Posudek vedoucího diplomové práce

Název: EM algorithm for truncated Gaussian mixtures

Autor: Bc. Adéla Nguyenová

Předložená práce se zabývá odvozením Expectation-Maximization algoritmu pro useknuté gaussovské směsi a podrobným prozkoumáním jeho vlastností. Tento model je relevantní například ve statistice bodových procesů, kde jednotlivá pozorování (body) znázorňují polohu nějakého objektu, přičemž proces je pozorován pouze v omezeném pozorovacím okně. O přítomnosti bodů za hranicí pozorovacího okna nemáme žádnou informaci a dochází tedy přirozeným způsobem k useknutí dat (truncation). Na oblíbený a často v praxi používaný Thomasové proces je pak možno nahlížet jako na náhodný výběr z useknuté gaussovské směsi a snažit se odhadnout parametry modelu pomocí EM algoritmu, který dovede zohlednit useknutí dat.

V první kapitole autorka popisuje standardní verzi EM algoritmu a jeho konkrétní podobu pro (neuseknuté) gaussovské směsi, kde jednotlivé kroky algoritmu mají analytické vyjádření. Ve druhé kapitole pak podrobně odvozuje podobu algoritmu pro useknuté gaussovské směsi. V tomto případě už nemají jednotlivé kroky analytické vyjádření a je potřeba využít numerických metod.

Studentka také v druhé kapitole komentuje algoritmus představený v článku Lee a Scott (2012). V tomto článku je uveden postup, který autoři vydávají za EM algoritmus pro useknuté gaussovské směsi, nicméně obsahuje jisté heuristické zjednodušení, které umožňuje dopracovat se k analytickému vyjádření jednotlivých kroků. To je jistě výhodné pro implementaci postupu, nicméně to znamená, že nejde o EM algoritmus, ale jakousi upravenou verzi, a nejde tedy spoléhat na platnost řady tvrzení odvozených pro obecný EM algoritmus. Například pro obecný EM algoritmus hodnota věrohodnosti v jednotlivých iteracích neklesá, ale v případě upraveného algoritmu z článku to není splněno, jak studentka ukazuje na simulovaných příkladech. Emailovou komunikací s autorem článku bylo potvrzeno, že motivací pro heuristickou úpravu bylo získat analytické vyjádření pro update odhadů parametrů a vyhnout se numerickému hledání řešení v každé iteraci.

Ve třetí kapitole studentka podrobně rozebírá několik praktických problémů souvisejících s použitím odvozeného algoritmu – volba počtu komponent gaussovské směsi, inicializace algoritmu, zastavovací kritérium.

Podstatnou částí práce je porovnání odvozeného algoritmu s upraveným algoritmem z odkazovaného článku v řadě počlivě zpracovaných simulačních experimentů ve čtvrté kapitole a v příloze. Následně pak autorka oba postupy aplikuje na reálná data z oblasti bodových procesů.

Autorka pracovala pečlivě a samostatně, s citem pro detail. Dokázala opravit chybný postup publikovaný v literatuře a podrobně oba postupy porovnat. Použité zdroje jsou řádně citovány, formální úroveň práce je vysoká, jazyková stránka by mohla doznat jistého zlepšení. Práci doporučuji přijmout jako diplomovou práci k obhajobě.

V Praze dne 24. 7. 2022

RNDr. Jiří Dvořák, Ph.D.