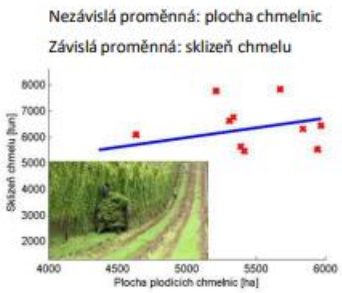
10. Regrese, analytické řešení metodou nejmenších čtverců, numerické řešení metodou největšího spádu

Regrese

= Matematická metoda umožňující odhadovat hodnotu náhodné veličiny (takzvané **závislé proměnné**) na základě znalosti jiných veličin (**nezávislých proměnných**)

Analytické řešení metodou nejmenších čtverců



Typy – Lineární, polynomická, exponenciální, logaritmická

Vícenásobná regrese

• Závislá veličina může záviset na více než jedné nezávislé veličině

• Vícenásobná regrese může být opět lineární i nelineární:

Lineární regrese

Jak funguje LR?

- Výstupní hodnota (nezávislá proměnná) se predikuje na základě znalosti vstupní hodnoty a **<u>regresního modelu</u>**

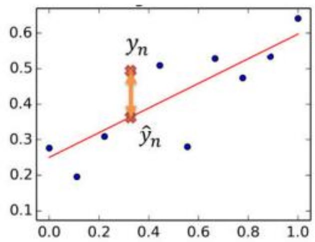
Čím je daný regresní model?

- Lineární funkcí ve tvaru
  $\hat{y} = \theta_0 + \theta_1 x_1$
  $\hat{y}$ … predikovaná veličina, závislá proměnná
  x_1 … nezávislá proměnná
- Parametry modelu představují hodnoty θ_0 a θ_1
  - Parametry modelu lze zapsat jako vektor $\boldsymbol{\theta}$ s hodnotami θ_0 a θ_1

• Regresní model není dokonalý, je zatížen chybou

Chyba modelu lze vyjádřit jako pro danou množinu N vzorků dat (množinu N bodů xi , yi) jako součet kvadrátů odchylek skutečných hodnot yi od predikovaných hodnot y (se stříškou) :

$$chyba\ modelu = J(\boldsymbol{\theta}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$



LR: Učení = hledání parametrů modelu

• Jedná se o úlohu učení s učitelem

• Model se určuje (trénuje)

1) Předem ve fází učení (trénování)

2) A na základě označkovaných dat

• Pro tato data známe hodnoty závislé i nezávislé proměnné (známe x i y)

• Máme k dispozici N dvojic x a y

• V rámci učení se **minimalizuje chyba modelu**

• Hledají se hodnoty **$\theta 0$ a $\theta 1$** tak, aby byla minimalizována chyba modelu

LR: Učení – minimalizace chyby modelu

Chyba modelu představuje **kriteriální funkci**

• Minimalizovat tuto funkci vzhledem k parametrům modelu znamená najít takové hodnoty parametrů, aby výsledná chyba byla minimální

• Kriteriální funkce je **kvadratická**

• Najít minimum této funkce znamená vyjádřit její **první derivaci** vzhledem k hledaným parametrům a položit ji **rovnu nule**

• Při hledání minima kvadratické funkce dochází k odhadu parametrů modelu metodou **„Nejmenších čtverců“ (LSE – least Squares Estimation)**

LR: Učení – LSE pro $\theta 0$

$$\boldsymbol{\theta_0 = \bar{y} - \theta_1 \bar{x} = mean(Y) - \theta_1 mean(X)}$$

LR: Učení – LSE pro $\theta 1$

$$\theta_1 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\boldsymbol{cov(X,Y)}}{\boldsymbol{var(X)}}$$

Vícenásobná LR

Jak funguje?

- Výstupní hodnota (nezávislá proměnná) se predikuje na základě znalosti vstupní hodnoty a **regresního modelu**

Čím je daný regresní model?

- Lineární funkcí ve tvaru

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_D x_D$$

$\hat{y}$ ... predikovaná veličina, závislá proměnná

$x_1 \dots x_D$ ... celkem D nezávislých proměnných

- Parametry modelu představují hodnoty θ_0 až θ_D
  - Parametry modelu lze zapsat jako vektor $\boldsymbol{\theta}$ s hodnotami θ_0 až θ_D

Vícenásobná LR: Učení

- Probíhá jako v případě jednonásobné lineární regrese metodou LSE

$$chyba\ modelu = J(\boldsymbol{\theta}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- Pro odvození je vhodné vyjádřit chybu modelu maticově:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{D1} \\ 1 & x_{12} & x_{22} & \dots & x_{D2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & x_{2N} & \dots & x_{DN} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_D \end{bmatrix} = \boldsymbol{y} - \widetilde{\boldsymbol{X}}\boldsymbol{\theta}$$

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - \widetilde{\boldsymbol{X}}\boldsymbol{\theta})^T (\boldsymbol{y} - \widetilde{\boldsymbol{X}}\boldsymbol{\theta})$$

- POZN.: $\widetilde{\boldsymbol{X}}$ je matice $\boldsymbol{X}$ rozšířená o jeden sloupec s hodnotami 1

$$\boldsymbol{\theta} = \left(\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{X}}\right)^{-1} \widetilde{\boldsymbol{X}}^T \boldsymbol{y}$$

Tento vztah platí i pro jednonásobnou LR

Lineární regrese: shrnutí

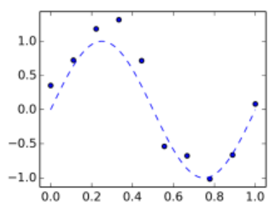
- Nalezení parametrů modelu se řeší principiálně stejně v případě jednonásobné i vícenásobné lineární regrese
- Existuje analytické řešení (vztahy) pro výpočet optimálních parametrů, které minimalizují kvadratickou chybu predikce
  - Jsou odvozené metodou LSE (Least-Squares Estimation) nebo jinými obdobnými metodami
- Nevýhody analytického řešení:
  - Pro vícenásobnou regresi je třeba vypočítat inverzi matice $\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{X}}$
  - To je pro velké N a velké D prakticky nemožné a výpočetně náročné
    - Lze použít jen část dat
    - Pro minimalizaci kriteriální funkce lze použít některou numerickou metodu

Polynomická regrese

• Polynomická regrese umožňuje modelovat nelineární závislost výstupních hodnot na vstupních datech

• Dochází k proložení polynomem k-tého řádu:

$$\hat{y}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots \theta_k x^k$$



• PR lze interpretovat jako speciální případ lineární regrese pro více nezávislých proměnných

• Každá vyšší mocnina x je v případě PR považována za další nezávislou proměnnou

• Regrese je pak polynomiální vzhledem k x ale stále lineární vzhledem k hledaným koeficientům

• Chyba modelu se vyjádří maticově a výsledné řešení je pak stejné

- ⇨ Model reprezentovaný polynomem vysokého řádu dosahuje nulové chyby na trénovacích datech
- ⇨ skutečnou funkci však dobře nereprezentuje
- ⇨ chybovost na nových/neviděných datech je vysoká
- ⇨ Hovoříme o **přetrénování** či špatné generalizaci

• Přetrénování lze zabránit volbou nižšího řádu či **Regularizací**:

Do funkce vyjadřující chybu modelu, se přidá penalizační člen, který je přímo úměrný kvadrátu velikosti jednotlivých koeficientů:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{k=0}^{K} \theta_k^2$$

Nová funkce $J(\boldsymbol{\theta})$ pak vyjadřuje celkové kritérium, které je cílem minimalizovat

Minimalizuje se tak nejen kvadratická odchylka ale i velikost naučených koeficientů – který je menší, pokud jsou menší naučené koeficienty

Čím větší je koeficient λ, tím větší důraz je při učení kladen na zabránění přeučení

Velikost λ se hledá na evaluační sadě dat, podobně jako velikost kroku α

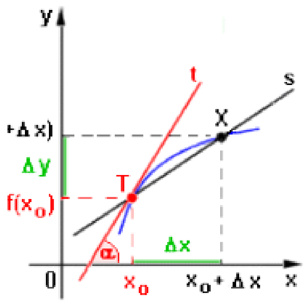
$$\boldsymbol{\theta} = \left(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + \lambda \boldsymbol{I}\right)^{-1} \tilde{\boldsymbol{X}}^T \boldsymbol{y}$$

Numerické řešení metodou největšího spádu

• Umožňuje najít minimum funkce numericky

• Bez znalosti analytického řešení (vzorce)

• V případě funkce jedné proměnné se hledání minima provádí na základě znalosti derivace funkce



Postup:

- Máme funkci $J(x)$
- Chceme najít takový bod x, pro který funkce $J(x)$ nabývá minimální hodnoty:
  1) Zvolíme počáteční hodnotu x
  2) Změníme hodnotu x podle vztahu

     $$x_{t+1} = x_t - \alpha \frac{d}{dx} J(x_t),$$

     kde α je volitelný koeficient a $\frac{d}{dx} J(x_t)$ je derivace funkce v daném bode a v kroku t

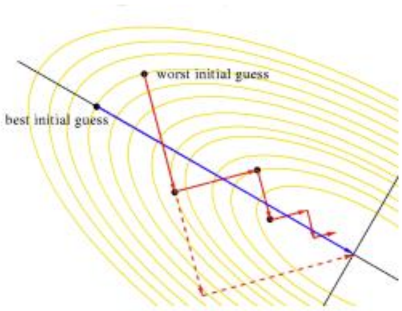
     Posuneme se tedy na ose x doprava nebo doleva (proti směru tečny!) o hodnotu $\alpha \frac{d}{dx} J(x_t)$

     Očekáváme, že funkce $J(x)$ bude v novém bodě x_{t+1} nabývat menší hodnoty než v bodě x_t

  3) Pokud je splněna podmínka konvergence, výpočet ukončíme, v opačném případě pokračujeme podle bodu (2)

Metoda se anglicky nazývá Gradient Descent (GD) nebo také Steepest Gradient Descent (SGD)

• Trpí tzv. **zig-zag** efektem:



Momentum

• K aktuálnímu posunu se přidává cca 90% předchozího

• Pokud se mezi minulým a aktuálním posunem změnil směr (nastala oscilace, zig-zag), je přičtena hodnota s opačným znaménkem a posun je menší = útlum oscilace

• Pokud je nový směr naopak stejný jako předchozí, přičte se hodnota se stejným znaménkem a posun je větší

$$\boldsymbol{v}_{t+1} = -\alpha \frac{d}{dx} J(\boldsymbol{x}_t) + \gamma \boldsymbol{v}_t,\ \gamma \approx 0{,}9$$
$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \boldsymbol{v}_{t+1}$$

PRO LR – minimalizace chyby metodou nej. spádu

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \sum_{n=1}^{\overset{\overline{n=1}}{N}} \tilde{\boldsymbol{x}}_n (\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_n - y_n)$$