Cvičení z RPZ – Principal Component Analysis

Vojtěch Franc

5. ledna 2005

1 Zadání

Cílem cvičení je implementovat a použít metodu Principal Component Analysis (PCA). Vypracujte následující body zadání:

1. Implementujte metodu PCA. Výsledkem má být funkce `Y = moje_pca(X,d)` jejíž vstupem je matice $X \in \mathbb{R}^{n \times m}$ obsahující vstupní (trénovací) data a dimenze $d < n$ nového prostoru. Výstupem je matice $Y \in \mathbb{R}^{d \times m}$ obsahující nízkodimezionální reprezentaci dat X.

2. Aplikujte PCA na synteticky vygenerovaná 2-dimenzionální data. Použijte funkci `creatdata` ze Statistical Pattern Recognition Toolboxu. Zobrazte si vstupní a rekonstruovaná data.

3. Aplikujte PCA pro kompresi obrázku lena.jpg. Obrázek rozdělte na nepřekrývající se čtvercové oblasti velikosti 16×16. Každou oblast reprezentujte jako vektor s dimenzí $256 = 16 \times 16$. Takto získané vektory aproximujte v prostoru nižší dimenze (volba dimenze kontroluje komprimaci a kvalitu obrázku). Vektory opět rekonstruujte a zobrazte získaný (dekomprimovaný) obrázek. V odstavci 3 naleznete podrobnější vysvětlení.

   Pro rozdělení obrázku použijte funkce `im2col` a `col2im`. Obrázek si zobrazte funkcí `imshow`.

2 Principal Component Analysis

Metoda PCA hledá nízkodimenzionální reprezentaci dat, tak aby rekonstrukční chyba byla minimální. Nechť matice $X = [x_1, \ldots, x_m] \in \mathbb{R}^{n \times m}$ obsahuje m vektorů v n-dimenzionálním lineárním prostoru. Nízkodimenzionální reprezentace dat X se získá lineární transformací

$$y = W\,x + b\,, \tag{1}$$

kde $W \in \mathbb{R}^{d\times n}$ je ortonormální matice a $b \in \mathbb{R}$ skalár. Promítnutím vektorů z X pomocí projekce (1) získáme matici $Y = [y_1, \dots, y_m]$ obsahující m vektorů v d-dimenzionálním prostoru, kde $d < n$.

Původní data lze z jejich nízkodimenzionální reprezentace rekonstruovat inverzí transformací k (1), t.j.,

$$\hat{x} = W^T (y - b) , \tag{2}$$

kde se využilo, že matice W je ortonormální. Nechť matice $\hat{X} = [\hat{x_1}, \dots, \hat{x_m}] \in \mathbb{R}^{n\times m}$ obsahuje rekonstruovaná data. Rekonstrukční chyba je definována jako

$$\varepsilon = \frac{1}{m} \sum_{i=1}^{m} \|x_i - \hat{x}_i\|^2 .$$

Metoda PCA hledá parametry lineární transformace (1) a (2) tak, aby rekonstrukční chyba ε byla minimální. Optimální řešení této úlohy se získá z vlastních vektorů matice

$$S = \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T ,$$

a z vektoru

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x_i .$$

K získání vlastních vektorů a čísel je třeba vyřešit problém nalezení skaláru $\lambda > 0$ a vektoru $w \neq 0$ tak, že

$$\lambda w = S w .$$

Řešení tohoto problému je v Matlabu implementováno ve funkci `eig`. Nechť $\lambda_1 \geq \lambda_2 \geq \dots \lambda_d$ je posloupnost d největších vlastních čísel a $w_1, w_2, \dots w_d$ jsou jejich odpovídající vlastní vektory. Optimální řešení je rovno

$$W = [w_1, w_2, \dots, w_d]^T , \quad \text{a} \quad b = -W \mu . \tag{3}$$

Dosazením (3) do vztahu pro dopřednou (1) a zpětnou (2) transformaci dostaneme

$$y = W (x - \mu) , \quad \text{a} \quad \hat{x} = W^T y + \mu .$$

3 Aplikace PCA pro komprimaci obrázku

Vstupní obrázek `lena.jpg` velikosti 512×512 pixelů.

Trénovací množina $X = [x_1, \ldots, x_m]$ obsahuje $m = 1024$ čtvercových oblastí 16×16 representovaných jako vektory $x_i \in \mathbb{R}^n$, $n = 256$.

- **Učení PCA modelu:** Hledá se nová báze $W = [w_1, \ldots, w_d]$, $d < n$, v které se dobře aproximují oblasti X.
- **Komprimace obrázku:** Vypočítá se nová reprezentace $Y = [y_1, \ldots, y_m]$ oblastí X v bázi W.
- **Dekomprimace obrázku:** Z nové reprezentace Y se rekonstruují oblasti $\hat{X} = [\hat{x_1}, \ldots, \hat{x_m}]$.

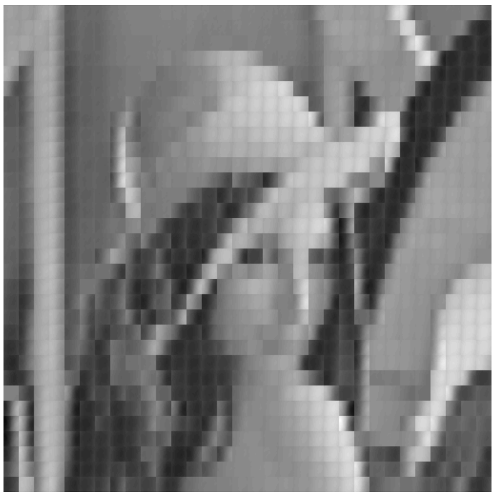
$comp$ ... kompresní poměr $comp = \dfrac{(d \cdot m) + (d \cdot n)}{n \cdot m}$.

d ... počet bázových vektorů - nová dimenze.

ε ... střední kvadratická chyba mezi $\varepsilon = \dfrac{1}{m} \sum_{i=1}^{m} \|x_i - \hat{x}_i\|^2$.

Obrázek 1: **Příklady komprimace obrázku** v závislosti na počtu bázových vektorů.

$d = 2$, $comp = 1.0\%$, $\varepsilon = 1.25$



$d = 5$, $comp = 2.4\%$, $\varepsilon = 0.67$

$d = 10$, $comp = 4.9\%$, $\varepsilon = 0.40$

$d = 50$, $comp = 24.4\%$, $\varepsilon = 0.07$