

Výběr modelu, ROC křivka

Návod ke cvičení
Jan Hrdlička, hrdlij1@fel.cvut.cz

ZS 2010/2011

1 Úvod

Cílem tohoto cvičení je seznámit se způsobem výběru modelu na základě ROC analýzy. Seznámit se s ROC křivkou a způsobem jejího výpočtu. V závěru je zmíněna plocha pod křivkou, její významem pro výběr modelu a také její pravděpodobnostní interpretace.

1.1 ROC křivka

ROC (Receiver operating characteristic) křivka je založena na kontingenční tabulce správného resp. špatného zařazení pozitivního nebo negativního vzorku:

	Pos examples	Neg examples	
Predicted positive	TP	FP	→ PPV
Predicted negative	FN	TN	→ NPV
	↓ TPR	↓ FPR	

Terminologie:

TP: True positive, zásah

TN: True negative, správné zamítnutí

FP: False positive, falešný alarm, chyba 1. typu

FN: False negative, minutí, chyba 2. typu

TPR: True positive rate, Sensitivita = $TP/(TP+FN)$

FPR: False positive rate, 1-Specifičnost = $FP/(TN+FP)$

PPV: Positive Predictive Value = $TP/(TP+FP)$

NPV: Negative Predictive Value = $TN/(TN+FN)$

Mějme klasifikátor (test, model), který se snaží zařadit příklady do dvou tříd, negativní resp. pozitivní $C \in \{0; 1\}$ a má spojitý výstup Y .

Tento výstup porovnáváme s prahem θ a podle výsledku zařadíme příklad (provedeme rozhodnutí D) na vstupu jako pozitivní nebo negativní.

$$D = \begin{cases} 1 & \text{if } Y > \theta \\ 0 & \text{if } Y \leq \theta \end{cases} \quad (1)$$

Křivka ROC je pak grafickým znázorněním závislosti TPR na FPR pro měnící se práh θ , její příklad je v m-file tohoto cvičení.

1.2 Plocha pod křivkou (Area Under the Curve)

Plocha pod ROC křivkou (AUC) je jedním z kritérií pro výběr vhodného klasifikátoru (modelu) a určuje kvalitu separace tříd v daném modelu resp. schopnost separace tříd daným klasifikátorem. AUC má také pravděpodobnostní interpretaci. Při dvou náhodně vybraných vzorcích, z nichž jeden je pozitivní $C_i = 1$ a druhý negativní $C_j = 0$, je AUC pravděpodobnost, že výstup klasifikátoru Y_i u prvního vzorku je větší než výstup Y_j odpovídající druhému vzorku.

$$AUC = P\{Y_i > Y_j\}; \quad (2)$$

2 Zadání

2.1 Data

V první sekci jsou vygenerována data do dvou tříd, které nejsou plně separabilní. Každá třída má trojrozměrné normální rozdělení určené střední hodnotou. Rozptyl je stejný pro všechny dimenze i obě třídy a je roven jedné. Experimentujte s různými nastaveními parametrů.

2.2 Klasifikátory, ROC křivka a AUC

K dispozici jsou 2 klasifikátory :

- Naivní Bayes předpokládající normální rozdělení dat. Jeho výstup je posteriorní pravděpodobnost příslušnosti k třídě.

- Jednoduchý klasifikátor, který pouze sečte naměřené hodnoty ve všech rozměrech.

Po vybrání trénovací množiny a naučení Naivního Bayese (odhadnutí parametrů pravěpodobnostního rozdělení dat) jsou na základě výstupů z klasifikátorů zkonstruovány ROC křivky a z nich je poté spočtena přibližnou integrací AUC. Který klasifikátor je úspěšnější a proč? Jaký bod na ROC křivce byste zvolili pro konkrétní klasifikaci a proč?

2.3 ROC convex hull

K dispozici jsou 4 klasifikátory, které odhadnou u jednotlivých vzorků příslušnost ke třídě:

- Naivní Bayes předpokládající normální rozdělení dat.
- Rozhodovací strom
- k-Nearest Neighbors algoritmus
- Klasifikátor, rozhodující se na základě Mahalanobisovy vzdálenosti

Každý z klasifikátorů odhadne příslušnost vzorků testovací množiny. Na základě těchto rozhodnutí se spočte TPR a FPR a vynesou se do grafu. ROC convex hull je pak konvexní křivka začínající v bodě $[0\ 0]$, jdoucí přes "úspěšnější" klasifikátory a končící v bodě $[1\ 1]$. Na této křivce jsou klasifikátory, které jsou nejúspěšnější pro různá vychýlení dat (toto vychýlení může být způsobeno rozdílnou apriorní pravděpodobností příslušnosti ke třídám, ale také různými cenami klasifikačních chyb).

2.4 ROC křivka z křížové validace

V další sekci je ROC křivka spočtená z 5-fold křížové validace. Křížová validace je v tomto případě prováděna na trénovací množině z předchozích sekcí. ROC křivka se získá tak, že po provedení křížové validace sloučíme výstupy klasifikátoru všech n testovacích množin (pro n běhů křížové validace) a z nich zkonstruujeme průměrnou ROC křivku. Tento způsob je algoritmicky snazší, ovšem předpokládá znalost konkrétních výstupů klasifikátoru, nejen odhady chyb.

Reference

- [1] *Fawcett T.: An introduction to ROC analysis*
Pattern Recognition Letters, 861-874, 2006