

Vytěžování dat: EM algoritmus

Radomír Černoch

October 4, 2011

- <http://demonstrations.wolfram.com/ExpectationMaximizationForGaussianMixtureDistributions>
- Bishop: Pattern Recognition and Machine Learning, str. 437

Gaussovské rozdělení

- Hustota pravděpodobnosti:

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Gaussovské rozdělení

- Hustota pravděpodobnosti:

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Odhad parametrů:

- Střední hodnota z aritmetického průměru:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n$$

Gaussovské rozdělení

- Hustota pravděpodobnosti:

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Odhad parametrů:

- Střední hodnota z aritmetického průměru:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n$$

- Variance ze střední kvadratické odchylky:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2$$

Gaussovské rozdělení

- Hustota pravděpodobnosti:

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Odhad parametrů:

- Střední hodnota z aritmetického průměru:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n$$

- Variance ze střední kvadratické odchylky:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2$$

Vygenerujte si v Matlabu náhodné vzorky z $P(x|\mu = 10, \sigma^2 = 5)$ pomocí `normpdf` a zpětně odhadněte jejich parametry, tentokrát použití `mean` a `var`.

Směs Gaussovských rozdělání (GMM)

- Mějme 2 normální rozdělání:

$$P(x | \mu_m = 180, \sigma_m = 10) \text{ a } P(x | \mu_z = 170, \sigma_z = 8)$$

Směs Gaussovských rozdělení (GMM)

- Mějme 2 normální rozdělení:

$$P(x | \mu_m = 180, \sigma_m = 10) \text{ a } P(x | \mu_z = 170, \sigma_z = 8)$$

s následujícími směsnými koeficienty:

$$P(m) = 0.9 \text{ a } P(z) = 0.1$$

Směs Gaussovských rozdělení (GMM)

- Mějme 2 normální rozdělení:

$$P(x | \mu_m = 180, \sigma_m = 10) \text{ a } P(x | \mu_z = 170, \sigma_z = 8)$$

s následujícími směsnými koeficienty:

$$P(m) = 0.9 \text{ a } P(z) = 0.1$$

- Výsledná hustota pravděpodobnosti:

$$P(x | \dots) = P(m) \cdot P(x | \mu_m, \sigma_m) + P(z) \cdot P(x | \mu_z, \sigma_z)$$

Směs Gaussovských rozdělení (GMM)

- Mějme 2 normální rozdělení:

$$P(x | \mu_m = 180, \sigma_m = 10) \text{ a } P(x | \mu_z = 170, \sigma_z = 8)$$

s následujícími směsnými koeficienty:

$$P(m) = 0.9 \text{ a } P(z) = 0.1$$

- Výsledná hustota pravděpodobnosti:

$$P(x | \dots) = P(m) \cdot P(x | \mu_m, \sigma_m) + P(z) \cdot P(x | \mu_z, \sigma_z)$$

Zkuste si z této distribuce vygenerovat vzorky pomocí `randn` a zobrazit je v histogramu pomocí `hist`.

GMM: Odhady parametrů

- Dokáži ze své výšky odhadnout, jestli jsem muž nebo žena?

GMM: Odhady parametrů

- Dokáži ze své výšky odhadnout, jestli jsem muž nebo žena?
- Souhlasíte s následující úvahou:
 $P(m|x) \sim P(m) \cdot P(x|\mu_m, \sigma_m)$? (pro $P(z|x)$ obdobně)

GMM: Odhady parametrů

- Dokáži ze své výšky odhadnout, jestli jsem muž nebo žena?
- Souhlasíte s následující úvahou:
 $P(m|x) \sim P(m) \cdot P(x|\mu_m, \sigma_m)$? (pro $P(z|x)$ obdobně)
- Aby platil součet $P(m|x) + P(z|x) = 1$, používá se normalizační konstanta (jmenovatel je stejný pro $P(m|x)$ i $P(z|x)$):

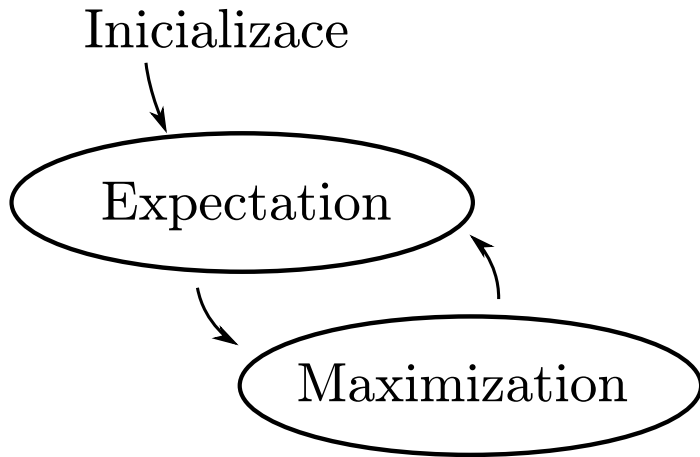
$$P(m|x) = \frac{P(m) \cdot P(x|\mu_m, \sigma_m)}{P(m) \cdot P(x|\mu_m, \sigma_m) + P(z) \cdot P(x|\mu_z, \sigma_z)}$$

GMM: Odhady parametrů

- Dokáži ze své výšky odhadnout, jestli jsem muž nebo žena?
- Souhlasíte s následující úvahou:
 $P(m|x) \sim P(m) \cdot P(x|\mu_m, \sigma_m)$? (pro $P(z|x)$ obdobně)
- Aby platil součet $P(m|x) + P(z|x) = 1$, používá se normalizační konstanta (jmenovatel je stejný pro $P(m|x)$ i $P(z|x)$):

$$P(m|x) = \frac{P(m) \cdot P(x|\mu_m, \sigma_m)}{P(m) \cdot P(x|\mu_m, \sigma_m) + P(z) \cdot P(x|\mu_z, \sigma_z)}$$

- Zjistěte, zda platí $P(m|x=160) > P(z|x=160)$ (pozn.: normalizační konstantu lze pro účel porovnání vynechat).



EM: 3 fáze

- 1 *Inicializace* náhodně nastaví parametry $P(m)$, $P(z)$, μ_m , σ_z , ...

EM: 3 fáze

- 1 *Inicializace* náhodně nastaví parametry $P(m)$, $P(z)$, μ_m , σ_z , ...
- 2 *Expectation* přiřadí instance oběma normálními rozděleními.

EM: 3 fáze

- 1 *Inicializace* náhodně nastaví parametry $P(m)$, $P(z)$, μ_m , σ_z , ...
- 2 *Expectation* přiřadí instance oběma normálním rozdělením.
- 3 *Maximization* odhadne parametry rozdělení na základě přiřazení z E fáze:

EM: 3 fáze

- 1 *Inicializace* náhodně nastaví parametry $P(m)$, $P(z)$, μ_m , σ_z , ...
- 2 *Expectation* přiřadí instance oběma normálním rozdělením.
- 3 *Maximization* odhadne parametry rozdělení na základě přiřazení z E fáze:

$$\mu_z \leftarrow \frac{1}{N_z} \sum_{n=1}^N P(z | x_n) x_n$$

$$\sigma_z^2 \leftarrow \frac{1}{N_z} \sum_{n=1}^N P(z | x_n) (x_n - \mu_z)^2$$

$$P(z) \leftarrow \frac{1}{N} \sum_{n=1}^N P(z | x_n)$$

- $N_z = \sum_{n=1}^N P(z | x_n)$... normalizační konstanta

Úloha (1/3)

- 1 Seznamte se s daty v souboru `height.csv`, který obsahuje tělesnou výšku vzorku 100 lidí, Američanů ve věku mezi 20 a 29 lety. Kromě výšky lidí (1. sloupec) obsahují data i jejich pohlaví (2. sloupec). Každý záznam tvoří jeden řádek tabulky.
- 2 Prohlédněte si dokumentaci k přiložené funkci `dataplot(data)`, která načtená data vykreslí do grafu:

```
>> data = csvread('height.csv'); dataplot(data);
```

Úloha (2/3)

- 4 Implementujte EM algoritmus pro maximum-likelihood optimalizaci parametrů směsi dvou normalních rozdělení. Popis algoritmu naleznete ve třetí přednášce (str. 21-24).
- Vstupem algoritmu bude první sloupec načtených dat (druhý sloupec můžete použít pro zpětnou kontrolu). Vhodně zvolte počáteční parametry obou rozložení.
 - Pokud Váš algoritmus vrátí matici 2×2 ve formátu

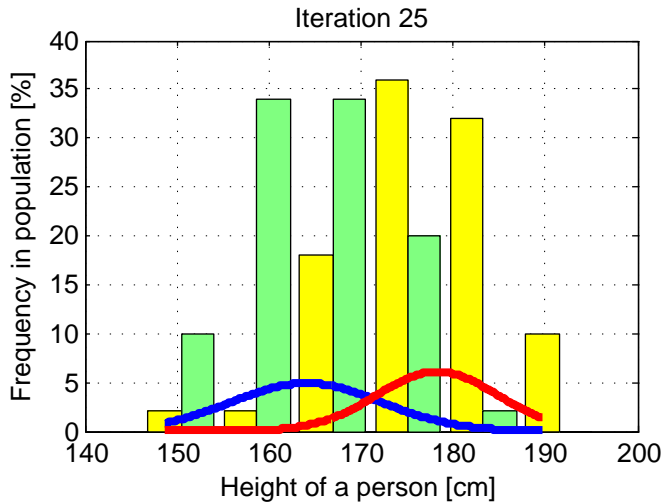
$$\text{params} = \begin{pmatrix} \mu_{\text{ženy}} & \sigma_{\text{ženy}} \\ \mu_{\text{muži}} & \sigma_{\text{muži}} \end{pmatrix}$$

můžete pro vykreslení obou rozdělení použít příkaz
`>> dataplot(data, params);`

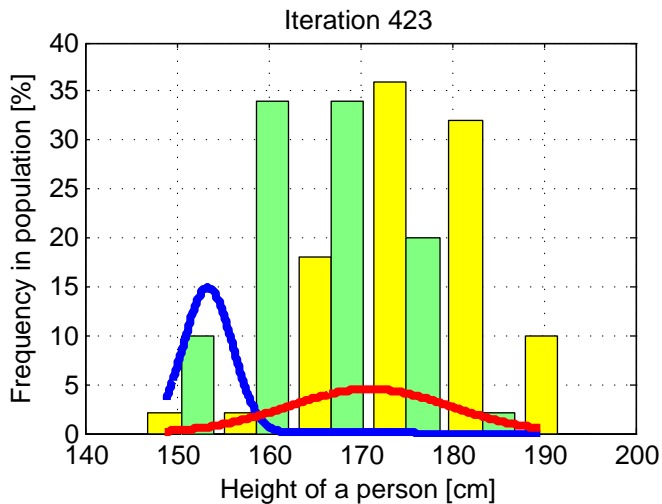
Úloha (3/3)

- 5 Vyvořte protokol o rozsahu cca. 1 strany A4, která shrne Vaši práci a analyzuje výsledky. Doporučený obsah:
 - grafy obou gaussovských rozložení v několika počátečních iteracích algoritmu a stav po konvergenci
 - počet iterací algoritmu (dochází-li k velkému rozptylu hodnot pro různá počáteční nastavení, spustte algoritmus několikrát a výsledek vyhodnoťte statisticky)
 - diskuze o vlivu prvotního přiřazení parametrů na jejich výsledné hodnoty.
 - rozbor, zda lze mezi výškou mužů a žen pozorovat statisticky významný rozdíl (využijte druhý sloupec vstupních dat a závěry z předchozích bodů)
 - poznámky k implementaci
- 6 Protokol odevzdejte do upload systému do 10.10.2011. Zdrojové kódy není nutné do systému nahrávat, ale můžete být požádáni o jejich ukázkou a předvedení během následujícího cvičení.

Úloha: Možný výsledek (1/3)



Úloha: Možný výsledek (2/3)



Úloha: Možný výsledek (3/3)

