1 Wprowadzenie

Dane wysokowymiarowe są często kłopotliwe do analizy. Może się zdarzyć że badane zjawisko z natury jest wysokowymiarowe i nic na to nie da się poradzić. Jednakże, często dane wykazują regularności i może je dobrze reprezentować (zredukować) przez dane w przestrzeni niższego wymiaru. Proste i często używane są modele liniowe: zakładamy że czyste dane są w hiperpłaszczyźnie niższego wymiaru, ale jest do nich dodany szum który powiększa wymiar. Inne modele są nieliniowe: zakładamy że obserwowane dane są wynikiem nieliniowego przekształcenia danych niskowymiarowych.

Redukcja wymiaru, tzn. odworowanie danych w przestrzeń mniejszego wymiaru może dać różne korzyści

- mniej danych zwykle zmniejsza koszt obliczeń
- redukcja wymiaru może zmniejszyć wpływ szumu
- redukcja wymiaru może pomagać w uogólnieniu

2 Metody

2.1 Losowe odwzorowania

Jedną z najprostszych metod redukcji wymiaru jest użycie losowego odwzorowania (macierzy). Na pierwszy rzut oka wygląda to podejrzanie, dlaczego losowe odwzorowanie mialoby dać sensowny wynik? Jednakże teoria compressed sensing mówi że jeśli A jest losową macierzą $m \times N$ o elementach normalnych (Gaussowskich) zaś prawdziwe dane mają wymiar k to przy odpowiedniej relacji k, m i N na podprzestrzeni wymiaru k macierz $B = \frac{1}{\sqrt{m}} A$ z dużym prawdopodobieństwem jest bliska izometrii. Dokładniej, niech $V \subset \mathbb{R}^N$ będzie podprzestrzenią wymiaru k, ι_V będzie włożeniem V w $\mathbb{R}^N$. Zachodzi lemat

Lemat 2.1 *Istnieje C takie że jeśli $\delta, \epsilon \in (0,1)$ i*

$$m \geq C\delta^{-2}(7k + 2\log(2\epsilon^{-1}))$$

to

$$\|\iota_V^* B^* B \iota_V - I_V\| < \delta$$

z prawdopodobieństwem co najmniej $1 - \epsilon$.

Komentarz: m wyżej jest mniejsze niż w typowym wyniku o compressed sensing, ale wyżej mamy słabszy warunek, tzn. jedną przestrzeń V.

Można oczekiwać podobnego wyniku dla rzutów na losowe podprzestrzenie wymiaru m. Losowe odwzorowania zwykle są gorsze od innych metod. Wymagają wyższego wymiaru i nie zmniejszają szumu. Ale są tańsze od innych metod i dlatego mogą się przydać do zmniejszania kosztu obliczeń. Warto to dodać że dla dużych i wysokowymiarowych danych koszt losowego odwzorowania ciągle może być duży (bo losowa macierz jest duża). Lecz zaobserwowano że można skombinować szybkie przekształcenie (jak FFT) z użyciem rzadkiej macierzy losowej co daje znaczne przyspieszenie obliczeń.

2.2 Pomijanie współrzędnych

Prostą metodą redukcji wymiaru jest pominięcie nieistotnych współrzędnych. Np. przy reprezentacji tekstu wektorami słów czy ogólniej cech popularną techniką jest pomijanie bardzo często i bardzo rzadko występujących słów, na zasadzie że częste słowa niosą mało informacji (słabo odróżniają teksty) zaś słowa bardzo rzadko występujące nie dają podstawy do porównania tekstów. Jednak proste kryteria pomijania współrzędnych działają niezbyt dobrze i podane dalej metody zwykle są lepsze.

2.3 PCA

Mając dane punkty x_i, $i = 1, \ldots, l$ metoda PCA stara się dobrać hiperpłaszczyznę V tak by zminimalizować sumę kwadratów odległości x_i do V:

$$V = \text{argmin}_V \sum_{i=1}^{l} d(x_i, V)^2.$$

Alternatywnie, z probablistycznego punktu widzenia maksymalizujemy wariancję z próby. Dokładniej, niech

$$z = \frac{1}{l} \sum_{i=1}^{l} \pi_V(x_i)$$

gdzie π_V oznacza rzut na V. Mamy $z \in V$ i

$$\|x_i - z\|^2 = d(x_i, V)^2 + \|\pi_V(x_i) - z\|^2$$

czyli minimalizacja sumy $d(x_i, V)^2$ wyżej jest równoważna maksymalizacji sumy $\|\pi_V(x_i) - z\|^2$, czyli maksymalizacja wariancji z próby $\pi_V(x_i)$.

Algorytm PCA

1. Niech
$$\bar{x} = \frac{1}{l} \sum_{i=1}^{l} x_i.$$
2. Niech $y_i = x_i - \bar{x}$.
3. Niech A będzie macierzą której wierszami są y_i.
4. Niech $B = A^*A$
5. Diagonalizujemy B.
6. Niech W będzie podprzestrzenią rozpiętą przez m wektorów własnych odpowiadających m największym wartościom własnym B.

Przykładowe zadanie: Mamy dane 4 punkty: $(-1, 1)$, $(1, -1)$, $(2, 2)$, $(-2, -2)$. Chcemy zredukować wymiar do 1.

Widać że $\bar{x} = 0$. Macierz A to
$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \\ 2 & 2 \\ -2 & -2 \end{pmatrix}$$
Następnie
$$B = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$$
Wartości własne B to 4 i 16. Wektor własny B odpowiadający wartości własnej 16 to $v = (1, 1)$. Przy rzutowaniu $(-1, 1)$ na v otrzymujemy 0. Również $(1, -1)$ rzutuje się na 0. $(2, 2)$ leży na prostej rozpiętej przez v, czyli jest swoim własnym rzutem. Również $(-2, -2)$ jest swoim własnym rzutem.

Koszt tworzenia macierzy AA^* może być znaczny. Tańszą metodą może być użycie SVD (rozkład na wartości osobliwe). Mianowicie, niech
$$A = UDV$$
gdzie U i V spełniają $U^*U = I$, $VV^* = I$, zaś D jest macierzą diagonalną z nieujemnymi elementami na diagonali. Mamy
$$B = A^*A = V^*DU^*UDV = V^*D^2V$$
czyli wiersze macierzy V daje wektory własne B, zaś D^2 daje wartości własne B. Czyli mamy następujący algorytm

1. Niech
$$\bar{x} = \frac{1}{l}\sum_{i=1}^{l} x_i.$$
2. Niech $y_i = x_i - \bar{x}$.
3. Niech A będzie macierzą której wierszami są y_i.
4. Niech $A = UDV$ będzie rozkładem na wartości osobliwe
5. Niech W będzie podprzestrzenią rozpiętą przez m wierszy V odpowiadających m największym wartościom na diagonali D.

Wariantem PCA jest kernel PCA. W tym wariancie punkty x_i odwzorowujemy nieliniowo w przestrzeń wysokiego wymiaru i (logicznie) stosujemy PCA w przestrzeni wysokowymiarowej. Oznaczając przez ϕ odwzorowanie w przestrzeń wysokowymiarową mamy

$$\bar{x} = \frac{1}{l}\sum_{i=1}^{l} \phi(x_i),$$

$$B_{i,j} = \langle \phi(x_j) - \bar{x}, \phi(x_i) - \bar{x} \rangle$$

tzn. by obliczyć macierz B wystarczy znać iloczyny sklarne wektorów $\phi(x_i)$ i $\phi(x_j)$ (iloczyny skalarne $\bar{x}$ wurażamy w terminach $\phi(x_i)$). W kernel PCA zakładamy że

$$\langle \phi(x_j), \phi(x_i) \rangle = K(x_j, x_i)$$

gdzie K jest znaną (zadaną) funkcją zwaną jądrem (ang. kernel). Oznacza to że macierz B można faktycznie obliczyć w przestrzeni niskiego wymiaru. Równierz efekt rzutu daje się obliczyć w przestrzeni niskiego wymiaru, co oznacza że koszt kernel PCA nie zależy od wymiaru obrazu ϕ. Dzięki temu że kernel PCA używa nieliniowe ϕ uzyskujemy nieliniowe odwzorowanie kosztem podobnym do liniowego (dodatkowy koszt to oblicznie wartości $K(x, y)$).

2.4 ICA

PCA używa przedstawienie

$$A = UDV = CV$$

gdzie $C = UD$, co oznacza że wiersze A są kombinacjami liniowymi wierszy V. Wiersze V są ortogonalne, czyli z probabilistycznego punktu widzenia są nieskorelowane. Innymi słowy, PCA przestawia x_i jako transformację

wielkości nieskorelowanych. Gdybyśmy mieli rozkład normalny to wielkości nieskorelowane byłyby niezależne. Ale ogólnie możemy mieć zależność. ICA transformuje nasze wielkości, tak by były możliwie bliskie niezależności. Mianowicie, jeśli W jest macierzą ortogonalną to

$$A = UDW^*WV = \tilde{C}Y$$

gdzie $\tilde{C} = UDW^*$ zaś $Y = WV$, co jest alternatywnym przedstawieniem x_i jako transformacji wielkości nieskorelowanych. Chcemy dobrać W tak by aminimalizować zależności między wierszami Y (idealnie wiersze Y byłyby niezależne, ale to zwykle jest niemożliwe). Jedna (teoretyczne) możliwość to powiar zależności przy pomocy informacji wzajemnej (używa się też nazwę odległość Kullbacka-Leiblera)

$$I(y) = \int P(y) \log_2 \left(\frac{P(y)}{\Pi_j P(y_j)} \right) dy$$

$$= -\int P(y) \log_2 \left(\frac{\Pi_j P(y_j)}{P(y)} \right) dy$$

$$= -E(\log_2 \left(\frac{\Pi_j P(y_j)}{P(y)} \right))$$

Mamy

$$E(\frac{\Pi_j P(y_j)}{P(y)}) = \int \frac{\Pi_j P(y_j)}{P(y)} P(y) dy$$

$$= \int \Pi_j P(y_j) dy = \Pi_j \int_{y_j} P(y_j) dy_j = 1.$$

Jako że logartym jest funkcją sciśle wklęsłą na mocy nierówności Jensena

$$I(y) = -E(\log_2 \left(\frac{\Pi_j P(y_j)}{P(y)} \right)) \geq \log(1) = 0$$

z równością tylko wtedy gdy

$$\frac{\Pi_j P(y_j)}{P(y)} = 1.$$

Innymi słowy $I(y)$ jest równe 0 wtedy i tylko wtedy gdy y_j są niezależne, w przeciwnym razie $I(y) > 0$.

W praktyce obliczanie entropii jest bardzo kłopotliwe, bo mamy do dyspozycji tylko próbkę. Dlatego zamiast entropii wzajemnej stosuje się funkcje

które są łatwiejsze do obliczania. Wariantem entropii wzajemnej jest $J(y)$ zadane wzorem

$$J(y) = \sum_j (-E(\log_2(z_j)) + E(\log_2(y_j)))$$

gdzie z_j ma rozkład normalny o tej samej wariancji jak y_j (u nas wariancje są równe 1). J też jest kłopotliwe do obliczania, ale stosuje się przybliżenie

$$J(y) \approx \sum_j (E(G(y_j)) - E(G(z_j)))$$

gdzie $G(t) = \frac{1}{a}\log(\cosh(at))$ dla wybranego $a \in [1, 2]$. W ostatnim wzorze wartość oczekiwaną można przybiżać przez średnią z próby co pozwala na stosowanie metod numerycznych.

Ograniczeniem ICA jest to że nie daje nic ponad PCA dla zmiennych o rozkładzie normalnym. Jednakże dla zmiennych nienormalnych działa dobrze, w szczególności jak obserwacje są kombinacjami liniowymi zmiennych niezależnych to ładnie wydziela składowe niezależne.

Przy redukcji wymiaru najprostsze podejscie napierw używa PCA do faktycznego zmniejszenia wymiaru a następnie ICA do wybrania lepszych współrzędnych. Inne podejście najpierw stosuje ICA w pełnym wymiarze a następnie używa dodatkową procedurę. Był też proponowany wariant w którym za istotne przyjmuje się składowe odległe od normalnych, zaś składowe bliskie normalnym uznaje się za szum.

2.5 MDS

MDS (multidimensional scaling) próbuje zachować odległości między punktami. Dokładniej, zakładamy że zadana jest macierz odległości i próbujemy ją zachować. Jednakże przestrzeń euklidesowa ma dość szczególne własności i zwykle nie jest możliwe dokładne (izometryczne) zanurzenie. Zamiast tego, chcemy by odległości były możliwie mało zniekształcane. Jeden z popularnych wariantów MDS to minimalizacja wyrażenia niżej

$$\sum_{i \neq j} (d_{i,j} - \|z_i - z_j\|)^2$$

gdzie $d_{i,j}$ to dana macierz odległości zaś szukamy werktrów $z_i \in \mathbb{R}^k$. To wyrażenie można minimalizować dość prostym algorytmem gradientowym. Obecnie jest szereg wariantów MDS, nie będziemy ich tu szczegółowo omawiać, zauważmy jednak że prowadzą one zwykle do odwzorowań nieliniowych.

3 Efekty

W zagadnieniu klasyfikacji tekstu gdzie tekst reprezentujemy przez wektory słów wymiary źródłowe zwykle są rzędu tysięcy. PCA czy ICA pozwala zredukować wymiar do rzędu 30 równocześnie poprawiając wyniki w porównaniu do bezpośredniego użycia danych źródłowych, przy tym ICA daje nieco lepsze wyniki. Losowe odwzorowania wymagają wysokiego wymiaru i działają niezbyt dobrze.

W przypadku obrazów PCA często pozwala zredukować wymiar z dziesiątków tysięcy do kilkuset. Z drugiej strony, wiele metod działa lepiej jeśli obrazy reprezentuje się bezpośrednio jako macierze punktów czy też gdy wyodrębnia się cechy używając macierz punktów bez redukcji wymiaru.

Ogólniej, mniejszy wymiar i lepsze wyniki zwykle osiąga się przy pomocy metod nieliniowych (o których niewiele było wcześniej).