Klątwa wielowymiarowości
The Curse of Dimensionality
2. część

Agnieszka Pocha
Michał Kowalik

18 marca 2015

na podstawie książki:
Bertrand Clarke, Ernest Fokoue, Hao Helen Zhang
Principles and Theory for Data Mining and Machine Learning

Agenda

Klątwa wielowymiarowości

Przy wysokim wymiarze przestrzeni, dane są zbyt rzadkie.
Przy wysokim wymiarze przestrzeni, liczba możliwych modeli do rozważenia rośnie zbyt szybko.

Liczba modeli rośnie super-wykładniczo (superexponential) wraz ze wzrostem rozmiaru.

Przykład

Dla $p = 1$ jest 7 możliwych różnych modeli:

$\mathbb{E}(Y) = \beta_0$,
$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1$,
$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$,

$\mathbb{E}(Y) = \beta_1 x_1$,
$\mathbb{E}(Y) = \beta_0 + \beta_2 x_1^2$,

$\mathbb{E}(Y) = \beta_2 x_1^2$,
$\mathbb{E}(Y) = \beta_1 x_1 + \beta_2 x_1^2$,

Dla $p = 2$ liczba możliwości wynosi 63.

Oczywistym jest, że problem się pogarsza dla wielomianów większego rzędu.

Principal Component Analysis - Analiza głównych składowych

Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję pierwszej współrzędnej, następnie wariancję drugiej współrzędnej, itd..

PCA jest często używana do zmniejszania rozmiaru zbioru danych statystycznych, poprzez odrzucenie ostatnich czynników.

Wariancja

Klasyczna miara zmienności. Intuicyjnie utożsamiana ze zróżnicowaniem zbiorowości.

$$D^2(X) = E(X^2) - [E(X)]^2$$

E - wartość oczekiwana

Algorytm

- Obliczenie wartości średniej dla każdej cechy: $u[m] = \frac{1}{N}\sum_{n=1}^{N} X[m,n]$
- Policzenie wartości odchyleń dla każdej komórki danych: $B[i,j] := X'[i,j] = X[i,j] - u[i]$
- Wyznaczenie macierzy kowariancji: $\mathbf{C} = \mathbb{E}\left[\mathbf{B} \otimes \mathbf{B}\right] = \mathbb{E}\left[\mathbf{B} \cdot \mathbf{B}^T\right] = \frac{1}{N}\mathbf{B} \cdot \mathbf{B}^T$
- Policzenie wartości własnych macierzy kowariancji: $\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}$
  wartość własna odpowiadająca temu wektorowi to skala podobieństwa tych wektorów.
  gdzie D jest macierzą przekątniową wartości własnych C.
- Wybór wartości własnych: można dokonać zawężenia wymiaru przestrzeni.

Algorytm

- Wyznaczenie wektorów własnych:
$$\begin{bmatrix} a_{11}-\lambda & a_{12}\cdots & a_{1n} \\ a_{21} & a_{22}-\lambda\cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn}-\lambda \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = 0$$
- Rzutowanie na wektory własne:
$$y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix} = V^T \cdot x = \begin{bmatrix} v_0^T \\ v_1^T \\ \vdots \\ v_{n-1}^T \end{bmatrix} \cdot x,$$ gdzie:
  - V to macierz wektorów własnych
  - x to wektor rzutowany
  - y to wektor w nowej przestrzeni
  - N to liczba wektorów własnych