Analýza dat genové exprese z pohledu strojového učení

Michael Anděl

25. dubna 2013

1 Zadání

K dispozici jsou data genové exprese (GE), měřená na 72 pacientech, resp. 72 vzorků krvetvorné tkáně z oblasti (?) kostní dřeně, pro každou tkáň byla současně měřena exprese 7129 genů. Z celkového počtu 72 pacientů bylo 25 postiženo akutní myeloidní leukemií (AML) a 47 postiženo akutní lymfoblastickou leukemií (ALL). Cílem je vytvořit model, který na základě měření exprese v krvetvorné tkáni rozhodne, resp. předpoví které buňky imunitného systému jsou resp. budou napadeny - zda myelocity (AML), nebo lymfocyty (ALL). Data jsou k dispozici v souboru `data.mat`, který je rozdělen na vektor tříd (`data.classes`: '1' - ALL, '2' - AML) a matici GE (`data.exprs`).

Data genové exprese je možno chápat jako matici $\mathbf{X} \in \mathbf{R}^{N \times M}$, kde N je počet pozorování (vzorků, tkání, pacientů) a M je počet atributů (příznaků), tedy měřených genů. Každý i-tý datový bod (vzorek) je tedy možno chápat jako objekt v M-rozměrném prostoru, tedy (řádkový) vektor o M složkách $\mathbf{x}_i \in \mathbf{R}^M$. Vzhledem k malému množství vzorků a nepoměrně velkému rozměru je záhodno data vyjádřit v prostoru o mnohem menším rozměru. Klíčovým problémem soudobé analýzy GE je jak takovouto transformaci *naučit*.

2 Jak na to

Jednou z nejjednodušších variant, jak vyjádřit datový bod $\mathbf{x}_i \in \mathbf{R}^M$ v prostoru o dimenzi K, je jeho ortogonální transformace

$$\mathbf{z}_i \in \mathbf{R}^K = \mathbf{x}_i \mathbf{V}$$

, kde $V \in \mathbf{R}^{M \times K}$ je ortogonální báze prováděné transformace. Projekci *celé* sady dat do redukovan0ho prostoru je možno maticově zapsat:

$$\mathbf{Z} = \mathbf{X}\mathbf{V}$$

, kde $\mathbf{Z} \in \mathbf{R}^{N \times K}$je datová sada v redukovaném prostoru o dimenzi K. Jedním z nejjednodušších způsobů je určit bázi V jako prvních K vlastních vektorů kovariance dat - viz *analýza hlavních komponent* (PCA).

Úkolem bude porovnat přesnost klasifikačního modelu, naučeném na a) GE datech plné dimenze M b) na datech transformovaných do redukovaného prostoru dimenze $K = 50$.

1. Na přiložených datech *plné* dimenze naučte rozhodovací strom. Výsledný klasifikátor vizualizujte a vyčíslete jeho (trénovací).

2. Transformujte data do redukovaného prostoru. Tj. naučte bázi metodou PCA, a data promítněte do redukovaného prostoru (viz výše). Na této projekci dat pak naučte rozhodovací strom, vizualizujte ho a vyčíslete chybu

3. Vyčíslete odhad *skutečné* chyby klasifikace na datech o plné dimenzi. Použijte metodu křížové validace. Tedy v každém cyklu validace rozdělte data na trénovací a testovací, naučte klasifikátor na testovacím oddílu a na trénovacím oddílu odhadněte jeho chybu.

4. Vyčíslete odhad *skutečné* chyby klasifikace na *transformovaných* datech:

   - V každém cyklu validace na *trénovacím* oddílu dat naučte transformační bázi metodou PCA.
   - Klasifikátor naučte na projekci trénovacích dat přes naučenou bázi.
   - Odhad chyby klasifikátoru spočítejte na projekci *testovacích* dat přes bázi *trénovacích* dat.

Pro učení a klasifikaci používejte třídu `ClassificationTree`, jako transformaci použijte přiloženou funkci `pca.m`, která pro zadaná data vrací prvních N bázových vektorů. Pro učení stromu použijte metodu `fit()`, pro predikci používejte funkci `predict()` a pro zobrazení stromu použijte funkci `view(your_model,'mode','graph')`. Vypracovat můžete do souboru `cv11.m`.