions.

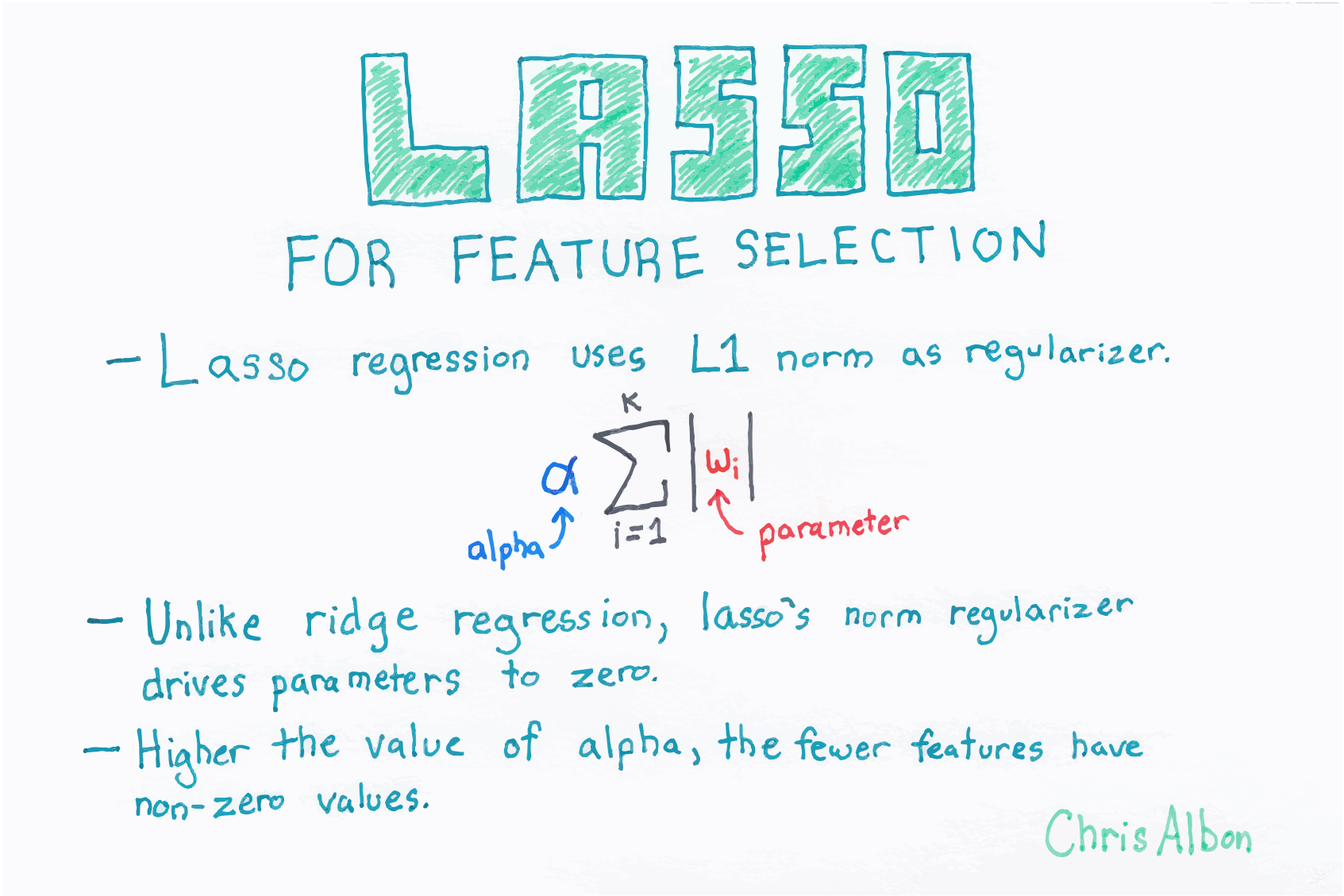
Regresja grzbietowa i LASSO



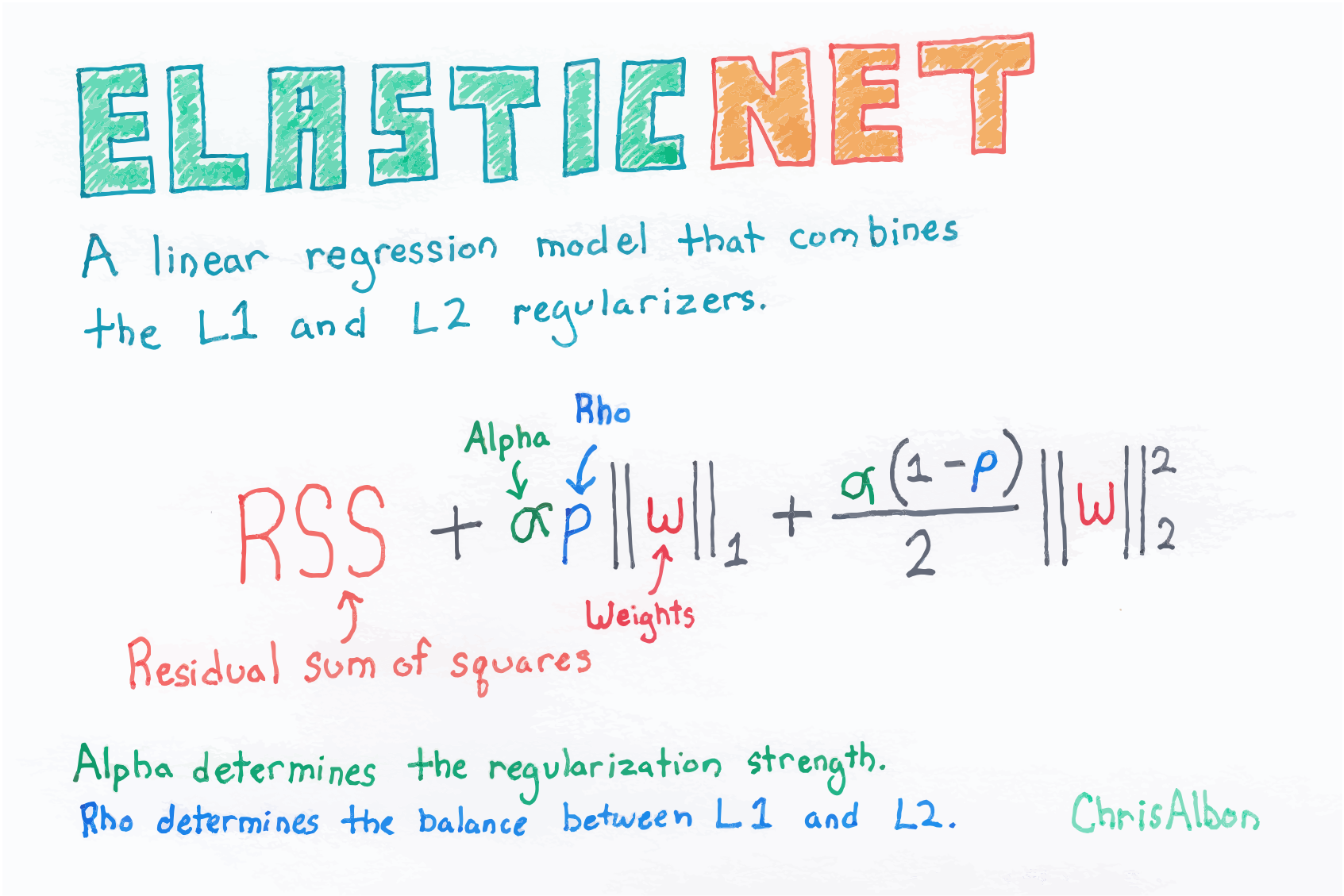
Regresja grzbietowa i LASSO



Regresja grzbietowa i LASSO



Regresja grzbietowa i LASSO



Regresja nieliniowa – wprowadzenie

W wielu zagadnieniach model regresji liniowej nie wyraża dobrze zależności między zmiennymi. Musimy wówczas zrezygnować z funkcji liniowej i wykorzystać regresję nieliniową. Modele takie można podzielić na:

- modele nieliniowe względem zmiennych objaśniających, ale liniowe względem parametrów,
- modele nieliniowe zarówno względem zmiennych objaśniających jak i parametrów, dla których istnieje transformacja do modelu liniowego,
- modele ściśle nieliniowe, tzn. modele nieliniowe zarówno względem zmiennych objaśniających jak i parametrów, dla których nie istnieje transformacja do modelu liniowego.

Regresja nieliniowa – przykład

Model liniowy względem parametrów, ale nieliniowy względem zmiennych objaśniających.

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2 + \frac{\alpha_3}{x_2} + \varepsilon$$

Modele takie możemy w prosty sposób sprowadzić do modelu liniowego poprzez odpowiednie podstawienie:

$$x_1' = x_1; x_2' = x_1^2; x_3' = \frac{1}{x_2}$$

otrzymując:

$$Y = \alpha_0 + \alpha_1 x_1' + \alpha_2 x_2' + \alpha_3 x_3' + \varepsilon$$

Regresja nieliniowa – przykład

Model wykładniczo-hiperboliczny (nieliniowy zarówno względem zmiennych objaśniających jak i parametrów):

$$Y = e^{\alpha_0 + \frac{\alpha_1}{x_1} + \varepsilon}.$$

Modele takie sprowadzamy do modelu liniowego poprzez transformacje zarówno zmiennych objaśniających jak i zmiennej objaśnianej. Logarytmując obustronnie otrzymujemy:

$$\ln Y = \alpha_0 + \frac{\alpha_1}{x_1} + \varepsilon.$$

$$Y' = \ln Y; x_1' = \frac{1}{x_1}$$

$$Y' = \alpha_0 + \alpha_1 x_1' + \varepsilon.$$

Regresja nieliniowa – przykład

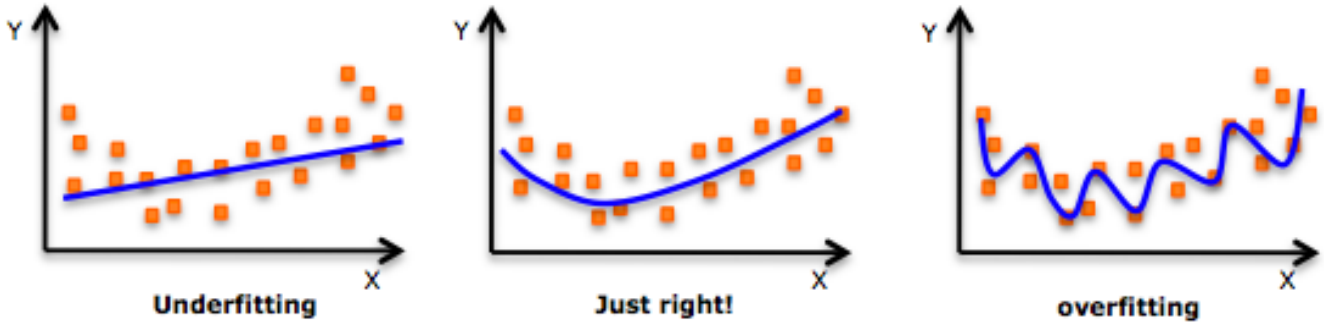
Model:

$$Y = \alpha_0 + \alpha_1 e^{\alpha_2 x_1} + \varepsilon$$

jest niesprowadzalny do modelu liniowego poprzez żadną transformację zarówno zmiennych objaśniających jak i zmiennej objaśnianej.

Regresja nieliniowa – uwaga

Jeśli tylko to możliwe zaleca się estymację parametrów regresji nieliniowej, a nie linearyzację modelu i estymację parametrów regresji liniowej.



Regresja nieliniowa w R

Model	# param.	Funkcja w R	Równanie
Michaelis-Menten	2	SSmicmen()	$y = \frac{ax}{b+x}$
Gompertz	3	SSgompertz()	$y = ae^{-bc^x}$
First-order compartment	3	SSfol()	$y = \frac{x_1 e^{a+b-c}}{e^a - e^b}\left(e^{-e^a x_2} - e^{-e^b x_2}\right)$
Asymptotic reg.	3	SSasymp()	$y = a + (b-a)e^{-e^c x}$
Asymptotic reg. with offset	3	SSasympOff()	$y = a(1 - e^{-e^b(x-c)})$
Asymptotic reg. $(c = 0)$	2	SSasympOrig()	$y = a(1 - e^{-e^b x})$
Weibull	4	SSweibull()	$y = a - be^{-e^c x^d}$
Biexponential	4	SSbiexp()	$y = ae^{-e^b x} + ce^{-e^d x}$
Logistic	4	SSfpl()	$y = a + \frac{b-a}{1+e^{(c-x)/d}}$
Logistic $(a = 0)$	3	SSlogis()	$y = \frac{a}{1+e^{(b-x)/c}}$

Regresja nieparametryczna

Czasami nie możemy zaproponować żadnej sensownej funkcji regresji lub też interesuje nas jedynie „wygląd". W takiej sytuacji możemy wyznaczyć linię trendu stosując nieparametryczne metody regresji:

- Lokalne wygładzanie wielomianami niskiego stopnia. Dzielimy zbiór wartości funkcji na rozłączne przedziały i na każdym kawałku dopasowujemy regresję wielomianową (najczęściej trzeciego stopnia).
- Wygładzanie jądrowe (regresja jądrowa).
- Regresja najbliższych sąsiadów. Wybierany jest parametr k, który wskazuje jaka część danych ma posłużyć do budowy modelu regresji liniowej. W celu oceny wartości x_i używane są obserwacje $x_{i-k/2}, \ldots, x_i, \ldots, x_{i+k/2}$.
- Ważona regresja lokalnie wielomianowa. Obserwacje otrzymują wagi (bliższe większe, dalsze mniejsze), a ocena x_i otrzymywana jest za pomocą odpornej regresji ważonej.

Regresja nieparametryczna w R

W R metody lokalne można uzyskać za pomocą funkcji: smooth.spline() (sześcienne funkcje sklejane, jest to wygładzona wersja funkcji spline()); supsmu() (regresja najbliższych sąsiadów); lowess() (ważona regresja lokalnie wielomianowa – *ang. locally weighted scatterplot smoothing*); scatter.smooth (punkty oraz trend, modyfikacja funkcji lowess()); ksmooth() (wygładzanie jądrowe – *ang. kernel regression, kernel smoother*).

Jądro

Definicja

Jądrem (funkcją jądrową – *ang. kernel*) będziemy nazywać każdą funkcję gładką (ma ciągłe pochodne wszystkich rzędów) K taką, że

- $K(x) \geq 0$ (nieujemna),
- $\int_{-\infty}^{\infty} K(x)dx = 1$ (unormowana),
- $K(x) = K(-x)$ (symetryczna).

Jądro

- Jądro jednostajne (*ang. uniform*):

$$K(x) = \frac{1}{2} I_{|x| \leq 1}(x),$$

- Jądro gaussowskie (*ang. Gaussian*):

$$K(x) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2}\right) I_{\mathbb{R}}(x),$$

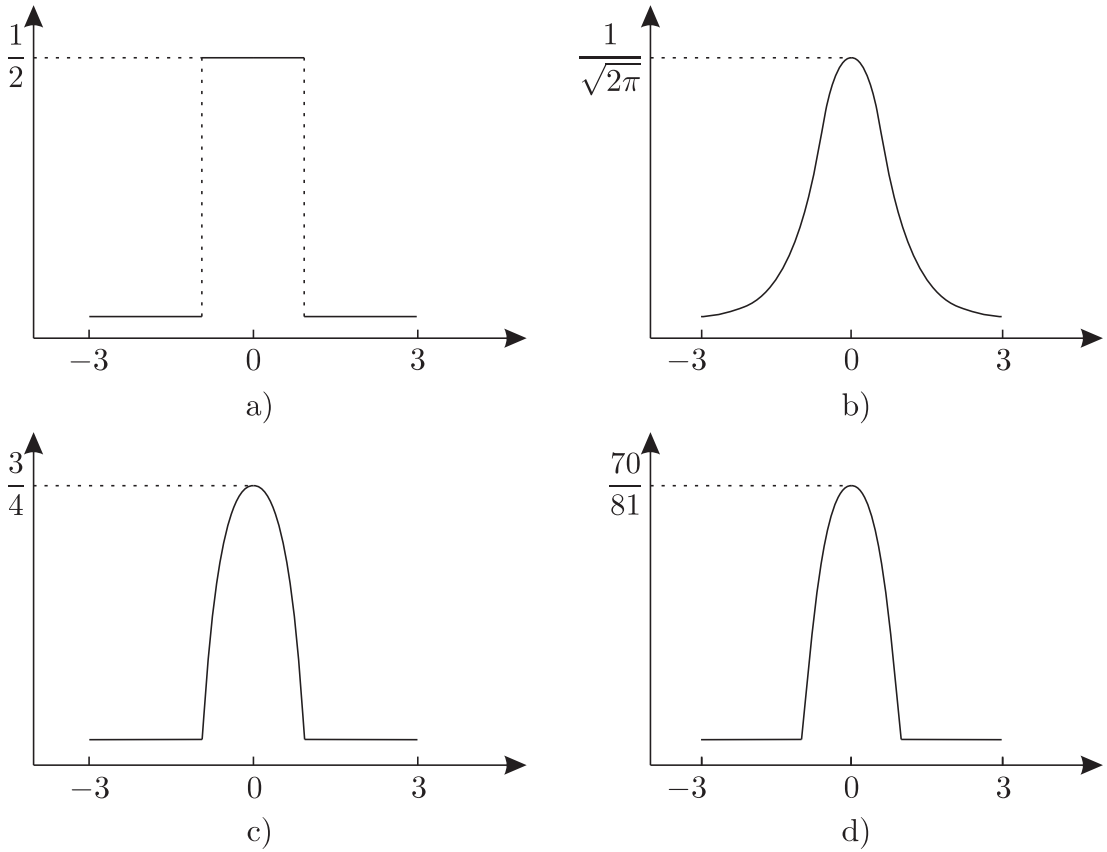
- Jądro Epanechnikowa (*ang. Epanechnikov*):

$$K(x) = \frac{3}{4}(1 - x^2) I_{|x| \leq 1}(x),$$

- Jądro stopnia trzeciego (*ang. tricube*):

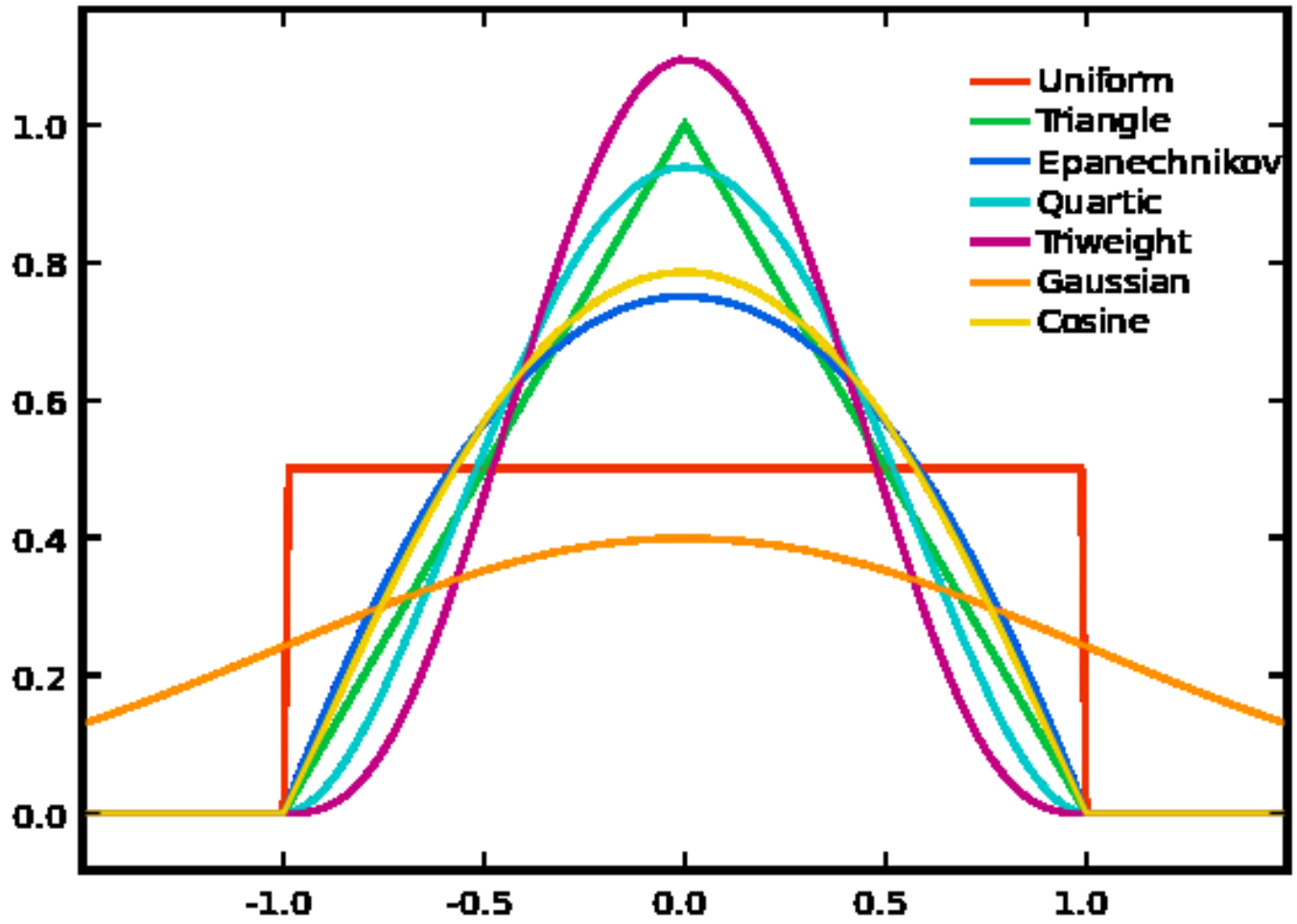
$$K(x) = \frac{70}{81}(1 - |x|^3)^3 I_{|x| \leq 1}(x).$$

Jądro



Przykłady funkcji jądrowych: a) jądro jednostajne b) jądro gaussowskie c) jądro Epanechnikowa *d) jądro stopnia trzeciego.*

Jądro


Uniform
Triangle
Epanechnikov
Quartic
Triweight
Gaussian
Cosine
1.0
0.8
0.6
0.4
0.2
0.0
-1.0
-0.5
0.0
0.5
1.0

Jądro

Definicja

Dla danego jądra K i dodatniej liczby h, zwanej współczynnikiem gładkości (ang. *bandwith*), jądrowy estymator Rosenblatta-Parzena gęstości f jest równy

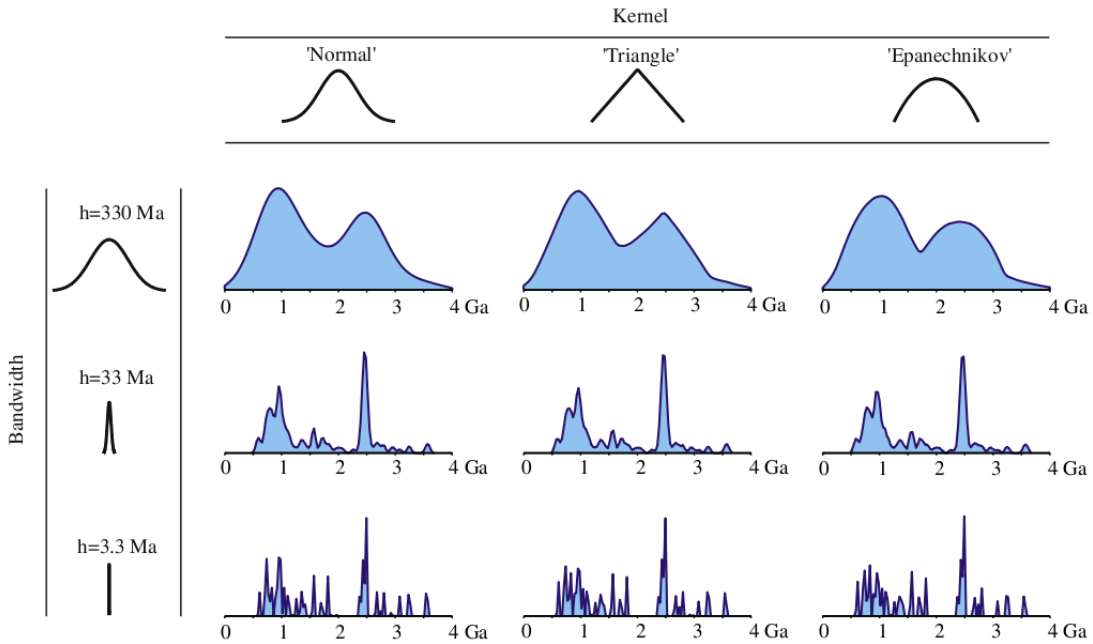
$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right).$$


Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics* 27(3):832–837.


Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* 33(3):1065–1076.

Jądro

W celu skonstruowania jądrowego estymatora gęstości, musimy wybrać jądro K i współczynnik gładkości h. Poniższy wykres pokazuje, że wybór jądra ma niewielkie znaczenie. Z drugiej strony decydująca jest szerokość okna (na górze zbyt długie okno prowadzi do nadmiernego wygładzenia, na dole zbyt wąskie okno prowadzi do zbyt małego wygładzenia).



Jądro

Najpopularniejszym estymatorem jądrowym funkcji regresji jest estymator NADARAYI-WATSONA postaci:

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)},$$

gdzie

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right).$$


Nadaraya, E.A. (1964). On Estimating Regression. *Theory of Probability and its Applications* 9(1):141–2.


Watson, G.S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A* 26(4):359–372.

Jądro

Popularne metody wyznaczania h (dwie pierwsze dla rozkładu normalnego):

- SILVERMAN's rule of thumb (nrd0)

$$0.9 \cdot \min\left(\hat{\sigma}, \frac{\text{IQR}}{1.34}\right) n^{-1/5}.$$

- SCOTT's rule of thumb (nrd)

$$1.06 \cdot \min\left(\hat{\sigma}, \frac{\text{IQR}}{1.34}\right) n^{-1/5}.$$

- SHEATHER-JONES (SJ)

Scott, D. W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B* 53(3):683–690.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.*. Chapman and Hall.

Regresja logistyczna – wprowadzenie

W wielu sytuacjach nie możemy założyć, że zmienna objaśniana jest ciągła. W takiej sytuacji powinniśmy wykorzystać uogólnione modele liniowe (*ang. generalized linear models*), w których na zmienną zależną nakłada się rozkład (dopuszczalne są rozkłady pochodzące z tzw. wykładniczej rodziny rozkładów: np. rozkład normalny, wykładniczy, gamma, POISSONA, dwumianowy, geometryczny oraz wielomianowy). Poza tym, aby uwzględnić również nieliniowy charakter zależności wprowadza się tzw. funkcję wiążącą (*ang. link function*) h, która ma następującą własność:

$$h(\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})) = \boldsymbol{X}\boldsymbol{\beta}.$$

Zauważmy, że jeśli funkcja wiążąca jest identycznością ($h(x) = x$), a zmienna objaśniana ma rozkład normalny, to model ten sprowadza się do modelu regresji liniowej.

Regresja logistyczna – wprowadzenie

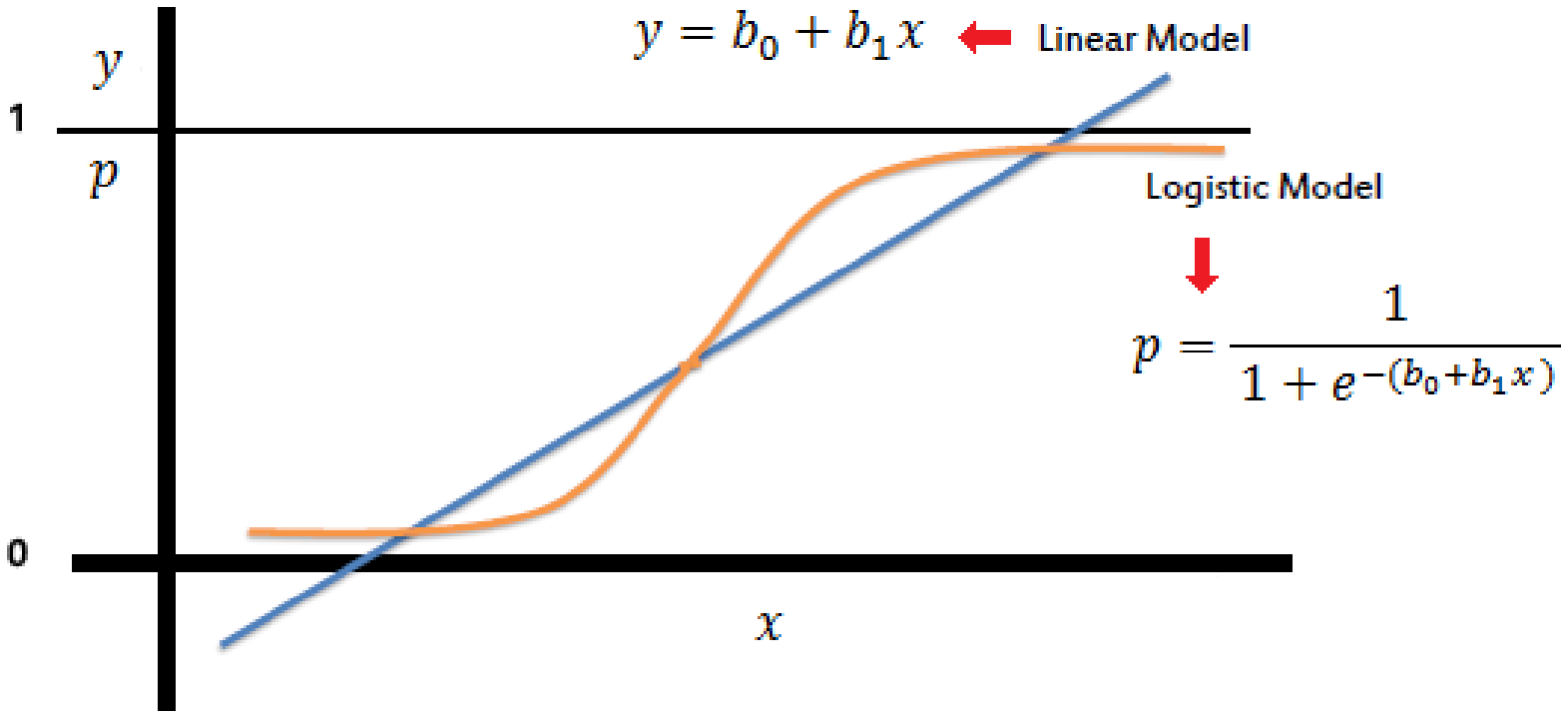
Szczególnym i bardzo ważnym przykładem uogólnionego modelu liniowego jest regresja logistyczna (*ang. logistic regression*). Formalnie w tym przypadku zakładamy, że $Y \sim b(p)$. Oznacza, to, że zmienna objaśniana przyjmuje tylko dwie wartości (najczęściej jest to zmienna binarna). Modelujemy prawdopodobieństwo wystąpienia sukcesu p. Jako funkcja wiążąca używana jest funkcja logitowa (*ang. logit function*):

$$\text{logit}(p) = \ln \frac{p}{1-p}.$$

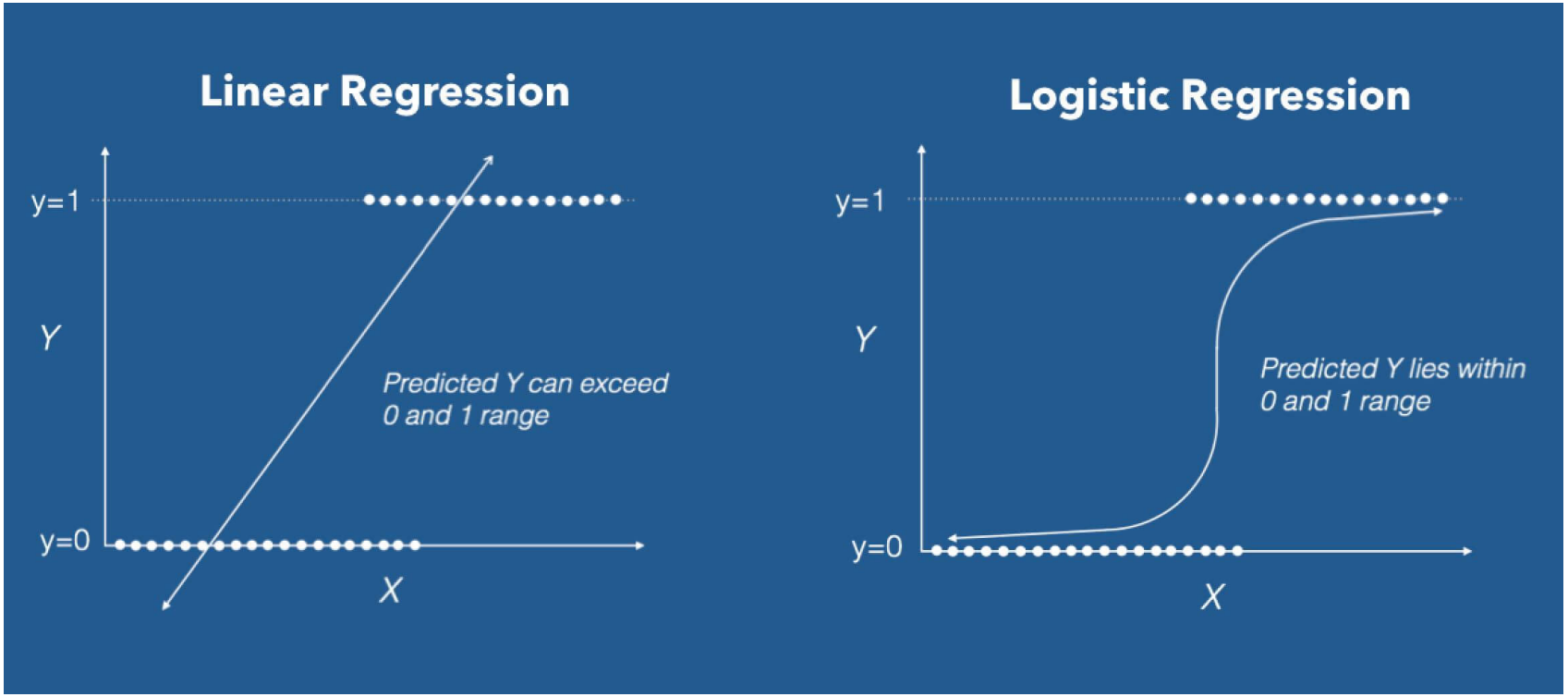
Prawdopodobieństwo p jest następnie szacowane jako:

$$p = \frac{1}{1 + \exp(-\boldsymbol{X}\beta)}.$$

Regresja logistyczna – wprowadzenie



Regresja logistyczna – wprowadzenie



Regresja logistyczna – wprowadzenie

Regresja logistyczna – wprowadzenie

Logistic Regression

$z = b + a_1x_1 + a_2x_2 + a_3x_3$
$p = 1.0 / (1.0 + e^{-z})$

Ex:

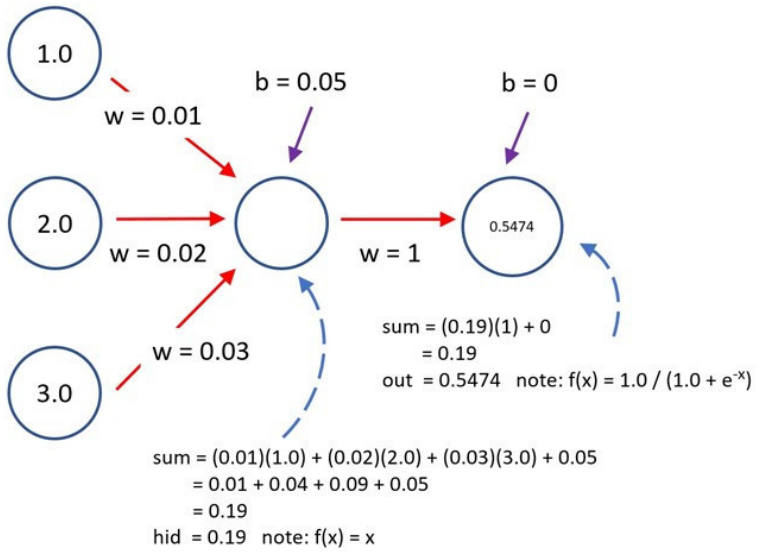
$x_1 = 1.0 \quad a_1 = 0.01$
$x_2 = 2.0 \quad a_2 = 0.02$
$x_3 = 3.0 \quad a_3 = 0.03$
$b = 0.05$

$z = (0.05) + (0.01)(1.0) + (0.02)(2.0) + (0.03)(3.0)$
$= 0.05 + 0.01 + 0.04 + 0.09$
$= 0.19$

$p = 1.0 / (1.0 + e^{-0.19})$
$= 0.5474$ (predicted class = 1)

Neural Network

single hidden layer, identity activation f(x) = x
single output node, logistic sigmoid activation $f(x) = 1 / (1 + e^{-x})$



Regresja logistyczna, a sieć neuronowa

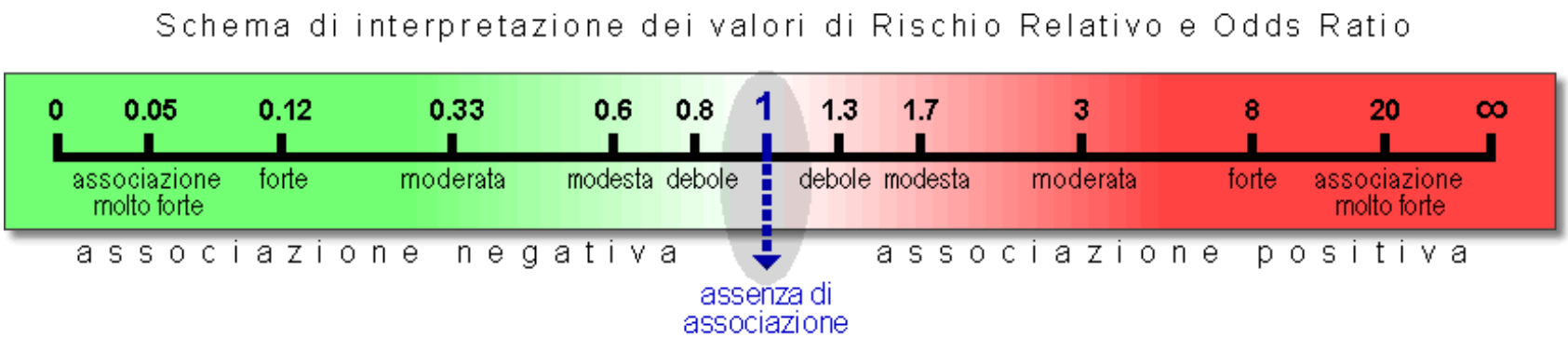
Regresja logistyczna jest bardzo prostą siecią neuronową!!!

Regresja logistyczna – iloraz szans

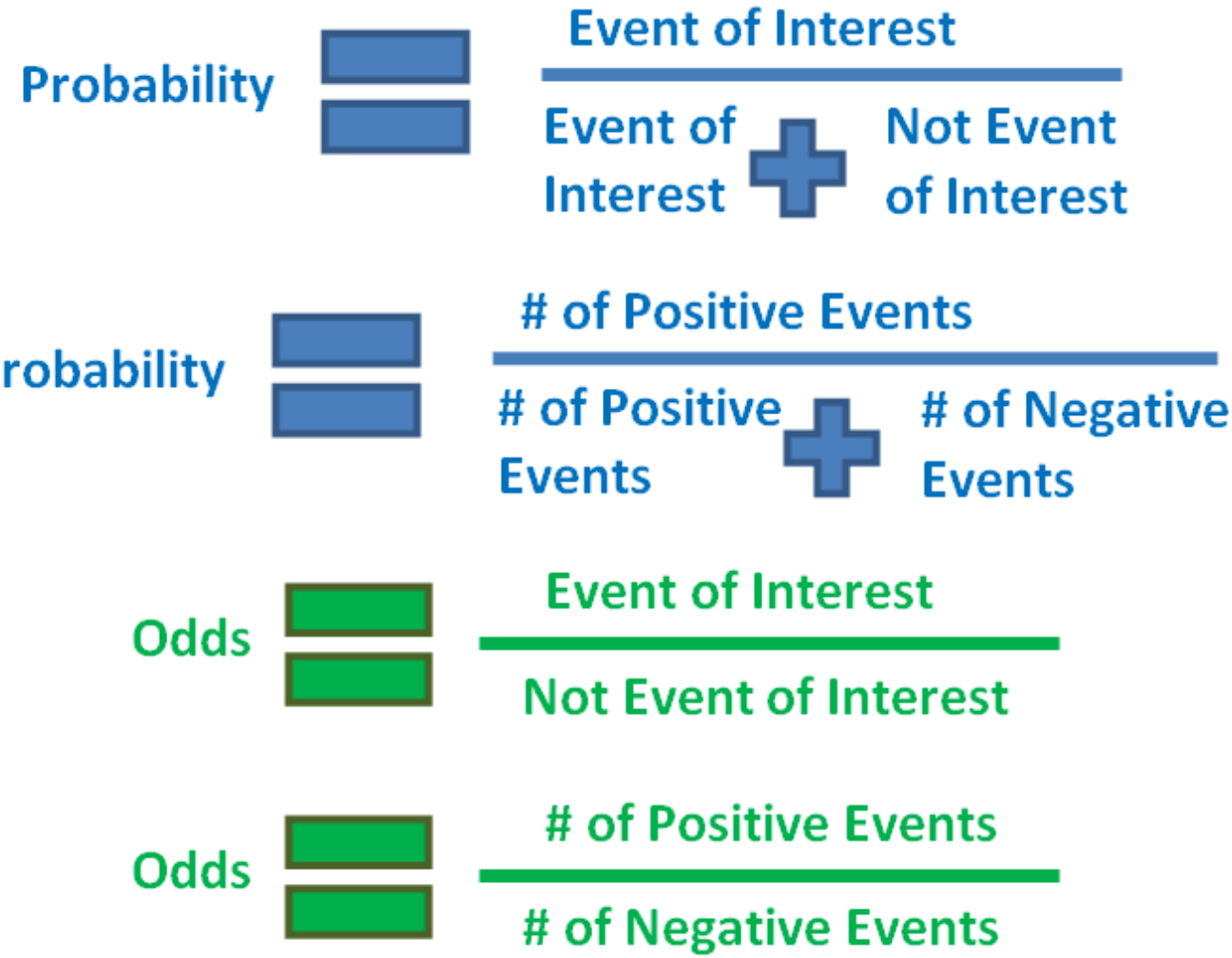
Interpretacji podlega iloraz szans (*ang. odds ratio*), który można wyrazić jako

$$\ln(\mathrm{OR}) = \ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

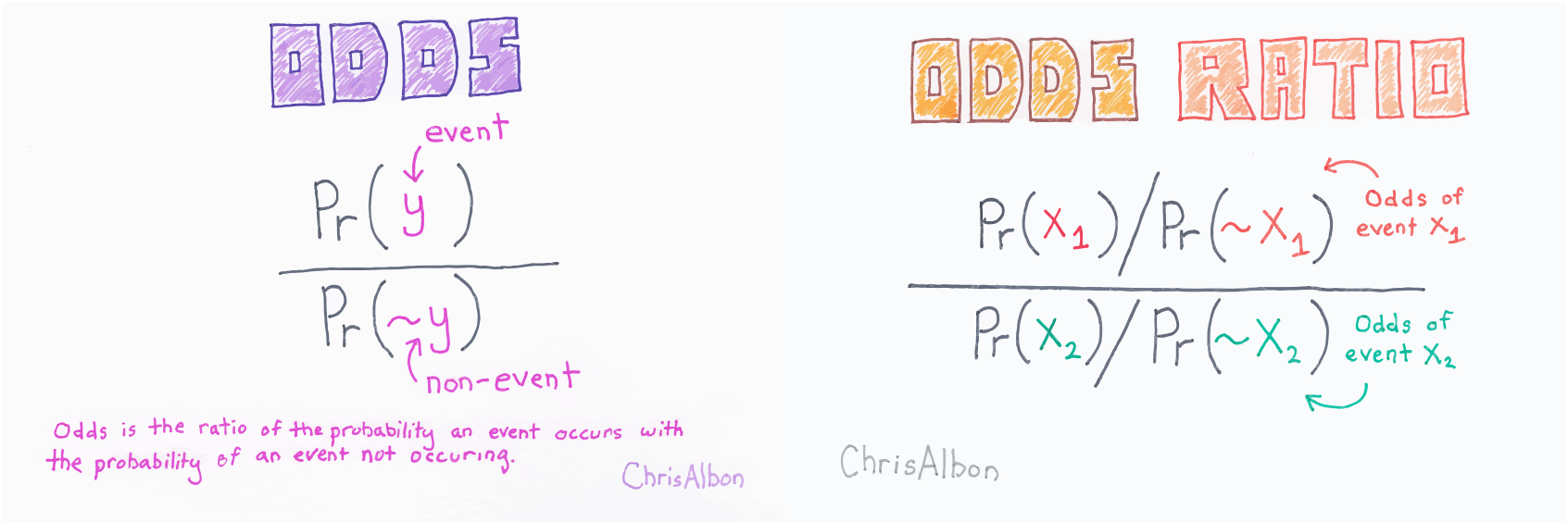
Jeżeli $e^{\beta_j} > 1$, to zmienna X_j działa stymulująco na możliwość wystąpienia badanego zjawiska, w przeciwnym razie działa ograniczająco (jeżeli $e^{\beta_j} = 1$, to zmienna X_j nie ma wpływu na badane zjawisko).



Regresja logistyczna – iloraz szans



Regresja logistyczna – iloraz szans



Regresja logistyczna – krzywe ROC

Jakość dopasowania można jak poprzednio zbadać za pomocą kryterium informacyjnych, jednak w przypadku regresji logistycznej bardziej efektywne są inne kryteria.
Jednym z nich są krzywe charakterystyczne zwane też krzywymi ROC (*ang. receiver operating characteristic curve*). Na model regresji logistycznej można spojrzeć jak na model, który służy do zdiagnozowania dwóch stanów: dobry/zły. Model liczy prawdopodobieństwa stanu „dobry". Wybieramy pewien próg $0 < t < 1$, jeżeli prawdopodobieństwo uzyskane z modelu jest powyżej t diagnozujemy stan jako „dobry", w przeciwnym razie jest „zły".

Regresja logistyczna – krzywe ROC

Mamy zatem cztery możliwości:

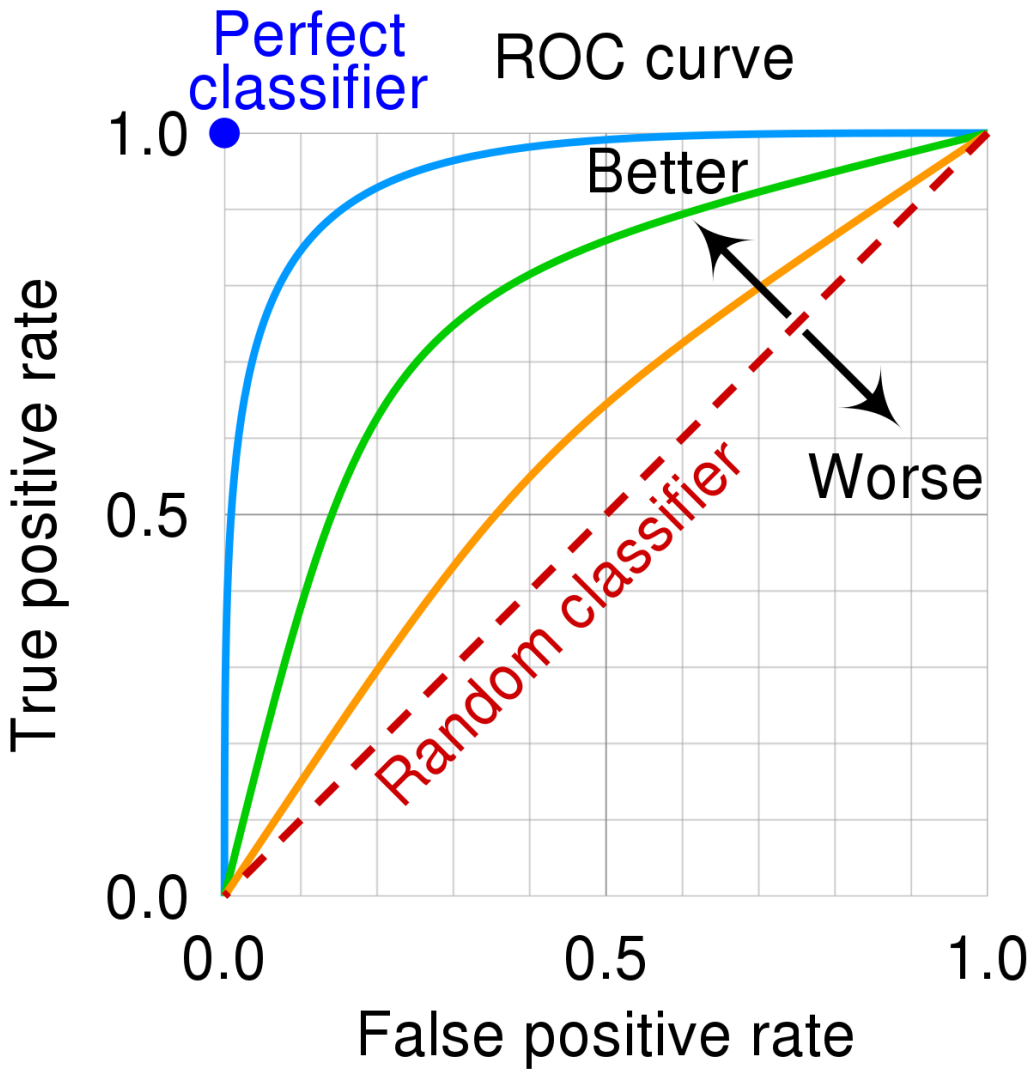
- TP (*ang. true positive*) – model przewidział „dobry" oraz zaobserwowano „dobry",
- TN (*ang. true negative*) – model przewidział „zły" oraz zaobserwowano „zły",
- FP (*ang. false positive*) – model przewidział „dobry" oraz zaobserwowano „zły",
- FN (*ang. false negative*) – model przewidział „zły" oraz zaobserwowano „dobry".

		Zaobserwowano	
		dobry	zły
Przewidziano	dobry	TP	FP
	zły	FN	TN

Regresja logistyczna – krzywe ROC

Jeśli teraz przez n_g oznaczymy liczbę zaobserwowanych „dobry", a przez n_b „zły", to TPR $= \mathrm{TP}/n_g$, TNR $= \mathrm{TN}/n_b$, FPR$= 1-$TNR oraz FNR$= 1-$TPR. Krzywa ROC jest to wykres współczynnika TPR, na osi pionowej przeciwko współczynnikowi FPR na osi poziomej dla wszystkich wartości progowych t. Krzywa ROC jest to zatem rodzina punktów (FPR, TPR) obrazująca zależność między zdolnością wyróżniania przypadków pozytywnych i negatywnych dla różnych parametrów modelu. Aby teraz zmierzyć jakość modelu liczy się pole pod krzywą ROC: AUC (*ang. area under curve*). Im wielkość tego pola bliższa 1 tym zdolność modelu do przewidywania stanu „dobry" lepsza, pole bliskie 0,5 oznacza model bardzo słaby (losowy).

Regresja logistyczna – krzywe ROC



Regresja logistyczna – współczynnik McFaddena

Jak wiemy klasyczny współczynnik dopasowania R^2 nie może być używany do porównywania modeli innych niż liniowe. Spośród wielu propozycji jego zastąpienia najpopularniejszy dla regresji logistycznej jest współczynnik pseudo-R^2 McFaddena (*ang. McFadden's pseudo* R^2) postaci:

$$R^2_{\text{McFadden}} = 1 - \frac{\ln L_M}{\ln L_0},$$

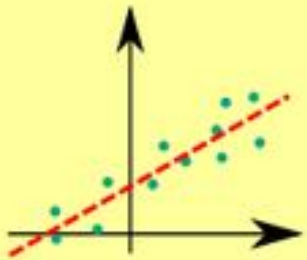
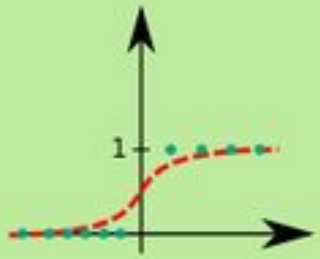
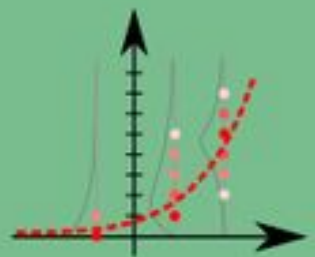
gdzie L_M jest funkcją największej wiarogodności dla badanego modelu, a L_0 jest funkcją wiarogodności modelu pustego (tylko średnia). Przyjmuje wartości między 0 a 1. Wartości pomiędzy 0.2 a 0.4 oznaczają już bardzo dobre dopasowanie.

Regresja logistyczna – współczynnik McFaddena

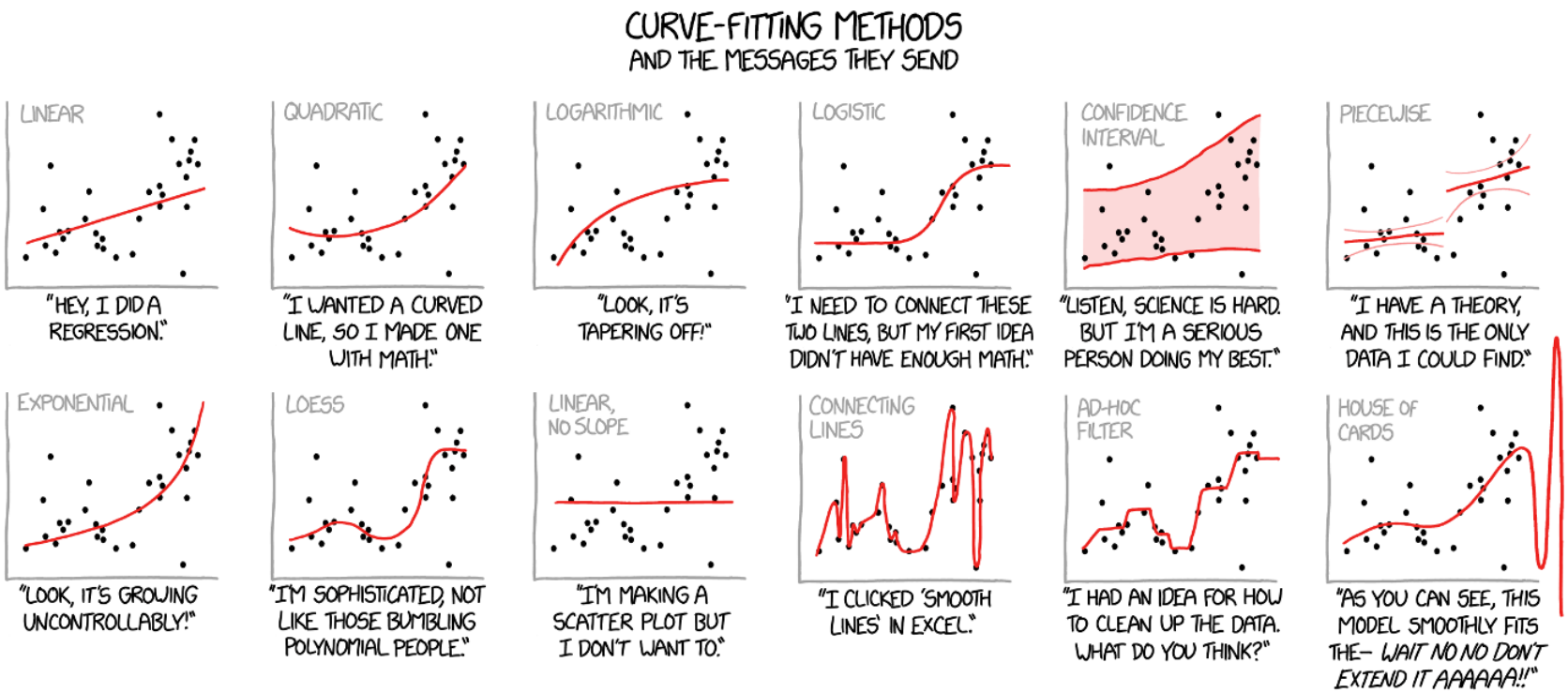
Daniel Little McFadden (1937-)

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometric* 105–142.

Porównanie

LINEAR REGRESSION	LOGISTIC REGRESSION	POISSON REGRESSION
❶ Econometric modelling ❷ Marketing Mix Model ❸ Customer Lifetime Value	❶ Customer Choice Model ❷ Click-through Rate ❸ Conversion Rate ❹ Credit Scoring	❶ Number of orders in lifetime ❷ Number of visits per user
		
Continuous ⇒ Continuous	Continuous ⇒ True/False	Continuous ⇒ 0,1,2,...
$y = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$	$y = \frac{1}{1+e^{-z}}$ $z = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$	$y \sim Poisson(\lambda)$ $ln\lambda = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$
lm(y ~ x1 + x2, data)	glm(y ~ x1 + x2, data, family=binomial( ))	glm(y ~ x1 + x2, data, family=poisson( ))
1 unit increase in x increases y by α	1 unit increase in x increases log odds by α	1 unit increase in x multiplies y by e^{α}

Porównanie

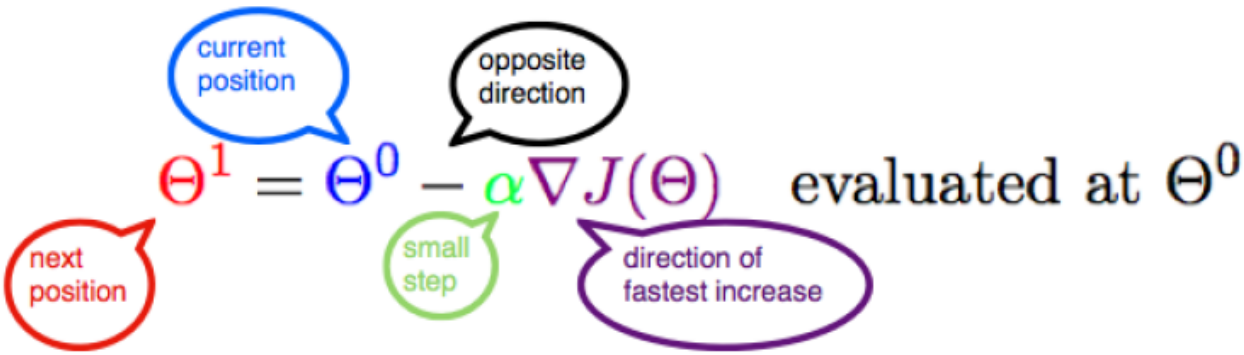

CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Gradient prosty (metoda największego spadku)

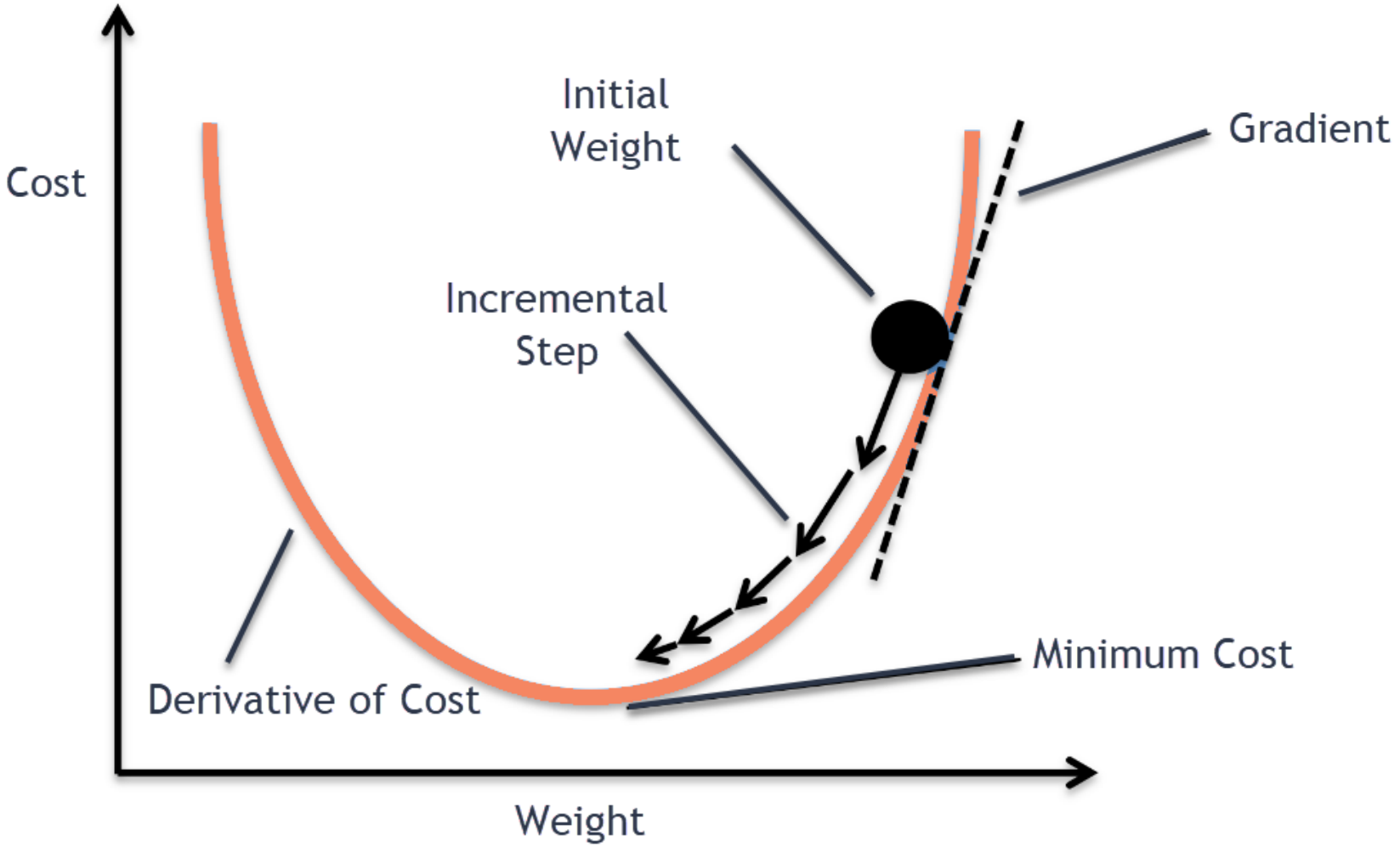
Metoda gradientu prostego (*ang. gradient descent*) jest iteracyjnym algorytmem, o liniowej zbieżności, wyszukiwania minimum zadanej funkcji celu F (ciągła, różniczkowalna i wypukła). Jest ona podstawą działania głębokich sieci neuronowych. W kolejnych iteracjach mamy (punkt startowy $\boldsymbol{x}_0$ wybieramy dowolnie):

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla F(\boldsymbol{x}_k),$$

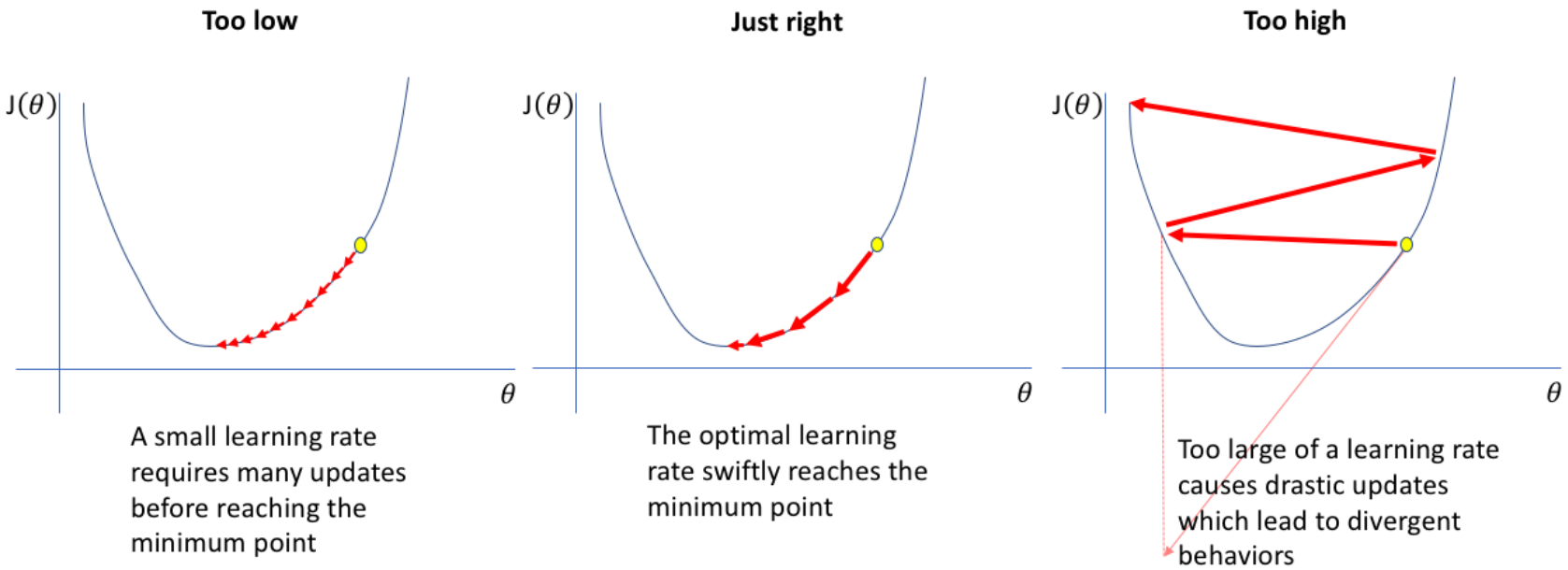
gdzie $\eta > 0$ jest tzw. współczynnikiem uczenia (*ang. learning rate*).



Gradient prosty (metoda największego spadku)



Gradient prosty (metoda największego spadku)



Gradient prosty (metoda największego spadku)

Dla typowej funkcji straty (błąd średniokwadratowy):

$$L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

mamy

$$\frac{\partial L}{\partial \theta_j} = \frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\frac{\partial \hat{y}}{\partial \theta_j}.$$

Zatem do aktualizacji wag będziemy używać całego zbioru danych. Proces jest powtarzany aż do osiągnięcia pewnego kryterium stopu. Taki wariant metody nosi nazwę *batch gradient descent (BGD)*.

Stochastyczny gradient prosty

BGD jest niepraktyczny dla dużych zbiorów danych. Z tego powodu wprowadzono modyfikację polegającą na wykorzystaniu jedynie jednego punktu z próby ($n = 1$) dla którego liczony jest gradient i od razu modyfikowane są wagi. Punkt ten jest wybierany losowo z całego zbioru. Po przeliczeniu dla wszystkich danych proces jest powtarzany dla tych samych punktów, ale w innej (losowej) kolejności. Taki wariant metody nosi nazwę *stochastic gradient descent (SGD)* (online gradient descent).

Mini-batch gradient descent

Niestety metoda ta jest bardzo niestabilna. Zaproponowano więc kolejną modyfikację, która wykorzystuje pewną liczbę obserwacji (zazwyczaj 2^k, $k = 1, 2, \ldots$) i modyfikuje wagi na każdej takiej losowej próbce danych. Podobnie jak poprzednio procedurę powtarzamy dla kolejnych losowych podziałów. Taki wariant metody nosi nazwę *mini-batch gradient descent (MBGD)*.

Mini-batch gradient descent

Batch gradient descent
Mini-batch gradient Descent
Stochastic gradient descent

Gradient prosty dla regresji prostej

$$\hat{y}_i = a + bx_i$$

$$L(a,b) = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - y_i)^2$$

$$\frac{\partial L}{\partial a} = \frac{2}{n}\sum_{i=1}^{n}(a + bx_i - y_i) = \frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial b} = \frac{2}{n}\sum_{i=1}^{n}(a + bx_i - y_i)x_i = \frac{2}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)x_i$$

$$a_{n+1} = a_n - \eta\frac{\partial L}{\partial a}$$

$$b_{n+1} = b_n - \eta\frac{\partial L}{\partial b}$$