Univerzita Pavla Jozefa Šafárika v Košiciach,
Prírodovedecká fakulta

Strojové učenie
Lineárna regresia

Gabriela Andrejková, Ľubomír Antoni

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Úvod

Dôležitým problémom v strojovom učení, je odvodenie funkcionálnych vzťahov (príprava algoritmov na učenie sa) medzi vstupnými atribútovými premennými a s nimi asociovanou odozvou alebo cieľovými premennými.

Tak bude možné predpovedať odozvu pre ľubovoľnú množinu hodnôt vstupných atribútových premenných.

Napríklad, chceme vytvoriť „model“, ktorý bude predpovedať príchod geomagnetickej búrky (GB) o niekoľko hodín na základe niekoľkých meraných parametrov.

https://www.startitup.sk/k-zemi-sa-blizi-obrovska-geomagneticka-burka-zo-slnka-slovensku-ma-priniest-nebeske-divadlo/

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Úvod

Modelom v strojovom učení bude vlastne výsledok programu (nastavenie parametrov vo vybranom type modelu), ktorý rieši daný problém na základe dát.
Tento program nájde funkcionálne vzťahy na základe dát. Taký typ modelu vieme navrhnúť, ak očakávame určité vzťahy medzi vstupmi a výstupmi.

V prípade GB modelom je funkcia, ktorá búrky predpovedá.

Na to je potrebné

- mať k dispozícii namerané údaje atribútov, od ktorých je cieľová premenná závislá a poznať ich vlastnosti.
- predpokladať nejaký vzťah medzi nimi.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Lineárne modelovanie

Najjednoduchším príkladom učiacich sa algoritmov je **lineárne modelovanie**, t. j. učenie sa lineárnych vzťahov medzi vstupnými atribútmi a odozvami.

V tabuľke na ďalšej stránke sú rekordné výsledky v skoku do diaľky u mužov wikipedia: *https://sk.wikipedia.org/wiki/Skok_do_dialky* .

Na obrázku sú tieto výsledky znázornené graficky v paneli A.

Rekordy tvoria rastúcu funkciu. Náš cieľ je použiť tieto údaje na učenie pre model funkcionálnej závislosti (ak existuje) medzi rokmi rekordov a dosiahnutými výsledkami.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Rekordy v skoku do diaľky

Tabuľka: Muži, skok do diaľky

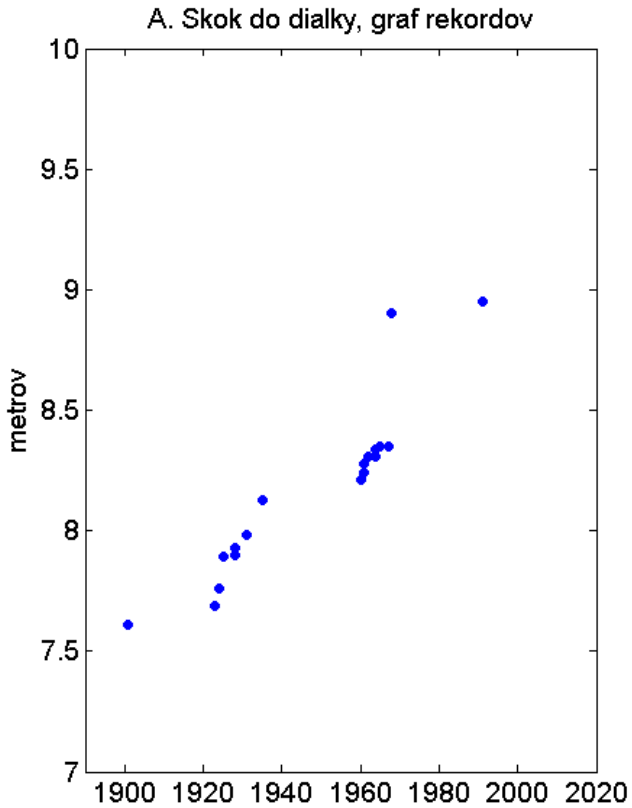
Výkon (m)	Atlét	Miesto	Dátum
7,61	Sp. kráľovstvo Peter O'Connor	Dublin	5. 8. 1901
7,69	USA Edwin Gourdin	Cambridge	23. 7. 1923
7,76	USA Robert LeGendre	Paríž	7. 7. 1924
7,89	USA William DeHart Hubbard	Chicago	13. 6. 1925
7,90	USA Edward Hamm	Cambridge	7. 7. 1928
7,93	Haiti Sylvio Cator	Paríž	9. 9. 1928
7,98	Japonsko Chuhei Nambu	Tokio	27. 10. 1931
8,13	USA Jesse Owens	Ann Arbor	25. 5. 1935
8,21	USA Ralph Boston	Walnut	12. 8. 1960
8,24	USA Ralph Boston	Modesto	27. 5. 1961
8,28	USA Ralph Boston	Moskva	16. 7. 1961
8,31	ZSSR Igor Ter-Ovanesyan	Jerevan	10. 6. 1962
8,31	USA Ralph Boston	Kingston	15. 8. 1964
8,34	USA Ralph Boston	Los Angeles	12. 9. 1964
8,35	USA Ralph Boston	Modesto	29. 5. 1965
8,35	ZSSR Igor Ter-Ovanesyan	Mexiko	19. 10. 1967
8,90	USA Bob Beamon	Mexiko	18. 10. 1968
8,95	USA Mike Powell	Tokio	30. 8. 1991

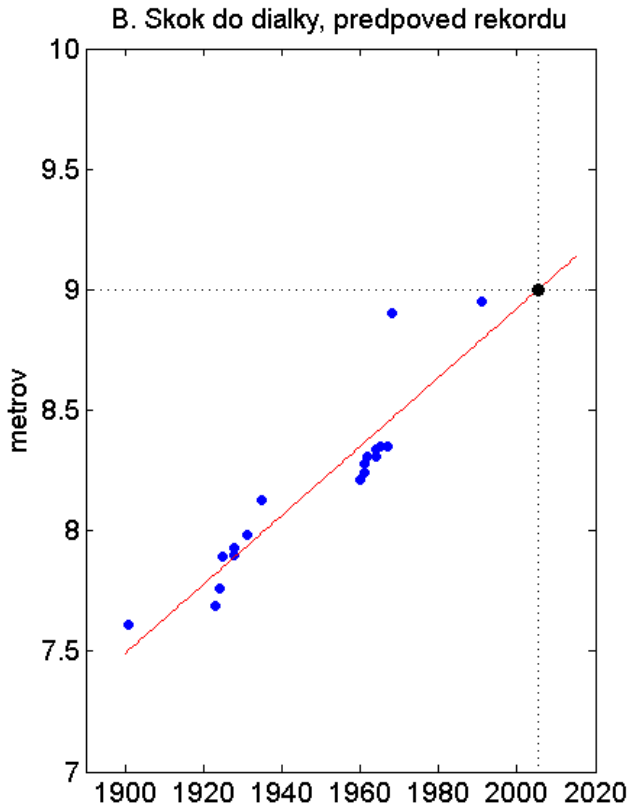
EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Rekordy v skoku do diaľky





EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Rekordy v skoku do diaľky

Uvažujeme o lineárnom modeli.
Náš cieľ je použiť tieto údaje na učenie pre model lineárnej závislosti medzi rokmi rekordov a dosiahnutými výsledkami.

Je nám jasné, že rok nie je jediným faktorom, ktorý ovplyvňuje rekord. Keby sme chceli predpovedať seriózne, mali by sme brať do úvahy aj iné faktory, napríklad formu súťažiacich. Avšak na prvý kontakt s lineárnym modelom nám tieto údaje postačia.

Nakoniec by sme chceli vedieť, v ktorom roku by sme mohli očakávať ďalší rekord.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Definícia modelu

Začneme definíciou nášho modelu ako funkcie, ktorá mapuje naše **vstupné atribúty**, v tomto prípade rok rekordov v skoku do diaľky na **výstupné alebo cieľové hodnoty** - dĺžku skokov.

Je mnoho funkcií, ktoré by sme mohli použiť na toto modelovanie.

Ak označíme x vstupný atribút (rok) a t výstupný atribút (dĺžku skoku), matematicky to budeme zapisovať

$$t = f(x)$$

.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Definícia modelu

Okrem vstupných atribútov funkcia spravidla obsahuje ďalšie parametre, z ktorých niektoré sú modifikovateľné (napríklad učením sa).

Nastavenie parametrov učiaceho sa modelu sú ústrednou témou strojového učenia.

Budeme používať zápis

$$t = f(\overline{x}; \overline{a})$$

pre vyjadrenie toho, že funkcia pracuje na vstupných atribútoch $\overline{x}$ a má parametre $\overline{a}$. Je zrejmé, že aj premenných aj parametrov môže byť viac.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Prvý model

Náš prvý model bude lineárny a bude mať jednu premennú a dva parametre,

$$y = f(x; k, q) = kx + q,$$

kde k, q sú konštanty a x je premenná a z geometrického hľadiska táto funkcia predstavuje priamku.

Aby sme vedeli vybrať model, potrebujeme určiť predpoklady charakterizujúce vzťah medzi x a t.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Prvý model

Pre prvý model (rekordy) sme stanovili, že vzťah bude lineárny. Dáta na obrázku v paneli A by mohli byť modelované pomocou priamky. Teda

$$y = f(x; k, q) = kx + q, \tag{1}$$

K tomu, aby sme pre k a q vybrali najlepšie hodnoty, potrebujeme definovať, čo to znamená **najlepšie**.

Naším cieľom bude, aby model pre všetky vstupné údaje dával výstupy, ktoré sa budú od skutočných výstupných hodnôt líšiť čo najmenej (celková chyba modelu bude čo najmenšia).

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Lineárna regresia

Predpokladajme, že je daná tréningová vzorka (dvojice vstupných a výstupných hodnôt) $\{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\}$.
Chyba modelu pre jednu dvojicu je

$$E_i(t_i; f(x_i; k, q) = (t_i - f(x_i; k, q))^2 \tag{2}$$

Pre všetky dvojice budeme používať kvadratickú chybovú funkciu

$$E = \frac{1}{N} \sum_{i=1}^{N} E_i(t_i; f(x_i; k, q)) \tag{3}$$

Najlepší modelom bude taký, pre ktorý bude chyba (3) najmenšia.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Matematicky to zapíšeme

$$\arg min_{k,q} \frac{1}{N} \sum_{i=1}^{N} E_i(t_i; f(x_i; k, q)), \tag{4}$$

čo znamená *nájsť argumenty, ktoré minimalizujú ....*

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Odvodenie vzťahov pre výpočet k a q

Chyba pre danú tréningovú vzorku $\{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\}$

$$E = \frac{1}{N}\sum_{i=1}^{N} E_i(t_i; f(x_i; k, q)) = \frac{1}{N}\sum_{i=1}^{N}(t_i - (kx_i + q))^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(t_i^2 - 2t_i(kx_n + q) + k^2x_i^2 + 2kqx_i + q^2)$$

$$= \frac{1}{N}\sum_{i=1}^{N}(k^2x_i^2 + 2kx_i(q - t_i) + q^2 - 2t_iq + t_i^2) \tag{5}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Lineárna regresia

K výpočtu minima vypočítame najprv parciálne derivácie podľa k a q

$$\frac{\partial E}{\partial k} = \frac{2k}{N}\sum_{i=1}^{N} x_i^2 + \frac{2}{N}\sum_{i=1}^{N}(x_i(q - t_i)) \tag{6}$$

$$\frac{\partial E}{\partial q} = \frac{2k}{N}\sum_{i=1}^{N} x_i + 2q - \frac{2}{N}\sum_{n=1}^{N} t_i \tag{7}$$

Označme $\overline{x}$ priemer hodnôt $x_1, \ldots, x_n$ a $\overline{t}$ priemer hodnôt $t_1, \ldots, t_n$, t. j.

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \overline{t} = \frac{1}{N}\sum_{i=1}^{N} t_i$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Lineárna regresia

Položíme (7) rovné 0

$$\frac{2k}{N}\sum_{i=1}^{N} x_i + 2q - \frac{2}{N}\sum_{i=1}^{N} t_i = 2k\overline{x} + 2q - 2\overline{t} = 0$$

teda $q = \overline{t} - k\overline{x}$.
Označme $\overline{x^2}$ priemer hodnôt $x_1^2, \ldots, x_n^2$ a $\overline{xt}$ priemer hodnôt $x_1 t_1, \ldots, x_n t_n$, t. j.

$$\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2, \quad \overline{xt} = \frac{1}{N}\sum_{i=1}^{N} x_i t_i$$

Položíme (6) rovné 0

$$\frac{2k}{N}\sum_{i=1}^{N} x_i^2 + \frac{2}{N}\sum_{i=1}^{N}(x_i(q - t_i)) = 2k\overline{x^2} + 2q\overline{x} - 2\overline{xt} = 0$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Dosadením $q = \overline{t} - k\overline{x}$ dostaneme

$$2k\overline{x^2} + 2(\overline{t} - k\overline{x})\overline{x} - 2\overline{xt} = 0 \tag{8}$$

Úpravou dostaneme výrazy pre k a q

$$k = \frac{\overline{xt} - \overline{x}\,\overline{t}}{\overline{x^2} - \overline{x}^2}, \quad q = \overline{t} - k\overline{x} \tag{9}$$

Výpočtom druhej derivácie sa presvedčíme, že ide o minimum.

Pre náš príklad dostávame výsledky uvedené v nasledujúcej tabuľke

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Muži, skok do diaľky a vypočítané hodnoty

Por. č	Rok	Skok	Rok*Skok	Rok*Rok
1	1901	7,61	14466,61	3613801
2	1923	7,69	14787,87	3697929
3	1924	7,76	14930,24	3701776
4	1925	7,89	15188,25	3705625
5	1928	7,90	15231,20	3717184
6	1928	7,93	15289,04	3717184
7	1931	7,98	15409,38	3728761
8	1935	8,13	15731,55	3744225
9	1960	8,21	16091,60	3841600
10	1961	8,24	16158,64	3845521
11	1961	8,28	16237,08	3845521
12	1962	8,31	16304,22	3849444
13	1964	8,31	16320,84	3857296
14	1964	8,34	16379,76	3857296
15	1965	8,35	16407,75	3861225
16	1967	8,35	16424,45	3869089
17	1968	8,90	17515,20	3873024
18	1991	8,95	17819,45	3964081
$(1/N)\sum_{n=1}^{N}$	1947,7	8,1739	15927,40	3793921

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA

Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Hodnoty pre k, q a E sú

$$k = 0.0143; \quad q = -19.7139; \quad E = 0.0193$$

V paneli B na predchádzajúcom obrázku sa nachádza vypočítaná priamka aj nakreslená a vidíme, aký trend mali rekordy v skoku do diaľky.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Predikcia rekordu

Trendovú priamku môžeme použiť na predikciu, aký rekord by sme mohli očakávať v niektorom z nasledujúcich rokov.

Očakávaný rekord
v roku 2018 je $y = 0.0143 * 2018 - 19.7139 = 9,14\ m$.
a v roku 2023 je $y = 0.0143 * 2023 - 19.7139 = 9,21\ m$.

Treba tu poznamenať, že tieto predikcie sú len orientačné. Toto sa prejavuje aj v predchádzajúci rokoch.

Ak príde ďalší rekord priamku bude treba modifikovať.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Polynomiálne modely

Kvadratický model

$$y = f(x; \overline{\mathrm{w}}) = w_0 + w_1 x + w_2 x^2, \tag{10}$$

Model je stále lineárny v parametroch, ale kvadratický v dátach.

Polynomiálny model Všeobecnejší model je lineárny v parametroch, polynomiálny v dátach

$$f(x; \mathrm{w}) = \sum_{k=0}^{K} w_k x^k, x^0 = 1; \tag{11}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Polynomiálne modely

Ak by sme vytvorili vektory a maticu

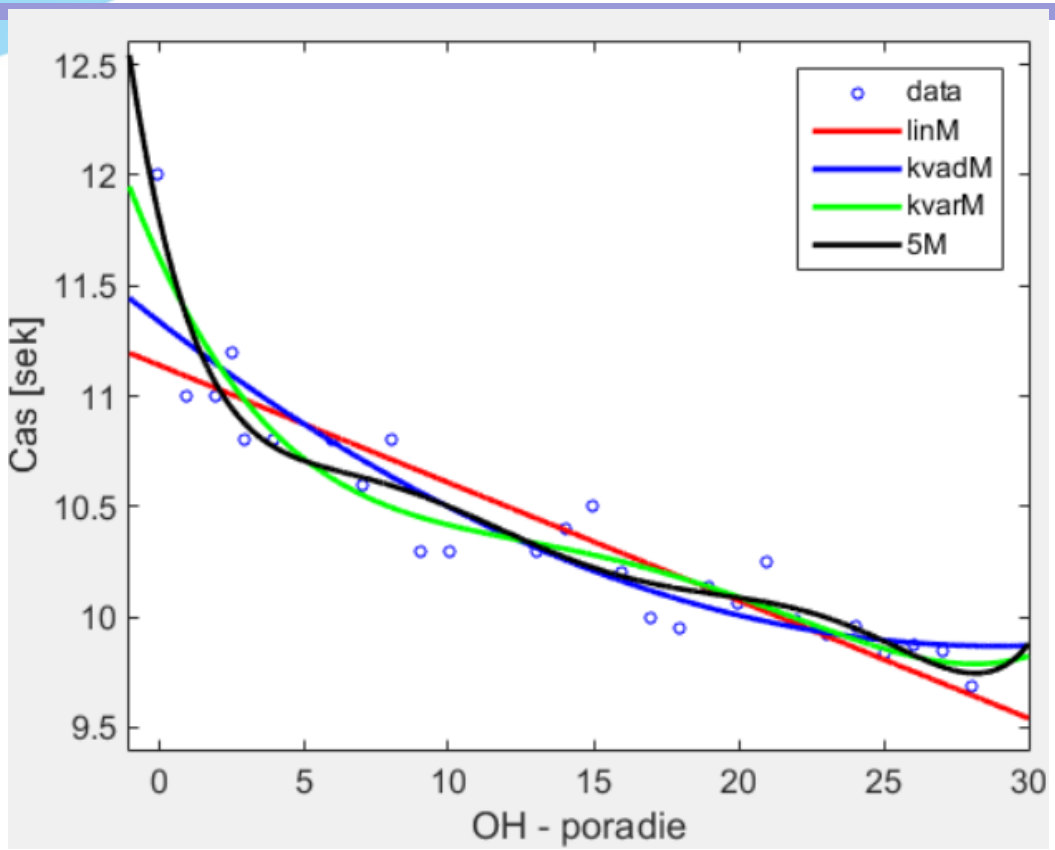
$$\mathrm{w} = (w_0, w_1 \dots w_K),\ \mathrm{x} = (x^0, x^1 \dots x^K),\ \mathrm{t} = (t_1, t_2 \dots t_N)^T$$

$$\mathrm{X} = \begin{pmatrix} x_1^0 & x_1^1 & x_1^2 & \dots & x_1^K \\ x_2^0 & x_2^1 & x_2^2 & \dots & x_2^K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N^0 & x_N^1 & x_N^2 & \dots & x_N^K \end{pmatrix}$$

$$\mathrm{X}\mathrm{w}^T = \mathrm{t}, \quad \mathrm{X}^T\mathrm{X}\mathrm{w}^T = \mathrm{X}^T\mathrm{t}, \quad \mathrm{w}^T = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{t}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Zovšeobecnenie lineárnych



Obr.: Rekordy v behu mužov, 100m, na OH. Modelovanie pomocou rôznych funkcií.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Polynomiálne modely

Matica dát má tvar

$$X = \begin{pmatrix} x_1^0 & x_1^1 & x_1^2 & \dots & x_1^K \\ x_2^0 & x_2^1 & x_2^2 & \dots & x_2^K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N^0 & x_N^1 & x_N^2 & \dots & x_N^K \end{pmatrix}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Polynomiálne modely so zavedenífunkcií

Nie je nutné používať len polynomiálne funkcie, je možné použiť K iných funkcií, $h_k(x)$. Matica dát má potom tvar

$$X = \begin{pmatrix} h_1(x_1) & h_2(x_1) & h_3(x_1) & \dots & h_K(x_1) \\ h_1(x_2) & h_2(x_2) & h_3(x_2) & \dots & h_K(x_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ h_1(x_N) & h_2(x_N) & h_3(x_N) & \dots & h_K(x_N) \end{pmatrix}$$

$$f(x; \mathrm{w}) = \sum_{k=0}^{K} w_k * h_k(x); \tag{12}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Generalizácia a prefitovanie

Naším originálnym cieľom je hľadanie modelu, ktorý bude robiť najlepšie (presnejšie) predikcie. Takým modelom bude ten, ktorý bude **generalizovať** (zovšeobecňovať) a dávať dobré výsledky aj okrem trénovacích dát.

To znamená model by mal minimalizovať chybu aj na dátach, na ktorých nebol trénovaný.

Ak navrhneme modely, ktoré sú zložitejšie, tieto modely sa môžu dostať bližšie k tréningovým dátam. Budú fitovať s menšou chybou, ale ich predikcia sa rapídne zhorší. Sú **prefitované** (over-fittinfg).

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Bias – variance tradeoff

Hľadanie optimálneho modelu, ktorý je schopný generalizovať bez prefitovania, je veľmi ťažká úloha. Tento vzťah sa nazýva **bias-variance tradeoff**.

- Predpokladajme, že poznáme pravdepodobnostné rozdelenie dát, $p(x, t)$, z ktorých trénovacia vzorka bola vybraná.
- Použitím tohto rozdelenia by sme teoreticky mohli vypočítať očakávanú hodnotu kvadratickej odchýlky medzi odhadnutými hodnotami parametrov a správnymi hodnotami.
- Chceme, aby táto hodnota, označená $\overline{\mathcal{M}}$, bola čo najmenšia.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Bias- - variance tradeoff

- Hodnota $\overline{\mathcal{M}}$ môže byť dekomponovaná do dvoch termov nazvaných **bias**, $\mathcal{B}$, a **variancia**, $\mathcal{V}$,

$$\overline{\mathcal{M}} = \mathcal{B}^2 + \mathcal{V}.$$

- Bias predstavuje rozsah, v akom sa priemerná predikcia v rámci všetkých súborov údajov líši od požadovanej regresnej funkcie.
- Variancia meria rozsah, v akom sa riešenia pre jednotlivé súbory údajov líšia od ich priemeru, a teda meria rozsah, v akom je funkcia daná datasetom citlivá na konkrétny výber súboru údajov.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Bias-variancia tradeoff

- Model, ktorý je príliš jednoduchý, bude mať veľký bias (podtrénovanie – under-fitting).
- Je možné znížiť bias tým, že použijeme zložitejší model.
- Zložitejší model bude mať ale vyššiu varianciu
- Teda hľadanie správneho vyváženia medzi generalizáciou a prefitovaním/podfitovaním (over/under-fitting) znamená aj hľadanie správneho vyváženia medzi biasom a varianciou.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

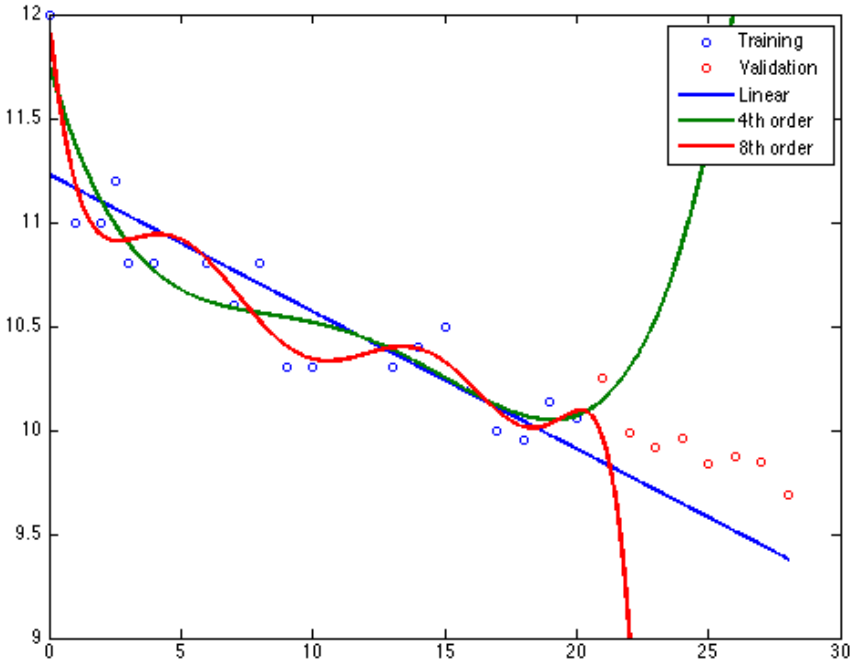
Validácia modelu

Jeden spôsob, ako sa vyhnúť tomuto problému, je použitie druhého datasetu, ktorý sa nazýva tiež **validačný** (potvrdzujúci). Je tak nazvaný preto, lebo sa používa na potvrdenie prediktívnej schopnosti modelu.

Validačné dáta možno pripraviť separovane alebo výberom z tréningovej vzorky.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Príklad:



Obr.: Rekordy v behu mužov, 100m, OH. Modelovanie pomocou rôznych funkcií

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje

Validácia – príklad, beh mužov, OH

Model order: 1, Validation loss: 0.101298
Model order: 2, Validation loss: 0.167631
Model order: 3, Validation loss: 1.06188
Model order: 4, Validation loss: 4.45706
Model order: 5, Validation loss: 5.5142
Model order: 8, Validation loss: 6023.29

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Krížová validácia

Je to technika, ktorá nám umožňuje efektívnejšie využitie daných dát. **K-zložková validácia** rozdelí dáta do K rovnakých podmnožín. Každá podmnožina sa stane raz validačnou množinou.

Spriemernenie K chybových hodnôt dáva výslednú chybovú hodnotu.

Špeciálny prípad, K=N: Leave One Out Cross Validation - LOOCV

$$\mathcal{L}^{CV} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \hat{\mathrm{w}}_{-n}^T \mathrm{x}_n)^2$$

kde $\hat{\mathrm{w}}_{-n}$ je odhad parametrov bez n-tého príkladu.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Regularizácia

Uvažujme triviálny model

$$f(x; \mathrm{w}) = \mathrm{w}^T \mathrm{X}, \text{ kde } \mathrm{w} = [0, 0, \ldots 0]^T.$$

Model predikuje vždy 0, ale je to najjednoduchší model.
Ak sa zmení niektorá hodnota vo w, model bude zložitejší. Uvažujme

$$f(x; \mathrm{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$$

Ak $w_0 \neq 0$ a ostatné parametre budú 0, model bude predpovedať konštantu w_0.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Regularizácia

$$f(x; \mathrm{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$$

Pridaním ďalších parametrov bude model zložitejší. Mohli by sme uvažovať súčet absolutných hodnôt parametrov.

Čím je súčet vyšší a má viac parametrov, tým je model zložitejší ...

Budeme predpokladať, že vyššia hodnota súčtov absolutných hodnôt váh znamená zložitejší model (nechceme kladné hodnoty kompenzovať zápornými).

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Regularizácia

$$f(x; \mathrm{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$$

Od absolutných hodnôt prejdeme ku kvadratickým hodnotám

$$\sum_i w_i^2 = \mathrm{w}\mathrm{w}^T$$

Ak nechceme mať model príliš zložitý, má zmysel mať túto hodnotu malú.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Regularizácia

Budeme minimalizovať regularizovanú chybu, ktorá je súčtom predchádzajúcej chyby a výrazu penalizujúceho zložitosť modelu

$$\mathcal{L}' = \mathcal{L} + \lambda \mathrm{w}\mathrm{w}^T$$

Parameter λ riadi vzťah (trade-off) medzi penalizáciou zle fitovaných dát ($\mathcal{L}$) a penalizáciou zložitosti modelov.Platí

$$\mathcal{L} = \frac{1}{N}(\mathrm{t} - \mathrm{X}\mathrm{w})(\mathrm{t} - \mathrm{X}\mathrm{w})^T$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Regularizácia

Parciálne derivácie $\mathcal{L}'$ podľa **w** nám dajú

$$\frac{\partial \mathcal{L}'}{\partial \mathrm{w}} = \frac{2}{N}\mathrm{X}^T\mathrm{X}\mathrm{w} - \frac{2}{N}\mathrm{X}^T\mathrm{t} + 2\lambda\mathrm{w}$$

Ak položíme výraz rovný 0, dostaneme

$$\frac{2}{N}\mathrm{X}^T\mathrm{X}\mathrm{w} - \frac{2}{N}\mathrm{X}^T\mathrm{t} + 2\lambda\mathrm{w} = 0$$

$$(\mathrm{X}^T\mathrm{X} + N\lambda\mathrm{I})\mathrm{w} = \mathrm{X}^T\mathrm{t}$$

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Regularizácia

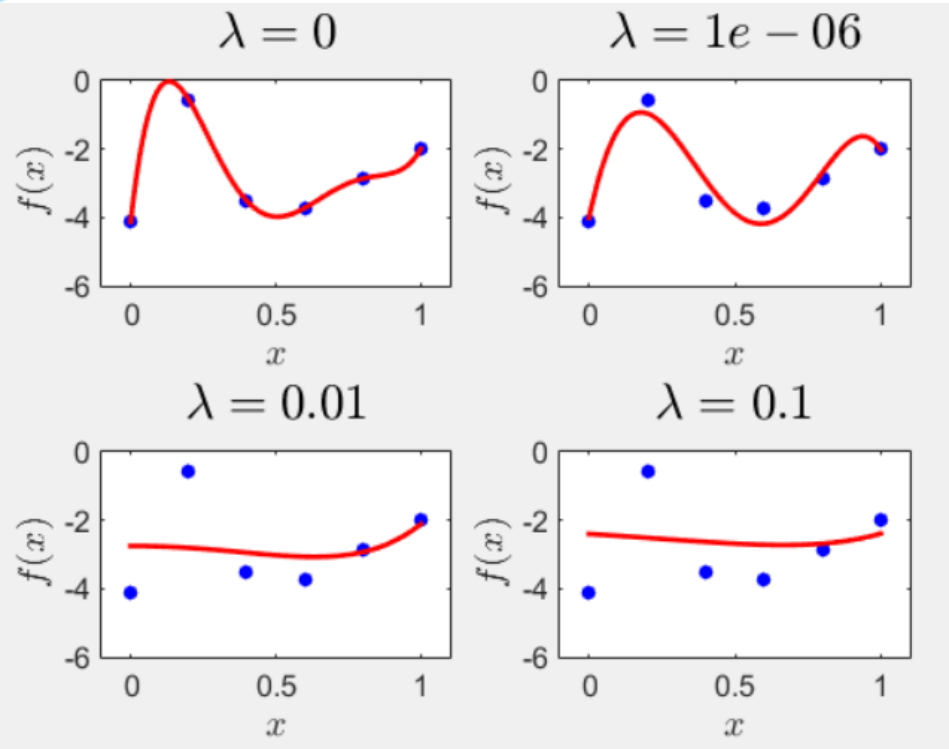
Dostávame regularizovanú hodnotu chyby najmenších štvorcov

$$\hat{w} = (XX^T + N\lambda I)^{-1} X^T t \tag{13}$$

Ak λ sa rovná 0, dostaneme originálne nastavenie parametrov. Inak sú parametre upravené.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Regularizácia



Obr.: Dôsledky meniaceho sa parametra λ pre polynóm 5. stupňa.

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk

Úlohy:

1. Na predchádzajúcich dátach (100 m muži) vyskúšajme použitie parametricky lineárneho modelu s funkciami:

$$h_1(x) = 1;\ h_2(x) = x;\ h_3(x) = \sin(\frac{x-a}{b})$$

   Zvoľte rôzne a a b.
2. Analyzujte skoky do diaľky u žien.
3. Ak je daný lineárny model pre n meraní, ako ho upraviť na $n+1$ meraní?

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY
UMB UNIVERZITA MATEJA BELA
Tento projekt sa realizuje vďaka podpore z Európskeho sociálneho fondu a Európskeho fondu regionálneho rozvoja v rámci Operačného programu Ľudské zdroje
www.minedu.sk www.employment.gov.sk/sk/esf/ www.itakademia.sk