Křížová validace

Návod ke cvičení

Jan Hrdlička, hrdlij1@fel.cvut.cz

ZS 2010/2011

1 Použití křížové validace pro výběr modelu

Pomocí křížové validace a jejímu odhadu generalizační chyby můžeme vybrat nejvhodnější model pro daná data. Různé modely jsou pro nás v této úloze představovány k-konjunkcí pro různé k. Postup pro výběr modelu je zachycen v m-souboru *cviceni11_1.m* a je následující:

Po vygenerování jsou rozdělena data na trénovací a testovací množinu. Trénovací množina je k dispozici při výběru modelu. Data z testovací množiny budeme mít k dispozici teprve po výběru modelu (představují data při běhu klasifikátoru "naostro").

- Pro všechny možné modely (v našem případě k-konjunkce pro k od 1 do k_{max}) odhadněte generalizační chybu pomocí křížové validace.
- Najděte model s nejmenším odhadem generalizační chyby (v našem případě by to měla být 2-konjunkce, protože podle ní jsou data generována)
- Model s nejmenším odhadem generalizační chyby natrénujte na datech z celé trénovací množiny (tedy najděte konkrétní 2-konjunkci)

Na závěr je možné zjistit nevychýlený odhad chyby na testovacích datech.

2 Porovnání n-fold křížové validace a leave-one-out křížové validace

S leave-one-out křížovou validací i jejím porovnáním s n-fold křížovou validací jste se setkali na přednášce. Cílem tohoto cvičení je ověřit jejich teoretický

rozdíl v odhadu generalizační chyby. Otestujte hypotézu, že leave-one-out křížová validace má větší varianci odhadu generalizační chyby než její n-fold verze a menší bias.

2.1 Úkoly

1. Stáhněte si soubor *cvičení11_2.m*, v něm jsou generována data pro tuto úlohu.

2. Rozdělte data na testovací a trénovací množinu pomocí funkce *testTrainSplit*. Na trénovací množině naučte klasifikátor (rozhodovací strom bez prořezávání) pomocí funkce *treefit* [1]. Na všech datech poté klasifikátor spusťte pomocí funkce *eval* [2] a získejte správnou generalizační chybu e_i (předpokladáme, ze množina všech dat má dostatečnou kardinalitu pro takové tvrzení).

3. Na trénovací množině spusťte n-fold křížovou validaci a leave-one-out křížovou validaci a získejte tak odhady generalizačních chyb $\hat{e}_{cv_i}$ a $\hat{e}_{loo_i}$. K tomu vám pomůže funkce *crossvalidation.m* [3].

4. Body 2 a 3 proveďte stokrát ($j = 100$) a spočtěte následující veličiny:

   Odchylku

   $$Bias_{cv} = \frac{1}{j}\sum_{i=1}^{j}\hat{e}_{cv_i} - e_i \tag{1}$$

   $$Bias_{loo} = \frac{1}{j}\sum_{i=1}^{j}\hat{e}_{loo_i} - e_i \tag{2}$$

   Rozptyl

   $$Var_{cv} = \frac{1}{j-1}\sum_{i=1}^{j}(\hat{e}_{cv_i} - \bar{\hat{e}}_{cv_i})^2 \tag{3}$$

   $$Var_{loo} = \frac{1}{j-1}\sum_{i=1}^{j}(\hat{e}_{cv_i} - \bar{\hat{e}}_{loo_i})^2 \tag{4}$$

   Kde $\bar{\hat{e}}$ je střední hodnota odhadu.

[1] Syntax: *tree = treefit(train.samples,train.labels,'prune','off')*
[2] Syntax: *out = eval(tree,data.samples)*
[3] Syntax: *meanGenError = crossvalidation(dataset,nFold,@learnFcnHndl,@evalFcnHndl)*

Střední kvadratickou odchylku

$$MSE_{cv} = \frac{1}{j}\sum_{i=1}^{j}(e_i - \hat{e}_{cv_i})^2 \tag{5}$$

$$MSE_{loo} = \frac{1}{j}\sum_{i=1}^{j}(e_i - \hat{e}_{loo_i})^2 \tag{6}$$

5. Zjistěte, zda platí $Bias_{cv} > Bias_{loo}$ a $Var_{cv} < Var_{loo}$. Zjistěte, jak se liší Střední kvadratické odchylky leave-one-out a n-fold křížových validací.