Rješenje zadatka 1.5 predmeta Strojno učenje

Siniša Biđin

18. studenog 2012.

Programski kod priložen je u direktoriju `/src`. Prije korištenja, potrebno je instalirati *cabal*-paket `hmatrix`.

(a)

$$\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{ML}}\right)\left(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{ML}}\right)^{\mathrm{T}}$$

(b) Učitavamo skup primjera te izdvajamo značajke 4, 5, 6, 7 i 8:

```
ghci> fs <- feats [4..8] `fmap` load
```

Računamo $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ i $\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}$:

```
ghci> let (m, c) = (mean fs, cov fs)
```

$$\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \begin{bmatrix}61.683 & 8.609 & 168.693 & 40.977 & 20.579\end{bmatrix}^{\mathrm{T}}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \begin{bmatrix} 68.2 & 26.481 & 29.985 & -12.048 & 1.782 \\ 26.481 & 22.697 & 25.683 & -6.574 & 1.934 \\ 29.985 & 25.683 & 1102.35 & -252.835 & 85.036 \\ -12.048 & -6.574 & -252.835 & 61.129 & -19.201 \\ 1.782 & 1.934 & 85.036 & -19.201 & 6.666 \end{bmatrix}$$

(c) (i) Odabiremo slučajan uzorak od 100 primjera za učenje, na temelju kojih računamo ML-procjene vektora srednje vrijednosti i kovarijacijske matrice. Pritom gledamo, kao i u (b), samo značajke od 4 do 8.

```
ghci> let sample = randrows 100 fs
ghci> let (m', c') = (mean sample, cov sample)
```

(ii) Ponovimo (i) 50 puta te dobijemo 50 procjena vektora srednje vrijednosti i kovarijacijske matrice, svaka temeljena na drugačijem uzorku od 100 primjera:

```
ghci> samples <- replicateM 50 (randrows 100 fs)
ghci> let (ms, cs) = (map mean samples, map cov samples)
```

Računamo srednju vrijednost i varijancu procjenitelja najveće izglednosti za vektor srednje vrijednosti:

```
ghci> let ms' = fromRows ms -- Vektore stapamo u matricu.
ghci> let (msM, msV) = (mean ms', var ms')
```

Slično činimo i u slučaju kovarijacijskih matrica, gdje ovaj put svaku matricu promatramo kao jedan primjer s 25 značajki:

```
ghci> let cs' = fromRows (map flatten cs) -- 25 stupaca.
ghci> let (csM, csV) = (mean cs', var cs')
```

Razlika između naše procjene vektora srednje vrijednosti (`msM`) i stvarne vrijednosti parametra $\boldsymbol{\mu}$ (`m`) multivarijantne Gaussove razdiobe koja je generirala podatke prikazuje nam pristranost procjenitelja.

```
ghci> let bias = msM `sub` m
```

(iii) Ponavljamo čitav proces, samo što sada svaki uzorak sadrži 200 primjera za učenje umjesto prijašnjih 100:

```
ghci> samples <- replicateM 50 (randrows 200 fs)
ghci> let (ms', cs') = (map mean samples, map cov samples)
ghci> let ms'' = fromRows ms'
ghci> let (msM', msV') = (mean ms'', var ms'')
ghci> let cs'' = fromRows (map flatten cs')
ghci> let (csM', csV') = (mean cs'', var cs'')
```

Uspoređujemo vrijednosti varijanci oba procjenitelja u odnosu na vrijednosti dobivene u (ii):

```
ghci> let msVdiff = msV' `sub` msV
ghci> let csVdiff = csV' `sub` csV
```

Vidimo da su, sada kada svaki uzorak sadrži dvostruko više primjera, varijance oba procjenitelja pale. Možemo zaključiti da su procjenitelji valjani.

(d) Odabiremo značajke 2 (*distance circularity*), 18 (*major axis distance circularity*) te još 4, 10 i 13. Na temelju tih značajki svih primjera ulaznog skupa za učenje, računamo determinantu ML-procjene kovarijacijske matrice:

```
ghci> fs <- feats [2, 18, 4, 10, 13] `fmap` load
ghci> det (cov fs)
-1.3756112677457232e-5
```

Toliko mala vrijednost $|\hat{\Sigma}_{\mathrm{ML}}|$ ukazuje na jaku korelaciju između značajki ulaznih primjera. Konkretno, na korištenom skupu za učenje možemo vidjeti da je vrijednost značajke *major axis distance circularity* uvijek dvostruko veća od vrijednosti značajke *distance circularity*, što je uzrok takve vrijednosti determinante.