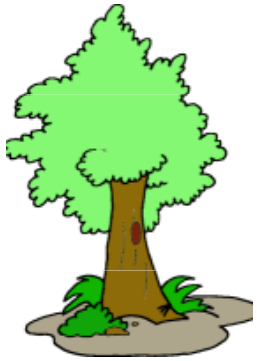
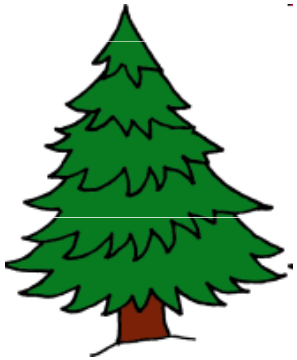
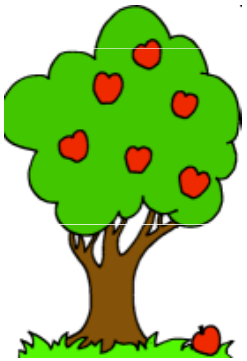
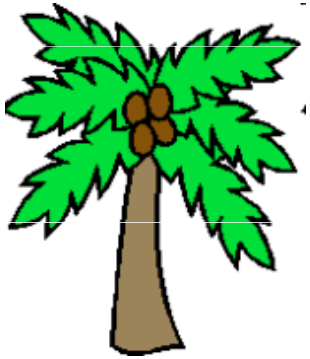
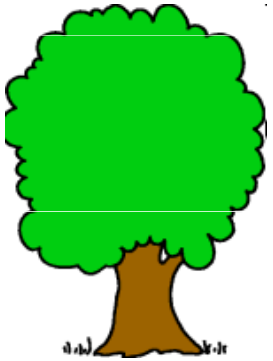










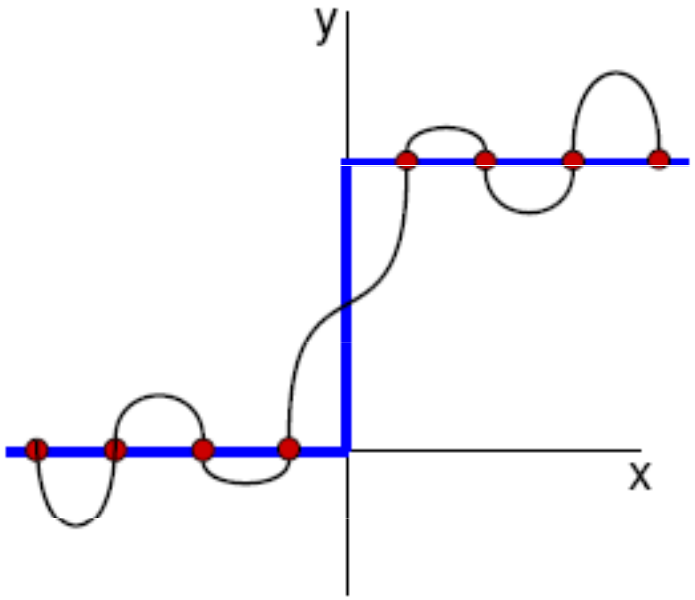
MARS - Multivariate Adaptive Regression Splines

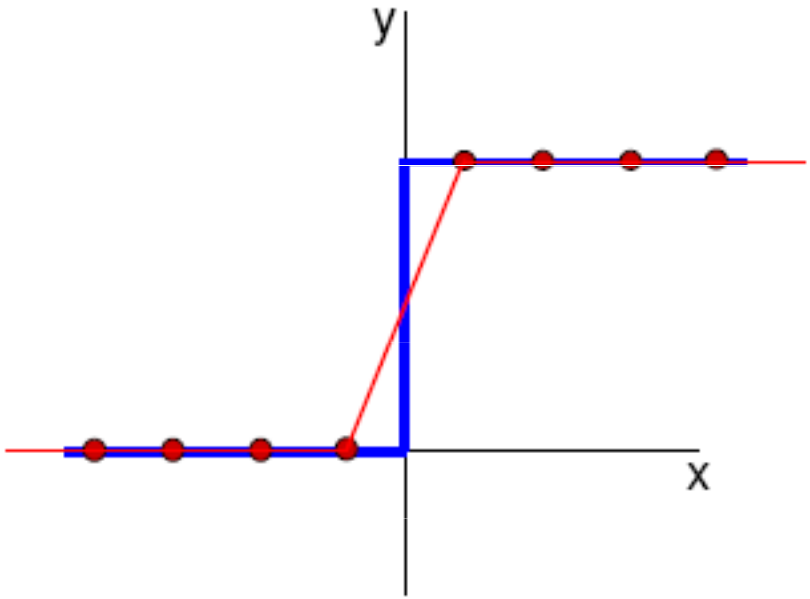
MARS - Multivariate Adaptive Regression Splines

- Friedman (1991)
- technika pro regresní problémy
- na rozhraní mezi stromovou technikou a parametrickou regresí→zobecnění postupné (*stepwise*) lineární regrese
- odstraňuje určité nedostatky binárních regresních stromů, především nespojitosti odhadnutých hodnot závisle proměnné
- prediktory mohou být spojité i kategoriální
- výsledkem metody je regresní rovnice → chybí stromová struktura a interpretace výsledků při velkém počtu proměnných může být obtížnější
- k rozdělení pozorování závisle proměnné se nepoužívá konstanta, ale lineární aproximace

Spline- interpolace

Interpolace - *n* body mohu proložit polynom *(n -1) řádu*
- větší stupeň polynomu - oscilace mezi body
daná množina bodů se aproximuje po částech = ***spline křivky***





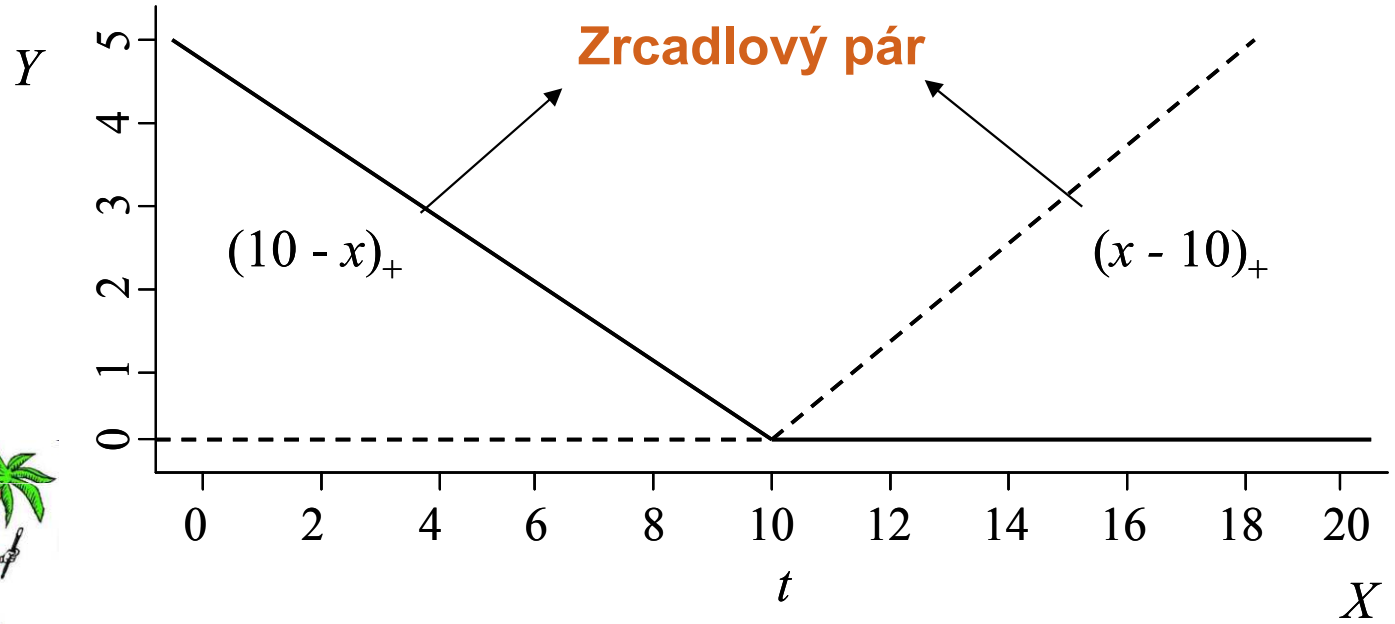
MARS

- spliny - po částech lineárních funkce $(x - t)_+$ a $(t - x)_+$, kde + je kladná část

$$(x - t)_+ = \begin{cases} (x - t), & \text{pokud } x > t \\ 0, & \text{jinak} \end{cases} \qquad (t - x)_+ = \begin{cases} (t - x), & \text{pokud } x < t \\ 0, & \text{jinak} \end{cases}$$

- se svým středem (uzel) v každé hodnotě x_{ij}, pro každý prediktor X_j.

Příklad funkce $(x - 10)_+$ a $(10 - x)_+$ Alternativní zápis: $max(0, x - t)$ a $max(0, t - x)$

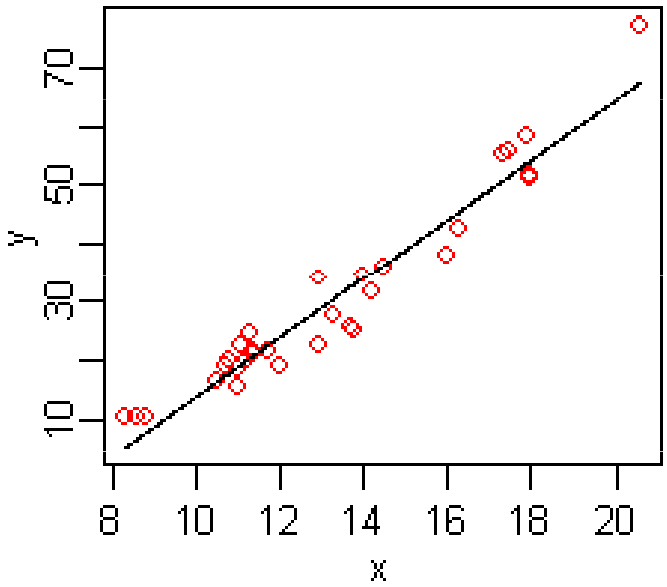


$t = 10$

MARS - příklad

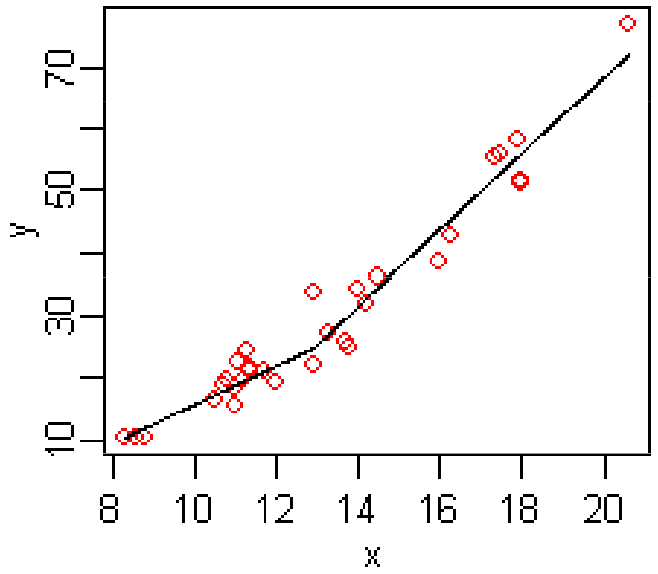
$$\hat{y} = -37 + 5.1x$$

Lineární regrese



$$\hat{y} = 25 + 6.1\max(0, x - 13) - 3.1\max(0, 13 - x)$$

MARS



MARS x lineární regrese

Mějme regresní rovnici:

$$Y = \beta_0 + \sum_{m=1}^{M} \beta_m \left(X_m \right) + \varepsilon$$

- kde Y je závisle proměnná, $X_1,\ldots, X_M$ jsou prediktory
- β_0 je intercept a $\beta_1,\ldots, \beta_M$ regresní koeficienty
- u jednorozměrné lineární regrese je k vyjádření závislosti Y na X použita přímka a koeficienty jsou odhadnuty metodou nejmenších čtverců

MARS

- předpokládejme model s jedním prediktorem a hodnotou uzlu $t = 10$, který můžeme zapsat pomocí dvou regresních rovnic:

$$Y = \beta_0 + \beta_1(X_1) + \varepsilon \qquad \text{pro } x > 10$$

$$Y = \beta_0 + \beta_2(X_1) + \varepsilon \qquad \text{pro } x < 10$$

Rovnice můžeme vyjádřit ve tvaru:

$$Y = b_0 + b_1(X_1 - t)_+ + b_2(t - X_1)_+ + \varepsilon$$

kde $b_0 \equiv \beta_0$, $b_1 \equiv \beta_1$ a $b_2 \equiv \beta_2$

MARS – interakce proměnných

- Stejně jako u lineární regrese lze i u metody MARS použít interakce proměnných
- pro dva prediktory X_1, X_2:

$$Y = b_0 + b_1(X_1 - t_1)_+ + b_2(t_1 - X_1)_+ + b_3(X_1 - t_1)_+(X_2 - t_2)_+ + \varepsilon$$

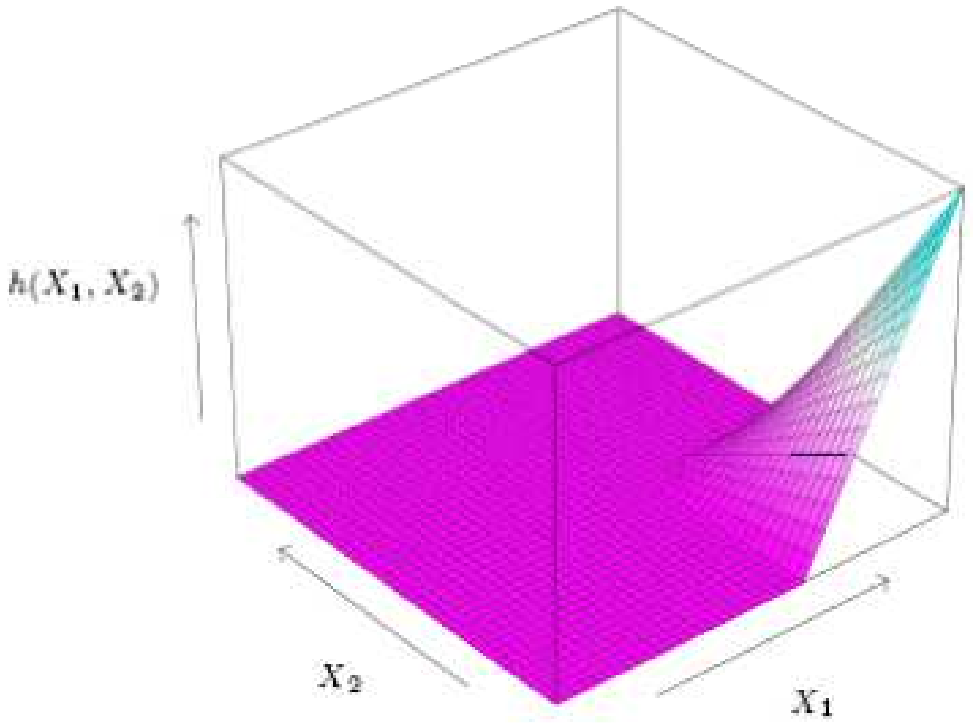
z čehož plyne:

$$Y = b_0 + b_1 X_1 - b_1 t_1 + \varepsilon \qquad pro \quad X_1 > t_1 \; a \quad X_2 < t_2$$

$$Y = b_0 - b_2 X_1 + b_2 t_1 + \varepsilon \qquad pro \quad X_1 < t_1$$

$$Y = b_0 + b_1 X_1 - b_1 t_1 + t_3(X_1 X_2 - t_1 X_1 - t_2 X_1 + t_1 t_2) + \varepsilon \qquad pro \quad X_1 > t_1 \; a \quad X_2 > t_2$$

MARS - interakce

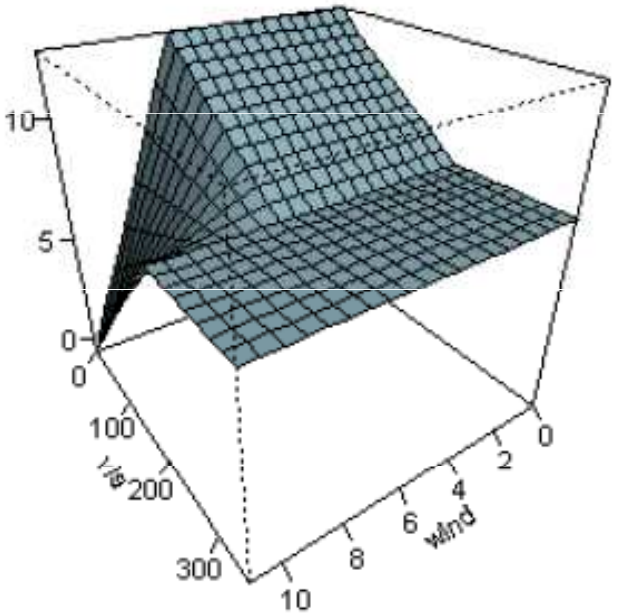


$$h(X_1,X_2) = (X_1\text{-}x_{51})_+ * (x_{72} - X_2)_+$$

(Hastie et. al, 2009)

MARS - příklad

% denní měření koncentrace ozonu, rychlosti větru, teploty vzduchu a intenzita slunečního záření v New Yorku



ozone = **25**
+ 3.1*max(0; temperature-85)
-1.28*max(0; 85-temperature)
-4.9*max(0; 13-wind)
- 0.09*max(0; radiation-139)
-0.049*max(0; radiation-112)*max(0; 13.21-wind)

MARS

Regresní funkci pro MARS můžeme tedy vyjádřit jako:

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

- kde h_m jsou bázové funkce nebo jejich interakce a koeficienty b_m pro dané h_m jsou odhadovány stejně jako u lineární regrese metodou nejmenších čtverců.
- Algoritmus MARS je velmi podobný postupnému dopřednému výběru (*forward stepwise selection*) vysvětlujících proměnných v regresním modelu → namísto proměnných se vybírají lineární splajny.
- Začínáme s nulovým modelem (bez prediktorů).
- Postupně se přidávají jednotlivé členy do rovnice (bázové funkce) → pouze takové, jejichž příspěvek k variabilitě vysvětlené modelem je statisticky významný.
- Tento příspěvek se určuje na základě snížení residuálního součtu čtverců modelu.

MARS- krovalidace

- krosvalidační kritérium *GCV* (*generalized cross-validation)* → vybere se model s optimálním počtem členů v rovnici.
- *GCV* lze použít i pro odhady relativních významností jednotlivých prediktorů.

$$GCV(\lambda) = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{(1 - M(\lambda)/N)^2}$$

kde N je počet pozorování, $\hat{y}_i$ je hodnota závisle proměnné odhadnutá modelem a $M(\lambda)$ je parametr složitosti modelu, který má tvar:

$$M(\lambda) = r + cK$$

kde r je počet nekonstantních bázových funkcí v modelu a K je počet uzlů t v modelu, kde již proběhl výběr parametrů pomocí dopředného výběru

Konstanta c je určena experimentálně:

c = 3 pokud nejsou zahrnuty interakce

c = 2 pro rovnici s interakcemi

MARS - krovalidace

- Datový soubor je rozdělen na testovací a trénovací v poměru zadaném uživatelem (často 70% trénovací a 30% testovací)
- Na trénovacím souboru je vytvořen model a je spočítána jeho přesnost (R^2) na testovacím souboru.
- Hodnota *GCV* je spočítána pro různé podmodely, mající různý počet členů v rovnici, který označuje parametr λ.
- Je vybrán podmodel s nejmenší hodnotou *GCV*.
- Analogie s CART a CHAID → optimální počet terminálních uzlů stromu a PRIM → okna optimální velikosti.

Algoritmus metody MARS

- Krok1: Algoritmus začíná s konstantní funkcí $h_m(X) = 1$

- Krok2: Vytvoří se splajny (zrcadlové páry) se svým středem (uzlem t) v každé hodnotě x_{ij}, pro každý prediktor $X_j \rightarrow$ získáme množinu všech „kandidátských“ bázových funkcí $C \rightarrow$ model je tvořen prvky z této množiny nebo jejich kombinací.

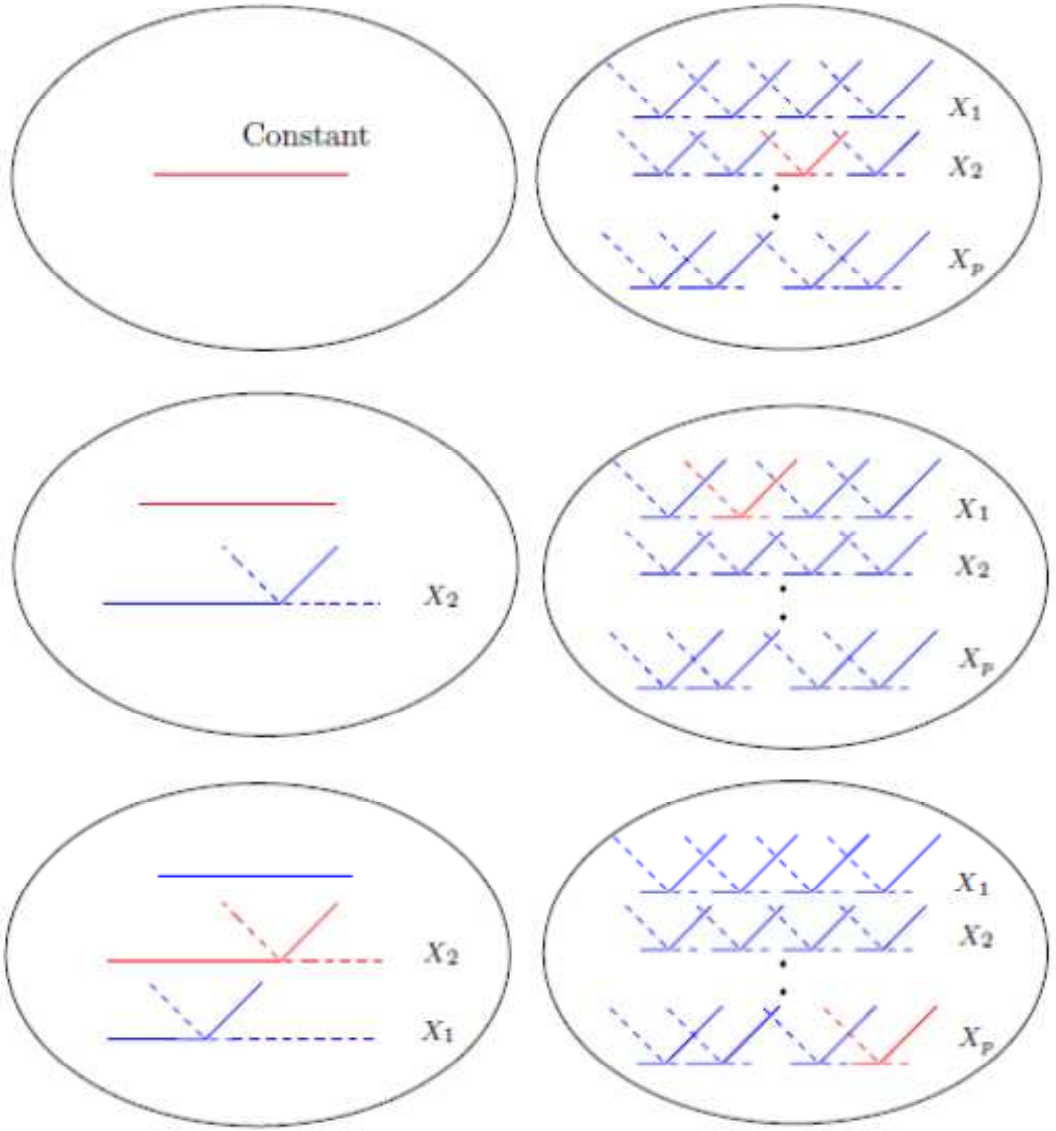
- Krok3: Z množiny C jsou do modelu přidávány pomocí postupného výběru významné bázové funkce, které snižují reziduální chybu modelu.

  !Proces postupuje hierarchicky, významné interakce jsou přidávány do modelu pouze z kombinace bázových funkcí, které již byly do modelu vybrány!

- Z kroku 1 - 3 jsme získali rovnici s vybranými členy → počet členů však bývá většinou velmi velký

- Krok4: procedura zpětného odstraňování.
  - Z rovnice jsou odstraněny ty členy, u kterých po jejich odstranění dojde k nejmenšímu zvýšení chyby modelu.
  - Zpětné odstraňování je učiněno pomocí krosvalidace. Hodnota *GCV* je spočítána pro různé velikosti modelu (s různým počtem členů v rovnici) a je vybrán model, pro který je hodnota *GCV* minimální.

MARS - algoritmus



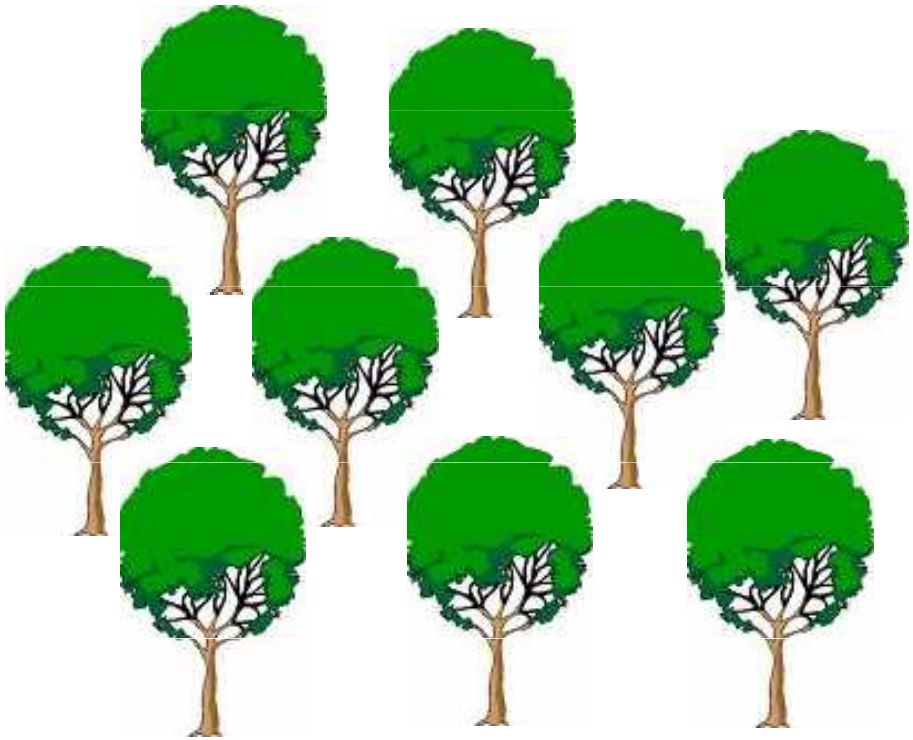
(Hastie et. al, 2009)

MARS

☺ modelovaná plocha je spojitá
☺ zahrnuje aditivitu proměnných
☺ zahrnuje interakci proměnných
☺ vhodná i pro větší počet prediktorů

☹ nevýhodou je méně názorná interpretace → chybí stromová struktura
☹ dopředný výběr proměnných je hierarchický
☹ každý vstup se může v modelu objevit pouze jednou

- PolyMARS (Stone et al., 1997) – pro klasifikaci



Skupinové modely

Klasifikační a regresní lesy

Moudrost davu (*Wisdom of Crowds*)

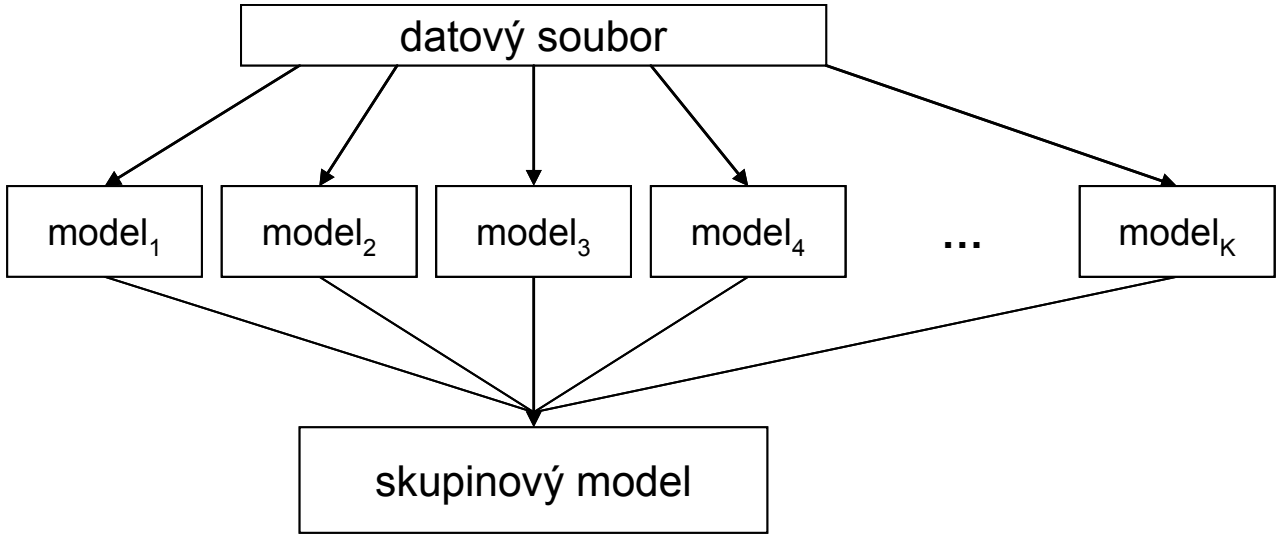
- James Surowiecki, 2004

„skupinový úsudek je daleko inteligentnější a přesnější než úsudek jednotlivce, v případech, kdy jde o hodnocení faktů“

- každý příslušník davu musí činit svůj úsudek na základě vlastních, nezávislých informací
- Výsledek je dán hlasováním

Skupinové modely (*ensamble models*)

- skupině modelů zadáme stejný problém, na kterém se naučí
- výstupy naučených modelů se kombinují
- výsledkem skupinového modelu je
  - v případě regrese → zprůměrování všech výsledků jednotlivých modelů
  - u klasifikace → většinové hlasování jednotlivých modelů (lze však použít průměrování)



Skupinové modely (*ensamble models*)

- Můžeme však kombinací modelů získat přesnější model?

- Podmínka → jednotlivé modely musejí být různé například použitím různých souborů pro učení modelu, které získáme náhodný výběrem z trénovací množiny dat.
- Modely tak budou vykazovat „odlišné“ chyby.
- Přesnost a stabilita těchto modelů se následně ověřuje na testovacích souborech.

- Označení skupinové modely se občas používá také pro kombinaci výsledků z různých modelů (např. neuronových sítí, rozhodovacích stromů a regrese) na stejném souboru.

Čím je způsobena chyba modelu…?

- Př: měříme náhodnou veličinu *Y* v populaci (např. váha člověka) a chceme vyjádřit její reprezentativní hodnotu pro celou populaci.

- Hledáme takový odhad $\hat{y}$, který minimalizuje střední hodnotu chyby $Ey(y-\hat{y})^2$ přes celou populaci.

- V ideálním případě bychom změřili všechny vzorky v populaci (zvážili všechny lidi) a zjistili jejich střední hodnotu E*y*(*y*) (např. průměr, medián), kterou bychom prohlásili za optimální odhad.

- V praxi však tento přístup není možný a pomůžeme si výběrem pouze určité skupiny pozorování z populace, který však musí mít stejné vlastnosti jako celá populace. Takovýto výběr vytvoříme náhodným výběrem.

Skupinové modely -Rozklad chyby

- analogie u modelů, kdy vybíráme pozorování pro trénovací soubor z množiny všech pozorování
- Odchylky pozorovaných od predikovaných hodnot (chybovost modelu) nebudou způsobeny pouze „přírodní“ variabilitou, kterou jsme modelem nevysvětlili, ale také rozdílem ve výsledcích pro různé náhodné výběry a celou populaci.

- Mějme soubor trénovacích dat:
  - $L = (\boldsymbol{y}_i, \boldsymbol{x}_i)$, $i = 1,\ldots,n.$

→ hledáme takovou funkci v prostoru všech prediktorů a hodnot závisle proměnné, aby predikční chyba byla malá.

Skupinové modely -Rozklad chyby

- Pokud mají (*Y*,*X*) stejné rozdělení a daná funkce *R* udává rozdíl mezi pozorovanou hodnotou y_i a predikovanou hodnotou $\hat{y}_i$ závisle proměnné *Y*, pak můžeme predikční chybu (*prediction error*) obecně vyjádřit jako:

$$PE(f, L) = E_{Y,X} R(Y, f(X, L))^2$$

- kde *f*(*X*,*L*) jsou predikované hodnoty $\hat{y}_i$ pro trénovací soubor *L*

Skupinové modely -Rozklad chyby

- Průměrná obecná chyba (*mean-squared generalization error*) na trénovacím souboru *L* je rovna:

$$PE(f,L) = E_{Y,X}\left(Y - f(X,L)\right)^2$$

- Optimální model by měl mít minimální průměrnou chybu pro různé výběry *L* → výsledky modelu pro jednotlivé výběry trénovacích souborů by se neměly příliš lišit.

- Vyjádříme průměr trénovacích souborů stejné velikosti ze stejného rozložení:

$$\bar{f}(x) = E_L f(x,L)$$

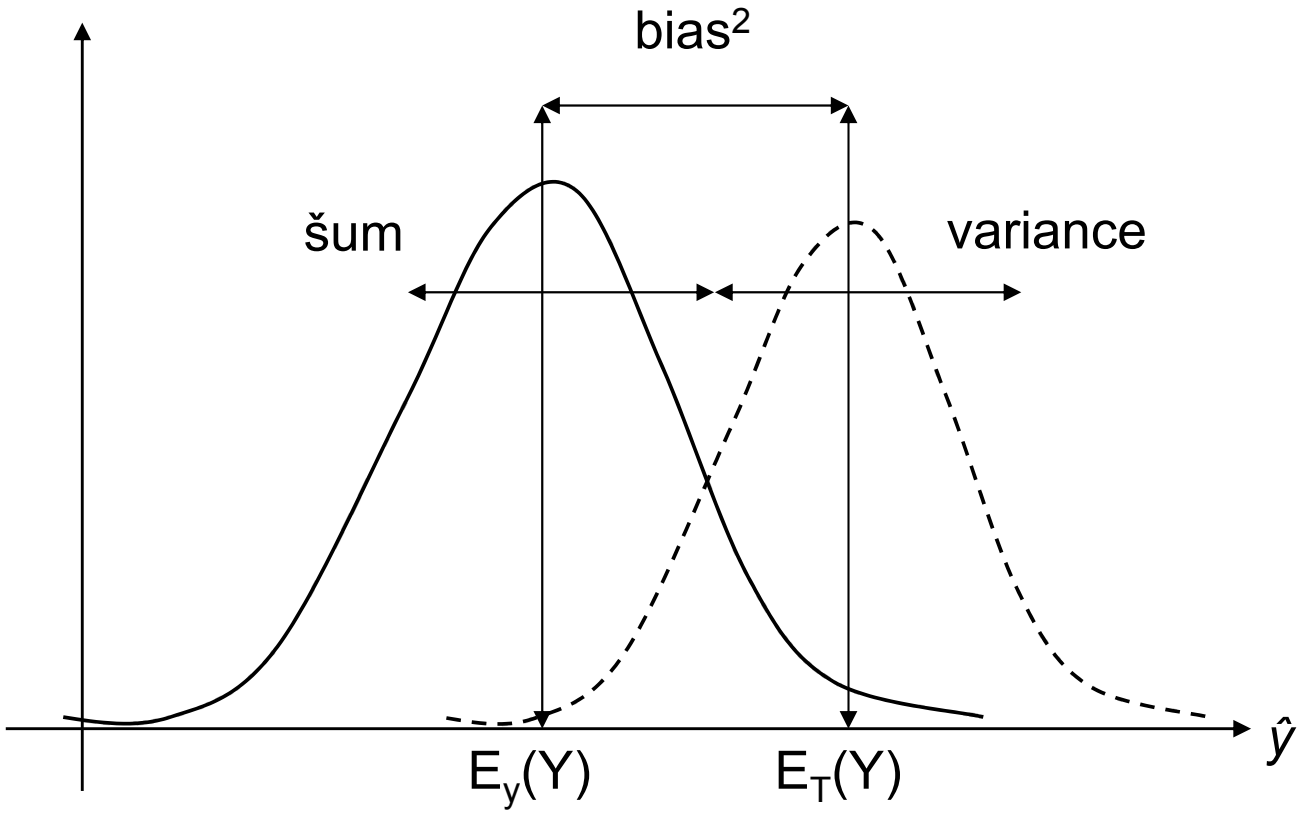
- kde $E_L f(x,L)$ je průměr přes všechny trénovací soubory *L* predikované hodnoty y_i v hodnotě x_i.

Rozklad na systematickou chybu a varianci (*Bias-Variance Decomposition*)

$$PE = \underbrace{E\varepsilon^2}_{\text{šum}} + E_{Y,X}\underbrace{\left(f(X) - E_L f(X,L)\right)^2}_{\text{zkreslení}^2} + E_{X,L}\underbrace{\left(f(X,L) - E_L f(X,L)\right)^2}_{\text{variance}}$$

- **Šum** – je reziduální chyba neboli minimální dosažitelná chyba modelu, kterou nejsme schopni modelem vysvětlit.
- **Zkreslení²**- určuje systematickou chybu modelu. Je to rozdíl optimálního modelu od průměrného modelu.
- **Variance –** je variabilita výsledků jednotlivých výběrů, jinými slovy, jak moc se predikované hodnoty $\hat{y}_i$ liší v rámci trénovacích podsouborů $L \rightarrow$ vysoká variance značí přeučený model.

Rozklad na systematickou chybu a varianci (*Bias-Variance Decomposition*)

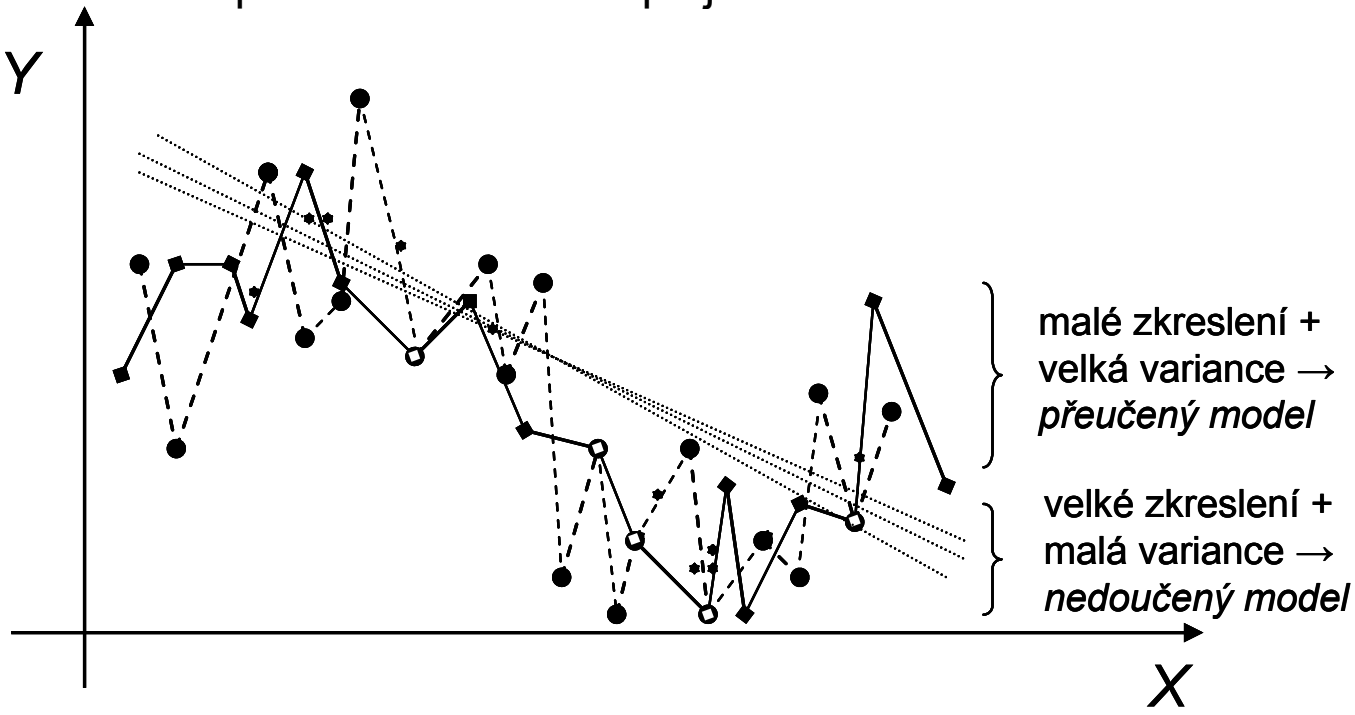


Šum – chyba modelu
Zkreslení²- systematická chyba modelu → optimální x průměrný
Variance –variabilita výsledků jednotlivých výběrů

Slabé modely

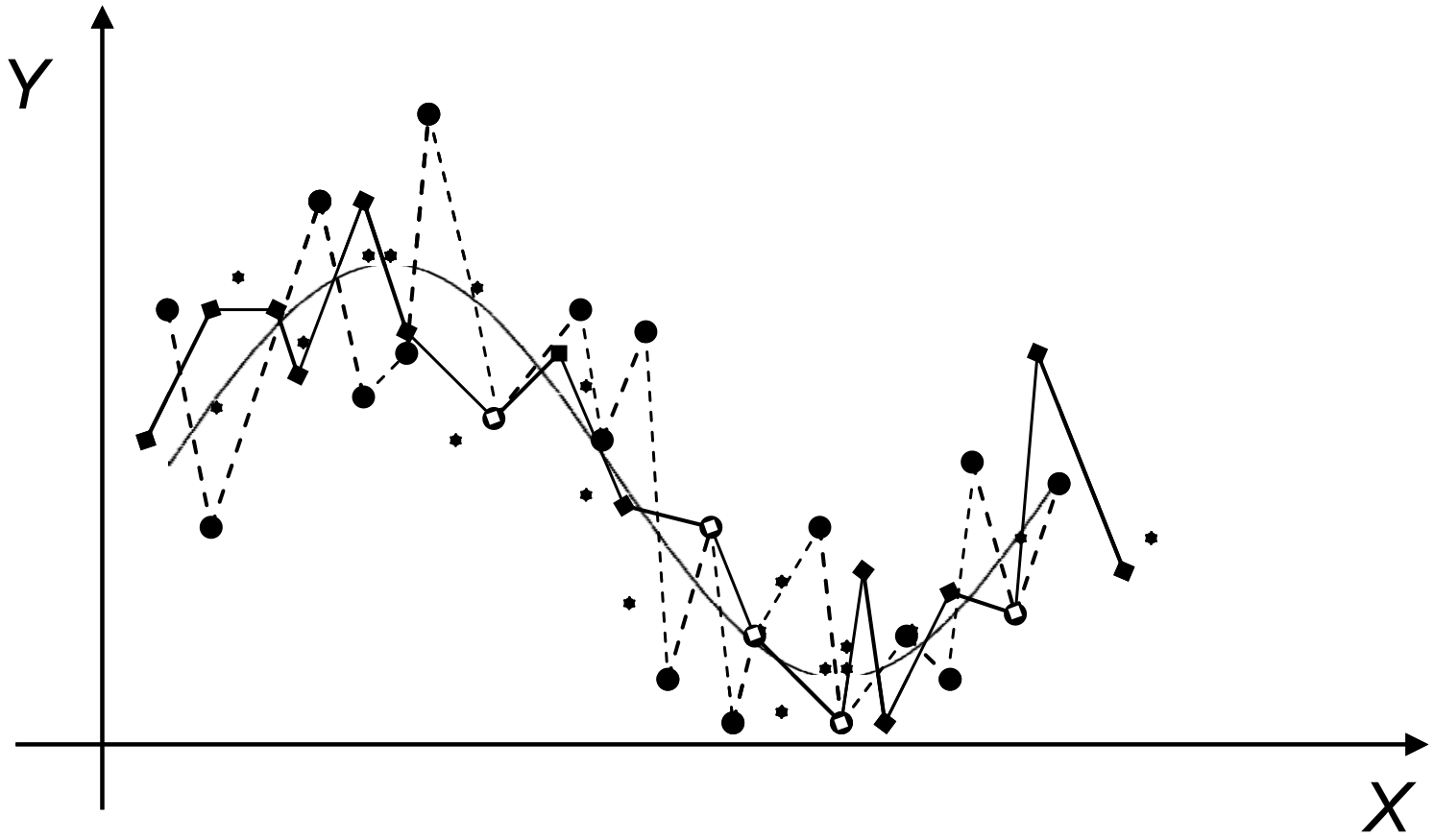
- Modely, které se používají ve skupinových modelech, se označují jako slabé modely neboli *weak learners* (slabý žák, u klasifikace také slabý klasifikátor).

- **Slabý model** je definován obecně jako model, který má malé zkreslení, ale vysokou varianci → mají velmi vysokou přesnost, ale pouze pro pozorování z trénovacího souboru
- Příkladem slabých modelů s velkým zkreslením, ale nízkou variancí může být interpolace bodů pomocí lineárních splajnů



Slabé modely – vytvoření skupinového modelu

- Hledáme tedy model, který by měl nízkou varianci i zkreslení. Kombinováním několika slabých modelů můžeme snížit obě tyto složky.
- Jak na to?



A co na to stromy?

- Rozhodovací stromy jsou dobrými kandidáty pro použití ve skupinových modelech.
- Neprořezané stromy mají totiž vysokou přesnost pro trénovací soubor (tedy nízký bias), ale vysokou varianci (výsledky mezi testovacím a trénovacím souborem se liší).
- Rozhodovací stromy, na které nejsou aplikovány metody pro hledání optimální velikosti stromu, jsou tedy podle výše uvedené definice slabými modely.
- u rozhodovacích stromů jsme pro určení jeho optimální velikosti museli rovněž najít kompromis mezi variancí a zkreslením!