0 Výběr příznaků

Často je výhodné počet příznaků před trénováním modelů snížit. Proces, který zvolí nějakou výhodnou podmnožinu příznaků, je feature selection. Tento proces spadá do části předzpracování dat, konkrétně se jedná o podoblast redukce dimenzionality (dimension reduction).

Výběrem příznaků řešíme hned několik problémů najednou:

- Zahozením nerelevantních a redundantních příznaků můžeme významně zlepšit schopnost generalizace modelu (model není zatížený šumem).
- Pomáhá s prokletím dimenzionality (curse of dim.), kdy vysoká dimenze způsobuje řídkost dat a nerelevantnost sousedství.
- Zlepšuje interpretovatelnost modelu.
- Snižuje výpočetní nároky pro trénování.

0.1 Základní metody výběru příznaků

0.1.1 Filtrační metody

Filtrační metody jsou jednoduché a často nevyžadují náročné trénovaní modelů:

- Vyhodit příznaky s příliš nízkým rozptylem (jsou téměř konstantní).
- Vyhodit příznaky, které mají příliš chybějících hodnot.
- Vyhodit redundantní příznaky, které mají vysokou korelaci s jiným příznakem (jsou v datasetu již zastoupeny).
- Vyhodit příznaky, které mají s cílovou proměnnou nízkou korelaci (je dobré v kombinaci s bázovými funkcemi – samotný příznak totiž nemusí vysvětlovanou proměnnou ovlivňovat pouze lineárně).
- U binárních příznaků rozdělit data na dvě populace a provést hypotézu o rovnosti středních hodnot obou populací (dvouvýběrový t-test).
- Provést test nezávislosti mezi příznakem a vysvětlovanou proměnnou.

0.1.2 Obalové metody

Obalové metody používají pro ohodnocení příznaků pomocný model, který na příslušné kandidátní množině příznaků natrénují a pak výkon porovnají se stejným modelem natrénovaný na jiné sadě příznaků.

Pokud se jako pomocný model použije finální model, je výhodou, že se zvolí tu sadu příznaků, která dobře pracuje s vybraným modelem. V takovém případě

však snadněji dochází k přeučení. Vybere-li se ale jiný model, sada příznaků může být zvolena nevýhodně pro finální model.

Pokud je kandidátních množin příznaků hodně, je tato metody výpočetně náročná, proto se často používají hladové algoritmy:

- Dopředný výběr (forward selection) začíná s prázdnou množinou a postupně přidává příznak, který v dané iteraci nejvíce zvýší výkonnost modelu.
- Zpětný výběr (backward selection) začíná se všemi příznaky a postupně odebírá ty, které nejméně sníží výkonnost modelu.
- Rekurzivní odebírání příznaků (recursive feature elimination) postupně odebírá podle vnitřního ohodnocení příznaků pomocného modelu (u norm. lineární regrese koeficienty, u stromu příslušný informační zisk)

Algoritmy běží dokud nemají požadovaný počet příznaků, případně mohou skončit i s menším počtem příznaků, pokud přidávání, resp. odstranění, příznaků nesnižuje výkonnost.

0.1.3 Vestavěné metody

Vestavěné metody (embedded methods) provedou výběr příznaků natrénováním modelu na celých datech a pak zahodí příznaky, které se naučil vůbec nepoužívat.

U lineární regrese se jedná o příznaky s koeficientem 0, u rozhodovacího stromu ty, které se nikde nepoužily, atd.

0.2 Lasso

Nejpoužívanější vestavěnou metodou je L_1 regularizovaná lineární regrese, která volí množinu příznaků s nenulovým koeficientem.

Na rozdíl od hřebenové regrese penalizuje absolutní hodnotu koeficientů. Pro $\lambda \geq 0$ minimalizuje

$$\mathrm{RSS}_\lambda^{\mathrm{Lasso}} = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 + \lambda \sum_{i=1}^{p} |w_i|$$

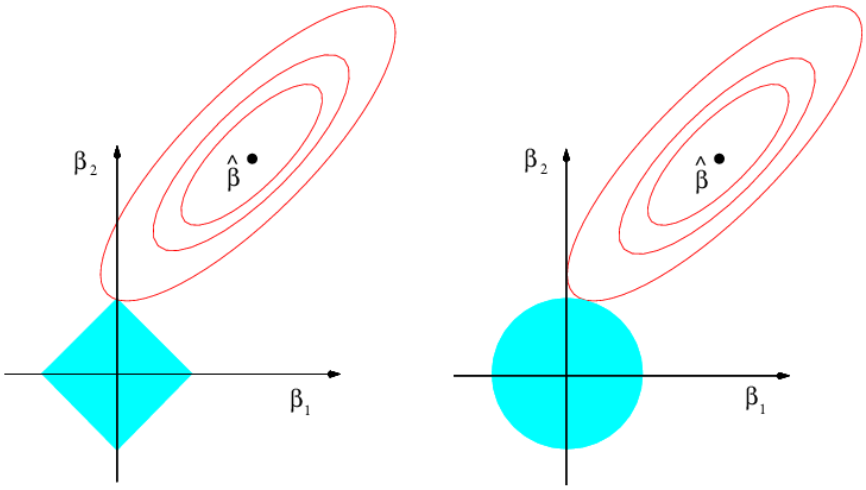
Pro $\lambda = 0$ se jedná o klasickou lineární regresi. Pro $\lambda > 0$ cílí na co nejmenší koeficienty, ale nevadí mu vysoké hodnoty.

$\mathrm{RSS}_\lambda^{\mathrm{Lasso}}$ není diferencovatelný, proto se řešení hledá iterativní metodou:

$$\hat{\boldsymbol{w}}_\lambda^{\mathrm{Lasso}} = \arg\min_{\boldsymbol{w}} \mathrm{RSS}_\lambda^{\mathrm{Lasso}}(\boldsymbol{w})$$

Výhoda modelu Lasso je, že $\hat{\boldsymbol{w}}_\lambda^{\mathrm{Lasso}}$ je řídké – hodně členů je rovno nule. S vyšší λ jsou nuly častější.

Formální důkaz tohoto tvrzení je složitý, ale pro představu lze znázornit obrázkem.



Červeně jsou vykreslené vrstevnice parabolické jámy neregularizované části $\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$. Přímo na nějakých osách bude hyperkrychle vrstevnici protínat celkem často, zatímco hypersféra prakticky nikdy (pokud neleží minimum přimo na ose).

U Lasso může být nežádoucí, že v případě kolinearity má tendenci volit pouze některé z příznaků. To se projevuje jako nevýhoda především u nových dat – příznak může chybět nebo být sám o sobě zatížený nějakým šumem. Proto existuje model, elastic net, který má oba regularizační členy (L_1 i L_2) a kombinuje výhody obou přístupů:

$$\mathrm{RSS}^{\text{Elastic net}}_{\lambda_1,\lambda_2} = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \lambda_1 \sum_{i=1}^{p} |w_i| + \lambda_2 \sum_{i=1}^{p} w_i^2$$

$$\hat{\boldsymbol{w}}^{\text{Elastic net}}_{\lambda_1,\lambda_2} = \arg\min_{\boldsymbol{w}} \mathrm{RSS}^{\text{Elastic net}}_{\lambda_1,\lambda_2}(\boldsymbol{w})$$