EM algoritmus a učení s poloučitelem

Návod ke třetímu cvičení

Andrea Szabóová, szaboand@fel.cvut.cz

ZS 2010/2011

Návod

Cílem cvičení je vyzkoušet si EM algoritmus a seznámit se s učením s poloučitelem. Na vypracování úkolů budete mít celé cvičení. Předpokládá se domácí příprava a odevzdání úlohy na dalším cvičení.

1 EM algoritmus

1. Vygenerujte náhodná čísla s normálním jednorozměrným rozdělením pomocí matlabovské funkce `randn`. Zobrazte empirickou hustotu pravděpodobnosti (použijte funkci `bar(centers, bins)`).

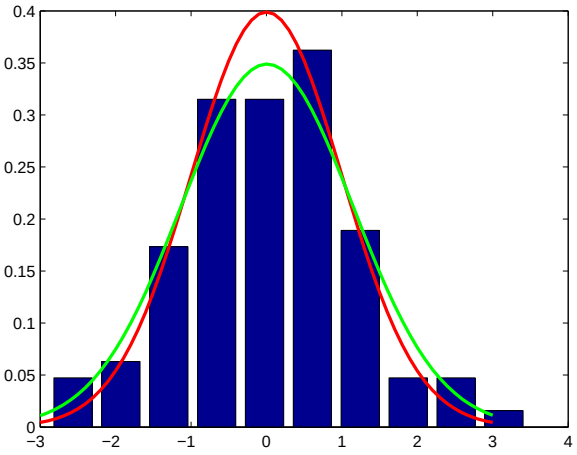
   Pozor: `[bins centers] = hist(x)`

   Funkce `hist` vám vrátí histogram s absolutními frekvencemi. Po normalizaci dostanete empirickou pravděpodobnostní funkci. Abyste dostali empirickou hustotu, musíte vydělit položky v histogramu šířkou binu. Proč? Zdůvodněte.

   Ve stejném obrázku zobrazte teoretickou hustotu pravděpodobnosti pomocí matlabovské funkce `normpdf`.

   Odhadněte parametry normálního rozdělení z vygenerovaných dat a odpovídající hustotu pravděpodobnosti zobrazte (pomocí funkce `normpdf`) ve stejném obrázku, ve kterem už jsou zobrazené empirická a teoretická hustota pravděpodobnosti.

   Očekávané výsledek můžete vidět na obrázku 1.



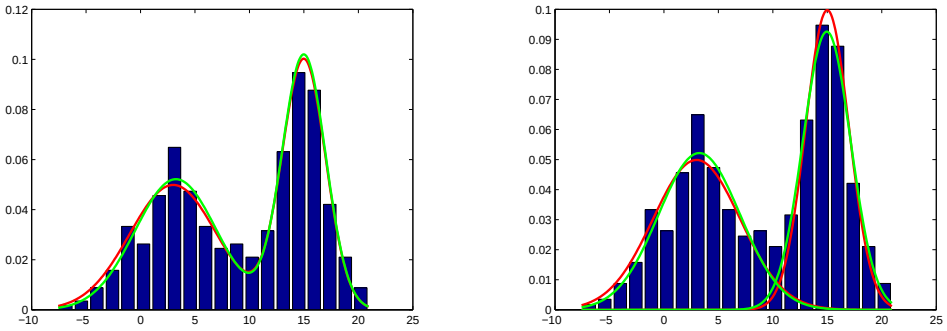
Obrázek 1: Očekávaný výsledek z úlohy 1. Křivka teoretické hustoty pravděpodobnosti je zobrazena červenou barvou, křivka odhadnuté hustoty pravděpodobnosti je zobrazena zelenou barvou.

2. Vygenerujte 200 vzorků ze směsi dvou normálních rozdělení ($\sigma_1 = 4$, $\mu_1 = 3$, $\sigma_2 = 2$, $\mu_2 = 15$).

   V případě směsi dvou různých normálních rozdělení (Gaussian Mixture Model) nelze pochopitelně vypočítat střední hodnotu a standardní odchylku pro jednotlivé složky směsi jako empirickou střední hodnotu a empirickou standardní odchylku z celého datasetu. Pozorovaná data musíte rozdělit na jednotlivá normální rozdělení, každé s vlastní střední hodnotou a standardní odchylkou. Jedna možnost, jak lze získat tyto parametry pro jednotlivá normální rozdělení, je použít EM algoritmus.

   Pomocí funkce "EM.m"najděte střední hodnoty a standardní odchylky pro složky směsi. Na jednom obrázku zobrazte empirickou, teoretickou a odhadnutou hustotu pravděpodobnosti pro směs normálních rozdělení. (Očekávaný výsledek můžete vidět v levé části obrázku 2.)

   Na dalším obrázku zobrazte empirickou, teoretickou a odhadnutou hustotu pravděpodobnosti pro jednotlivé složky směsi. (Očekávaný výsledek můžete vidět v pravé části obrázku 2.)

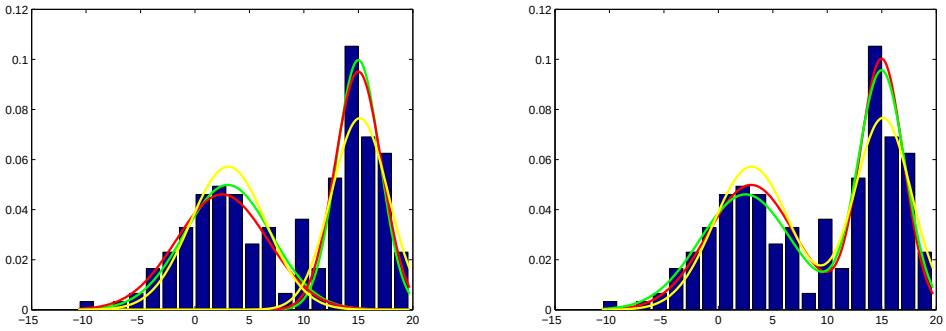


Obrázek 2: Očekávané výsledky z úlohy 2. Křivka teoretické hustoty pravděpodobnosti je zobrazena červenou barvou, křivka odhadnuté hustoty pravděpodobnosti je zobrazena zelenou barvou.

2 Učení s poloučitelem

3. Upravte EM algoritmus pro shlukování s poloučitelem (viz přednáška). Postupujte následujícím způsobem. K vstupním parametrům funkce EM přidejte parametr *klasifikace*, v němž budou obsažena přiřazení některých vzorků do shluků (1,2,...). Pro vzorky, pro něž nemáte přiřazení do shluků, použijete 0. V EM algoritmu fixujte přiřazení do shluků pro ty vzorky, pro něž bylo toto přiřazení zadáno na vstupu.

   Proveďte následující experiment. Upravte generování dat - přidejte klasifikaci, t.j. informaci o přislušnosti vzorků ke shlukům. Pro trénování použijete tuto informaci pouze pro náhodně vybranou část vzorků - zbytek použijete pro testování. Parametry směsi odhadněte dvěma způsoby. Za prvé, odhadněte parametry pouze ze vzorků, u nichž znáte klasifikaci (zde se EM algoritmus neuplatní). Za druhé, odhadněte parametry ze všech vzorků (i neklasifikovaných) pomocí upraveného EM algoritmu. Porovnejte přesnost přiřazení vzorků do shluků pomocí těchto dvou metod pro případ, že klasifikaci znáte pro 10%, 20% a 50% vzorků. K tomu využijete toho, že znáte klasifikaci pro všechny vzorky. Konkrétně použijte pro trénování všechny vzorky, přičemž jen pro některé z nich znáte přiřazení do tříd. Klasifikační přesnost spočítejte jako procento úspěšně zařazených vzorků, pro než jste na začátku přiřazení do tříd neznali.



Obrázek 3: Očekávané výsledky z úlohy 3. Křivka teoretické hustoty pravděpodobnosti je zobrazena červenou barvou, křivka hustoty pravděpodobnosti odhadnuté pomocí EM algoritmu je zobrazena zelenou barvou, křivka hustoty pravděpodobnosti odhadnuté pouze z anotovaných vzorků je zobrazena žlutou barvou.

Očekávaný výsledek můžete vidět na obrázku 3.