Pokročilé neparametrické metody

Validační techniky

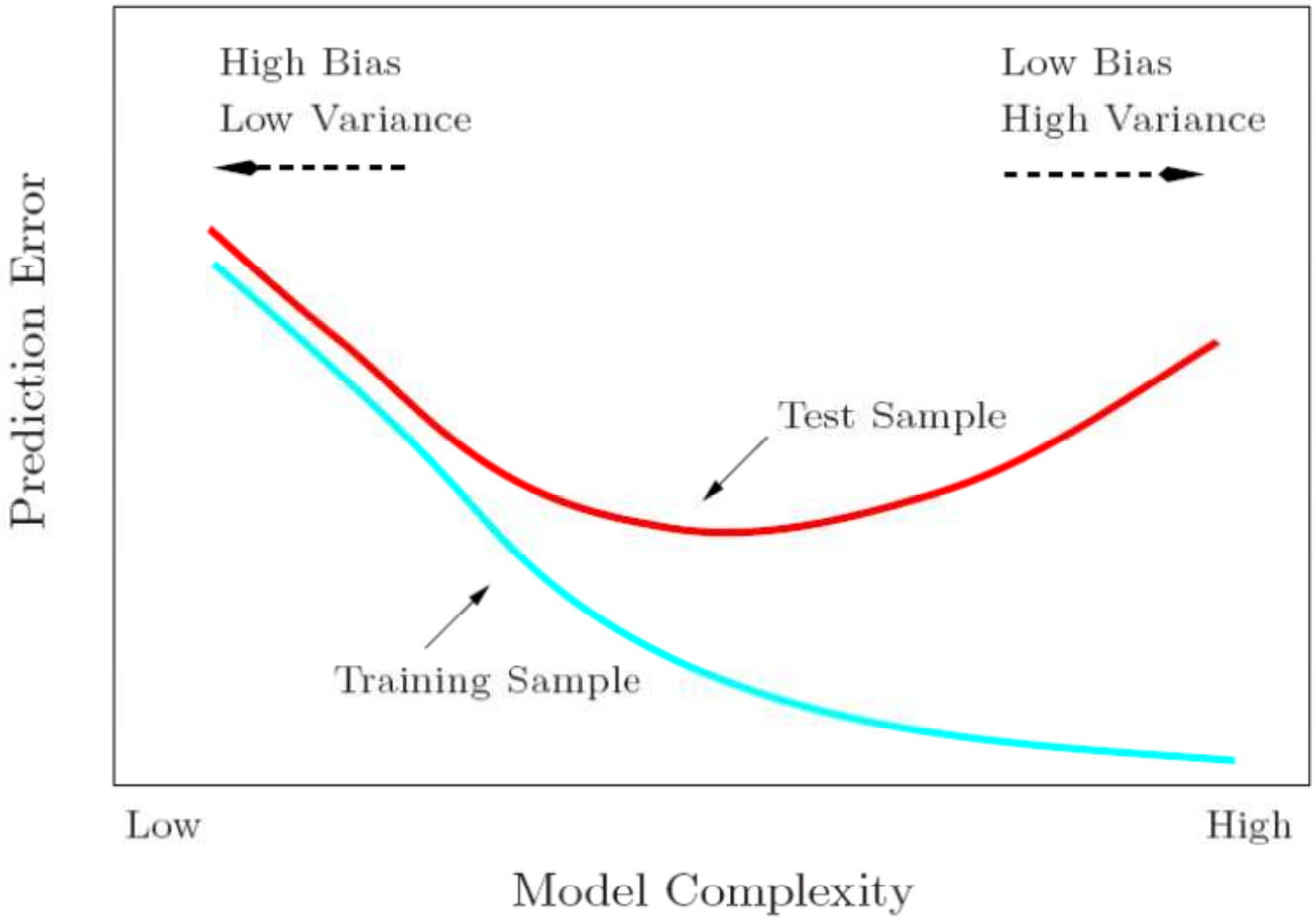
Validace modelů

- k objektivnějšímu a méně zkreslenému odhadu celkové chyby modelu
  - pro výběr mezi různými modely
  - stability modelu
  - jeho obecné platnosti – složitost modelu
  - výběru proměnných do modelu

  !vybrat „nejjednodušší“ model,
  vysvětlující největší množství informace!

- Validační techniky
  - analytické, zahrnující například informační kritéria (AIC, BIC)
  - Založené na opakovaném použití pozorování (krosvalidace, jednoduché rozdělení, bootstrap, jacknifing)

Složitost modelu



Hastie et al., 2009

Analytické validační techniky

- S množstvím proměnných přidávaných do modelu, můžeme sice zvýšit jeho přesnost, ale tím také roste nebezpečí nadhodnocení modelu
- Informační kritéria penalizují počet proměnných v modelu
- Výsledek je kompromisem mezi složitostí modelu a jeho přesností
- Informační kritéria se používají nejčastěji pro parametrickou regresi, kdy se vybírá optimální model z modelů, obsahující různý počet vysvětlujících proměnných; jsou však použitelné i pro neparametrické techniky

Informační kritéria

- AIC - Akaikovo informační kritérium (Akaike, 1974)

$$AIC = 2k - 2\ln(L)$$

- BIC – Bayesovo informační kritérium (Schwarz, 1978) někdy také jako Schwarzovo kritérium (SBC, SBIC)

$$BIC = -2\ln L + k\ln(n)$$

- kde *k* je počet parametrů modelu, *L* je maximální věrohodnostní funkce u GLM (u LM logaritmu residuální sumy čtverců) a *n* počet pozorování
- u BIC je penalizace přidaných proměnných větší než u AIC

Validační techniky II - „*resampling*“ metody

- jednoduché rozdělení, krosvalidace, bootstrap - techniky založeny na opakovaném použití pozorování

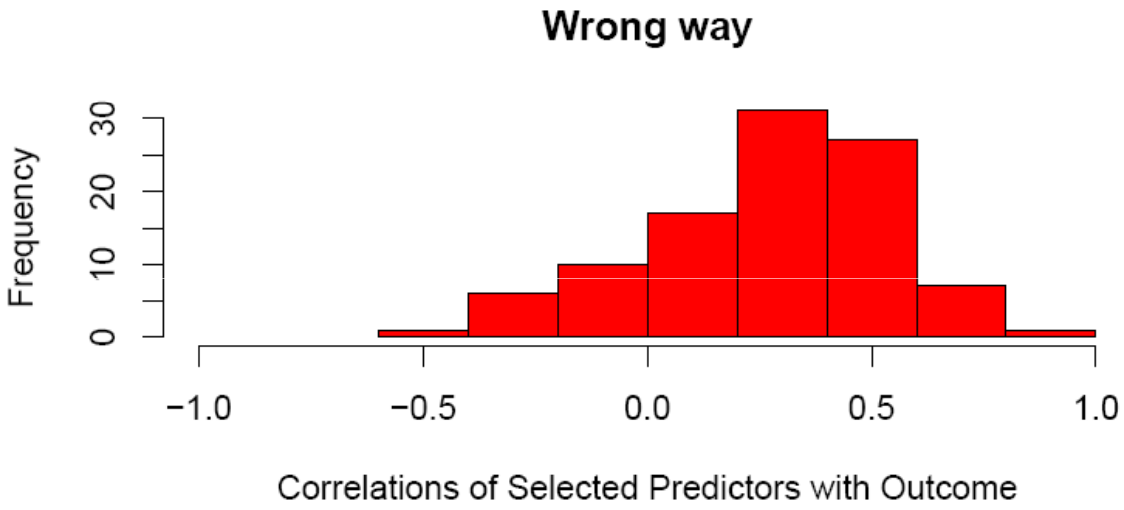
Jednoduché rozdělení (simple spliting)

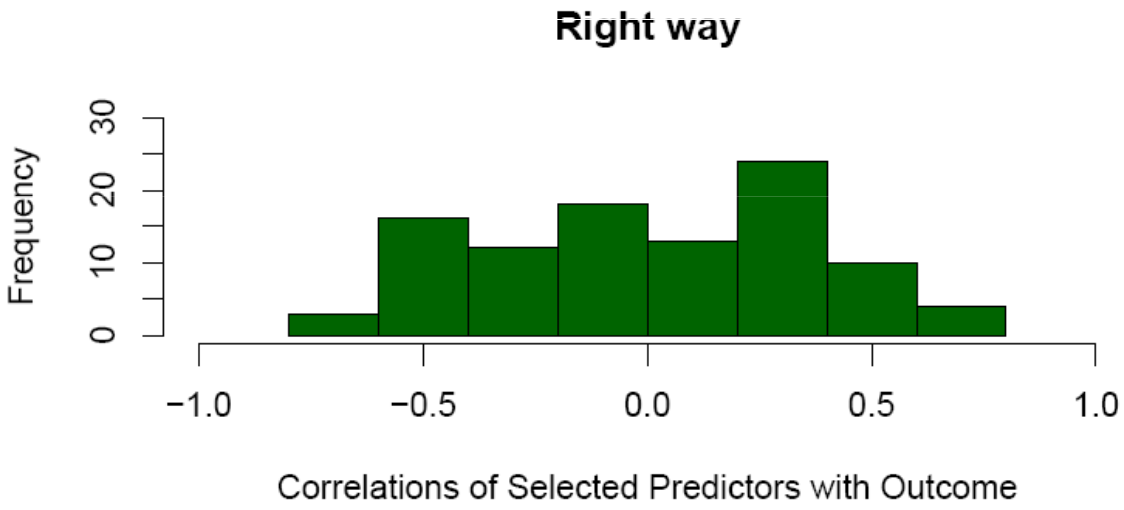
- rozdělení na testovací a trénovací soubor (split-sample,hold-out či simple splitting)
- pouze jeden podsoubor (testovací) je použit k odhadu celkové chyby (generalization error)
- je potřeba větší počet pozorování, aby při dělení nedošlo ke ztrátám informace
- Pokud by se následně vyměnily testovací a trénovací soubor, šlo by již o krosvalidaci pro $k = 2$.

Křížové ověřování - krosvalidace

- pozorování jsou rozdělena do *k* nezávislých podsouborů (obvykle *k*=10)
- jeden podsoubor se vždy použije pro testování (pozorování nejsou použity při tvorbě modelu) a všech ostatních *k*-1 skupin pro tvorbu modelu
- celkem je vytvořeno *k* modelů otestovaných na *k* testovacích souborech
- Z výsledků testovacích souborů můžeme učit stabilitu metody (spočítat např. průměr a směrodatnou odchylku přesnosti na testovacím souboru) a její predikční schopnost
- Výhoda krosvalidace - používáme pro otestování vždy nezávislý datový soubor

krosvalidace





Hastie et al., 2009

Křížové ověřování (krosvalidace)

Rozdělení datového souboru do *k* skupin (zde *k*=5)

testovací	trénovací	trénovací	trénovací	trénovací
trénovací	testovací	trénovací	trénovací	trénovací
trénovací	trénovací	testovací	trénovací	trénovací
trénovací	trénovací	trénovací	testovací	trénovací
trénovací	trénovací	trénovací	trénovací	testovací

Křížové ověřování - krosvalidace

- Pokud se počet krosvalidačních podsouborů rovná počtu pozorování, pak se jedná o "*leave-one-out*" (LOO) krosvalidaci
- LOO krosvalidace byla navržena pro velmi malé datové soubory
- Je vhodná pro odhad obecné chyby v modelu pro spojité funkce, jako je střední kvadratická chyba
- není optimální pro nespojité odhady chyby např. počet chybně zařazených pozorování
- Krosvalidace je velmi často používána k určení optimální velikosti při tvorbě rozhodovacích stromů
- Pro výběr podmnožiny proměnných v lineární regrese má 10-fold a 5-fold krosvalidace lepší výsledky než LOO

Bootstrap

- založen na náhodných výběrech s opakováním z původního výběru
- Soubor se v každém kroku náhodně rozdělí na testovací a trénovací, jako procento z celkového souboru
- Testovací soubory však nejsou nezávislé jako u krosvalidace
- Při každém novém náhodném výběru se vychází vždy ze všech dat
- Vzorky se tedy v jednotlivých testovacích souborech mohou opakovat
- Výhodou je možnost použití i pro menší datové soubory
- V moha případech funguje bootstrap lépe než krosvalidace
- pro rozhodovací stromy, dávají horší výsledky – odhady jsou příliš optimistické

- použití
  - v Random forest a baggingu se používají k tvorbě lesa, k odhadu celkové chyby, v kombinaci s randomizací k odhadu významnosti proměnných
  - u neuronových sítí je bootstrap používán pro výpočet intervalů spolehlivosti jejich výsledků