Principalkomponentanalyse

Peter Johansen

21. marts 2001

Vi vil beskrive en punktmængde i R^d ved at beskrive variationen fra punkternes middelværdi i enhver retning. Dette kaldes en principal komponent analyse (PCA). Vi begynder med at finde den retning gennem middelværdien som maksimerer kvadratsummen af punkternes projektion på retningen. Vi søger altså den akse gennem middelværdien, langs hvilken den samlede variation er størst.

Jeg vil i min notation lade middelværdien være trukket fra featurepunkterne, således at koordinatsystemets origo har alle koordinater 0. Lad n være en normalvektor i den givne retning. Lad de N punkter være x_i. Skalarproduktet af n og x_i betegnes $(n \cdot x_i)$. Den gennemsnitlige kvadratsum i retningen n kan skrives

$$\begin{aligned} \nu &= \frac{1}{N}\sum_{i=1}^{i=N}(n \cdot x_i)^2 \\ &= \frac{1}{N}\sum_{i=1}^{i=N}(n_1\, x_{i_1} + n_2\, x_{i_2})^2 \end{aligned}$$

For at finde de ekstreme retninger skal vi maksimere ν med hensyn til n under den bibetingelse at normalen har længde 1. Vi indfører nu en Lagrange multiplikator λ. I stedet for at maksimere ν med hensyn til n vil vi maksimere

$$\nu + \lambda(n \cdot n - 1)$$

men nu med hensyn til n og λ. Da $n \cdot n = 1$ giver dette også en gyldig løsning for n.

For at forenkle fremstillingen vil jeg delvis regne i koordinater, og jeg vil antage at rummet har to dimensioner, altså at $n = (n_1, n_2)$ og at $x_i = (x_{i_1}, x_{i_2})$. Beregningerne kan nemt generaliseres til et vilkårligt endeligt antal dimensioner.

For at finde maksimum differentierer jeg med hensyn til n_1 og med hensyn til n_2 og sætter resultaterne til 0.

$$\begin{aligned} \frac{1}{N}\sum(n \cdot x_i)x_{i_1} + \lambda n_1 = 0 \\ \frac{1}{N}\sum(n \cdot x_i)x_{i_2} + \lambda n_2 = 0 \end{aligned}$$

Jeg har nu to ligninger i de ubekendte n_1 og n_2. Koefficienterne i ligningssystemet er funktioner af de N punkters koordinater.

$$\frac{1}{N}\sum_i (n_1 x_{i_1} + n_2 x_{i_2}) x_{i_1} + \lambda n_1 = 0$$
$$\frac{1}{N}\sum_i (n_1 x_{i_1} + n_2 x_{i_2}) x_{i_2} + \lambda n_2 = 0$$

Omskrevet på matrix form får vi

$$\begin{pmatrix} \frac{1}{N}\sum(x_{i_1})^2 & \frac{1}{N}\sum x_{1i}x_{i_2} \\ \frac{1}{N}\sum x_{1i}x_{i_2} & \frac{1}{N}\sum(x_{21})^2 \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = -\lambda \begin{pmatrix} n_1 \\ n_2 \end{pmatrix}$$

Matricen som vi kalder C (engelsk *covariance matrix*) er reel og symmetrisk. Derfor har den reelle egenværdier og dens egenvektorer er orthogonale. Ligningssystemet har åbenbart to løsninger. De ubekendte - lagrangemultiplikatoren og normalen - er samhørende værdier af egenværdi og egenvektor.

Nu vil jeg skifte koordinatsystem. Den koordinattransformation som lader koordinatakserne være egenvektorerne i kovariansmatricen kaldes **Karhunen-Loéve** transformationen. Egenvektorerne kaldes de *principale akser* og en vektors koordinater i **KL** systemet kaldes dens *principale komponenter*. Den matrix A som transformerer en vektors koordinater til **KL** systemet er givet ved at dens rækker er egenvektorerne. Det kan nemt indses idet $Ax = (e_1 \cdot x, e_2 \cdot x)$ og første komponenten er derfor projektionen af x på e_1 og anden komponenten er projektionen af x på e_2. I dette koordinatsystem bliver punkternes koordinater $x'_i = Ax_i$, hvor A er matricen

$$A = \begin{pmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{pmatrix}$$

Lad os finde udtrykket for den gennemsnitlige variation v i det ny koordinatsystem.

$$\begin{aligned}
\nu &= \frac{1}{N}\sum(n \cdot x_i)^2 && (1)\\
&= \frac{1}{N}\sum(n_1 x_{i_1} + n_2 x_{i_2})^2 && (2)\\
&= \frac{1}{N}(n_1^2 \sum x_{i_1}^2 + 2n_1 n_2 \sum x_{i_1}x_{i_2} + n_2^2 \sum x_2^2) && (3)\\
&= \begin{pmatrix} n_1 n_2 \end{pmatrix} \begin{pmatrix} \frac{1}{N}\sum(x_{i_1})^2 & \frac{1}{N}\sum x_{1i}x_{i_2} \\ \frac{1}{N}\sum x_{1i}x_{i_2} & \frac{1}{N}\sum(x_{21})^2 \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} && (4)\\
&= n^t C n && (5)\\
&= (A^{-1}n')^t C (A^{-1}n') && (6)\\
&= n'^t (A^{-1})^t C A^{-1} n' && (7)\\
&= n'^t A C A^{-1} n' && (8)
\end{aligned}$$

Husk i step 7 at den transponerede af en orthonormal matrix er den inverse matrix.

Vi ser nu at i **KL**-koordinater er C en diagonalmatrix ACA^{-1} med egenværdierne som diagonalelementer.

Fejl ved Approximation

Lad os nu permutere koordinaterne, så egenværdierne er sorteret efter faldende størrelse. Egenvektoren e_1 er den retning i hvilken variationen er maksimal. Størrelsen af variationen er λ_1. Den gennemsnitlige variation ν i retning n' er $n'^t ACA^{-1} n'$. Hvis vi approksimerer punkterne $\{x_i\}$ med deres projektion på e_1 begår vi en gennemsnitlig kvadreret fejl som er $\frac{1}{N}\sum x_{i_2}^2 = \lambda_2$.

Dette argument kan generaliseres hvis der er flere dimensioner end $d = 2$. Hyppigt reducerer man dimensionen ved kun at beholde de k største egenvektorer. Den gennemsnitlige fejl som man begår er $\sum_{k+1}^{N}\sum x_k^2 = \sum_{k+1}^{N}\lambda_k$. Den matrix som transformerer en vektors koordinater til det reducerede **KL** system er givet ved at de første k rækker er de største k egenvektorer; de øvrige er nulvektoren.

Beregning

Lad M være en $N \times d$ matrix hvis i'te række er x_i. Vi bemærker at $C = \frac{1}{N}M^t M$ med størrelse $N \times N$. Det gør det nemt at skrive et program i matlab til principal komponent analyse. Der findes nemlig en funktion *eig* som finder egenvektorer og egenværdier for en matrix. Hvis antallet af punkter N imidlertid er mindre end dimensionen d, kan vi effektivisere beregningen ved at beregne egenvektorer for en $N \times N$ matrix i stedet for en $d \times d$ matrix.

Lad nemlig $D = \frac{1}{N}MM^t$ med størrelsen $d \times d$ og lad f_i være dens egenvektorer med egenværdier μ_i.

$$\begin{aligned}
C(M^t f_i) &= \frac{1}{N}(M^t M)(M^t f_i) \\
&= M^t(\frac{1}{N}MM^t)f_i \\
&= M^t D f_i \\
&= M^t \mu_i f_i \\
&= \mu_i(M^t f_i)
\end{aligned}$$

Dette vil sige at vektorerne $M^t f_i$ er egenvektorer for C med egenværdier μ_i. Vektorerne er dog ikke normeret til længde 1. Det skal gøres eksplicit.

Visualisering

Man plejer at angive en punktmængdes variation ved at tegne en flade som i hver retning ud fra middelværdien angiver variationens størrelse. Det viser sig at det giver et godt grafisk overblik hvis man i stedet for ν viser $\sqrt{\nu}$.[1]

En helt andet spørgsmål er hvorledes man visualiserer en d-dimensional punktmængde og en $d-1$ dimensional hyperflade. En enkel metode er at vise projektionerne på de $d(d-1)/2$ koordinatplaner.

Normalfordelingen

Formuleringen har indtil nu været i termer af beregning af den gennemsnitlige kvadratsumsafvigelse. Hyppigt ser man en anden formulering som involverer begrebet "probability density function" eller "sandsynlighedstæthedsfunktion". Man kan opfatte de observerede punkter som stikprøver fra en sandsynligheds-tæthedsfunktion i R^d. En sådan funktion er defineret overalt på R^d og har integralet 1. Sandsynligheden for at udfaldet af et forsøg ligger i et område er integralet af funktionen over området. Et spørgsmål man da må stille sig er hvilken tæthedsfunktion man skal vælge som model for de observerede punkter samt derefter hvor god modellen er til at beskrive observationerne. Dette begrebsapparat anvendes i statistik. Man taler om en covariansmatrix for en funktion. Her skal man så spørge hvorledes man kan estimere den ud fra observationerne. Udtrykket bliver faktisk (næsten) det samme, nemlig $C = \frac{1}{N-1}A^tA$. I retning af en egenvektor taler man om *variansen* v som er den tilhørende egenværdi. Man taler om *standardafvigelsen* som er $\sigma = \sqrt{\nu}$.

En hyppig anvendt model er den d-dimensionale Gaussfordeling eller den d-dimensionale Normalfordeling, som har funktionsudtrykket

$$G = \frac{1}{(2\pi)^{d/2}|C|^{\frac{1}{2}}}e^{-\frac{1}{2}x^tC^{-1}x}$$

når den er centreret omkring origo.

:

Litteratur

[1] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Processing*, 61(1):38–59, January 1995.

[1] Grunden er at man kan estimere den sandsynlighedsfordeling som med størst likelihood ville have givet de givne punkter x_i ved N forsøg. Denne fordeling viser sig at være en d-dimensional Normalfordeling med kovariansmatrix C