AGML, Sommersemester 2020 Vorlesung 22 (7 Seiten)

Dimensionsreduktion

Anne Driemel Letzte Aktualisierung: 14. Juli 2020

In vielen Anwendungen sind die Daten, die wir als Eingabe für unsere Lernalgorithmen bekommen, hochdimensional. In der Bildanalyse, zum Beispiel, ist jeder Datenpunkt eine Kombination von vielen Pixelwerten. Jeder einzelne Pixel ist dabei ein eigenes Merkmal und nimmt somit eine eigene Dimension im Merkmalsraum ein. Gleichzeitig kann man sich leicht vorstellen, dass der exakte Werte jedes einzelnen Pixels nicht unbedingt für die Analyse benötigt wird. In der Dimensionsreduktion geht es darum die Daten in einen geeigneten niedrig-dimensionalen Unterraum zu projizieren, um die Daten vereinfacht darzustellen, wobei die Datenpunkte trotzdem möglichst gut erhalten bleiben sollen.

1 Definition der Zielfunktion

Sei $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ eine Menge von Datenpunkten und sei $k \in \mathbb{N}$ ein Parameter mit $k \leq d$. Wir wollen S mithilfe einer Funktion $f : \mathbb{R}^k \to \mathbb{R}^d$ beschreiben, definiert durch $\mu \in \mathbb{R}^d$ und eine $d \times k$ Matrix V mit

$$f(\lambda) = \mu + V\lambda, \text{ mit } \mu \in \mathbb{R}^d$$

Wir verlangen außerdem von der Matrix V, dass sie orthonormal ist, das heißt

(i) für jeden Spaltenvektor v_i von V gilt, dass $\langle v_i, v_i \rangle = 1$

(ii) für je zwei Spaltenvektoren v_i und v_j von V gilt, dass $\langle v_i, v_j \rangle = 0$

Die Funktion f bildet auf eine k-dimensionale Hyperebene im $\mathbb{R}^d$ ab. Ziel ist es also, die Datenpunkte in S innerhalb einer k-dimensionalen Hyperebene angemessen darzustellen. Die Dimensionsreduktion geschieht hier indem wir jedes x_i über seinen Index dem Vektor λ_i assoziieren. Die Abbildung in den k-dimensionalen Unterraum wird also durch die Wahl der Vektoren $\lambda_1, \ldots, \lambda_n$ bestimmt. Wie gut unsere Repräsentation von S ist, messen wir mithilfe der Summe der quadratischen Abstände. Dies wird in der folgenden Zielfunktion ausgedrückt.

Wir wollen einen Vektor μ, eine Matrix V und Spaltenvektoren $\lambda_1, \ldots, \lambda_n$ finden, welche zusammen die Zielfunktion

$$\phi(\mu, V, \lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} \|x_i - f(\lambda_i)\|^2$$

minimieren. Diese Zielfunktion lässt sich noch vereinfachen. Dazu betrachten wir zunächst λ_i und halten dabei V und μ und λ_j mit $i \neq j$ fest. Man kann zeigen, dass ϕ für

$$\lambda_i = V^T(x_i - \mu) \tag{1}$$

minimiert wird. Insbesondere ist $f(\lambda_i)$, für diese Wahl von λ_i, die orthogonale Projektion von x_i auf die Hyperebene, die durch μ und V gegeben ist, und damit der Punkt in der Hyperebene mit dem kleinsten Abstand zu x_i. Im nächsten Schritt halten wir V und die $\lambda_1, \ldots, \lambda_n$ fest und

minimieren ϕ über alle Werte von μ. Hier können wir die partielle Ableitung nach μ wie folgt herleiten. Sei $\gamma_i \in \mathbb{R}^d$ definiert als $\gamma_i = x_i - V\lambda_i$ für jedes $1 \leq i \leq n$.

$$\begin{aligned}
\frac{\partial}{\partial \mu} \sum_{i=1}^{n} \|x_i - f(\lambda_i)\|^2 &= \frac{\partial}{\partial \mu} \sum_{i=1}^{n} \|x_i - \mu - V\lambda_i\|^2 \\
&= \frac{\partial}{\partial \mu} \sum_{i=1}^{n} \|\gamma_i - \mu\|^2 \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \mu} \langle \gamma_i - \mu, \gamma_i - \mu \rangle \\
&= \sum_{i=1}^{n} \left(\frac{\partial}{\partial \mu_1} (\gamma_{i,1} - \mu_1)^2, \ldots, \frac{\partial}{\partial \mu_d} (\gamma_{i,d} - \mu_d)^2 \right) \\
&= \sum_{i=1}^{n} (-2(\gamma_{i,1} - \mu_1), \ldots, -2(\gamma_{i,d} - \mu_d)) \\
&= \sum_{i=1}^{n} -2(\gamma_i - \mu) \\
&= \sum_{i=1}^{n} -2(x_i - \mu - V\lambda_i)
\end{aligned}$$

Setzen wir dies gleich dem Nullvektor, dann erhalten wir

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i - V \left(\frac{1}{n} \sum_{i=1}^{n} \lambda_i \right)$$

Sei $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. Setzen wir nun (1) ein, dann erhalten wir

$$\mu = \overline{x} - V \left(\frac{1}{n} \sum_{i=1}^{n} V^T (x_i - \mu) \right) = \overline{x} - VV^T (\overline{x} - \mu)$$

Das ist äquivalent zu

$$VV^T (\overline{x} - \mu) = \overline{x} - \mu$$

Wir können hier $\mu = \overline{x}$ wählen und diese Gleichung erfüllen, ohne dass die Wahl von V berührt ist.

Damit ergibt sich für unsere Zielfunktion

$$\phi(V) = \sum_{i=1}^{n} \|(x_i - \overline{x}) - VV^T (x_i - \overline{x})\|^2 \tag{2}$$

Wir können dies so interpretieren, dass wir eigentlich eine Funktion f für die zentrierte Menge $S' = \{x'_1, \ldots, x'_n\}$ mit $x'_i = x_i - \overline{x}$ finden wollen. Wir können vereinfachend annehmen, dass die Menge S schon zentriert ist. Dann ist $\overline{x}$ gleich dem Nullvektor und die optimale Hyperebene geht durch den Ursprung. In diesem Fall ist die Funktion f eine lineare Abbildung und bildet auf einen linearen Unterraum ab, die durch die Spaltenvektoren von V aufgespannt wird.

2 Beispiel

Wir wollen uns der Funktion f zunächst weiter anhand eines Beispiels nähern. Abbildung 1 zeigt eine zufällige Auswahl von Bildern einer handgeschriebenen Ziffer Drei, aus dem MNIST Datensatz. Jedes Bild ist durch einen hochdimensionalen Vektor von Pixelwerten gegeben. Ein Bild mit $h \times w$ Pixeln ist demnach ein Vektor im $\mathbb{R}^{h \cdot w}$. Wir wollen diesen Datensatz in der Parametrisierung einer 2-dimensionalen Hyperebene betrachten, welche die Zielfunktion ϕ minimiert. Das linke Bild zeigt den Vektor μ, also das Bild einer gemittelten handgeschriebenen

Ziffer Drei. Das mittlere Bild zeigt eine Darstellung des ersten Spaltenvektors v_1 der Matrix V, das rechte Bild zeigt eine Darstellung des zweiten Spaltenvektors v_2 der Matrix V. Beachte, dass der graue Hintergrund hier ein Artifakt der Darstellung ist. Die Pixelwerte sind in der Darstellung auf Grauwerte zwischen 0 und 1 abgebildet. Die hellen Pixel der Vektoren v_1 und v_2 sollten also als negative Werte interpretiert werden und dunkle Pixel als positive Werte.

Ein Punkt in der k-dimensionalen Hyperebene, die durch μ, v_1 und v_2 bestimmt ist, wird durch einen Parametervektor $\lambda = (t_1, t_2) \in \mathbb{R}^2$ als

$$f(t_1, t_2) = \mu + t_1 v_1 + t_2 v_2$$

dargestellt. Abbildung 2 zeigt das Ergebnis für eine Auswahl an Punkten im Parameterraum.

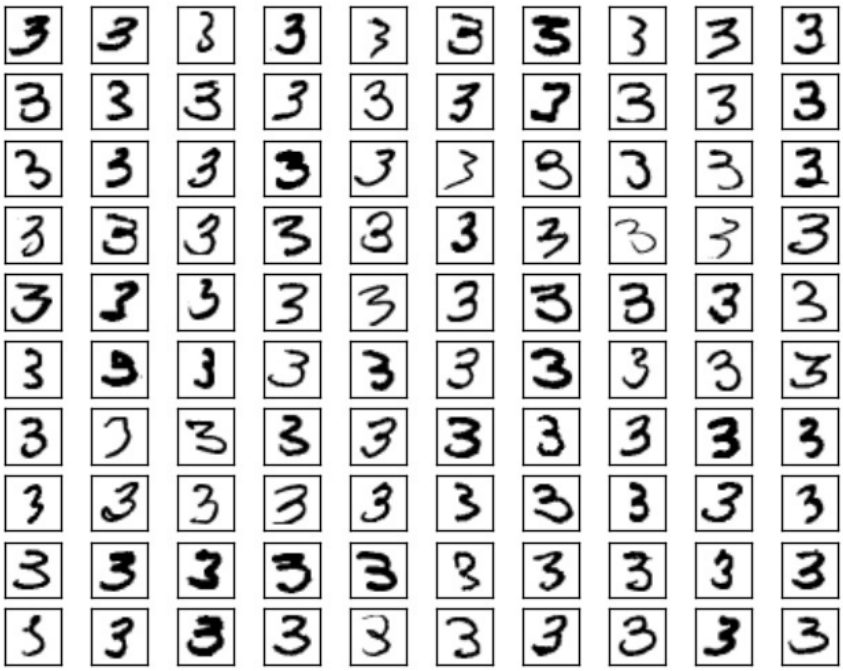


Abbildung 1: Zufällige Auswahl des MNIST-Datensatzes von Bildern von handgeschriebenen Ziffern. Hier ist eine Auswahl getroffen von Beispielen der Ziffer Drei.

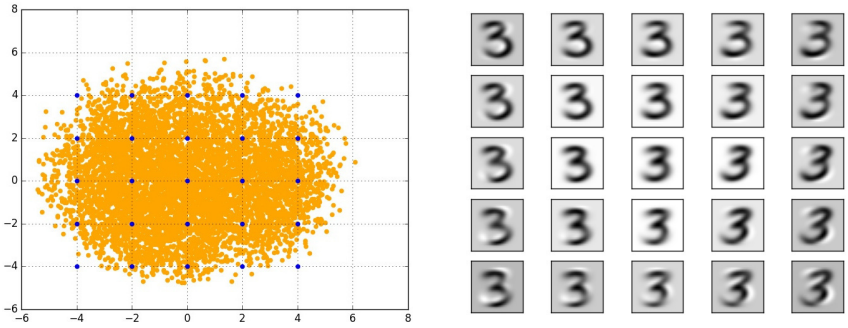


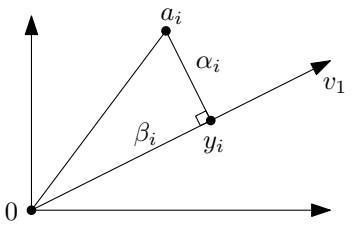
Abbildung 2: Links: Punktmenge (gelb) aus dem MNIST-Datensatz (nur Ziffer Drei) projiziert auf den Unterraum, der durch v_1 und v_2 gespannt wird. Rechts: Darstellung der Rekonstruktion durch die Funktion $f(t_1, t_2) = \mu + t_1 v_1 + t_2 v_2$ für die blauen Gitterpunkte (t_1, t_2) im Bild links.

3 Singulärwertzerlegung

Wir wollen eine Matrix V finden, welche die Zielfunktion ϕ in (2) minimiert. Dazu schreiben wir unsere Menge von Datenpunkten $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ in eine Matrix. Sei A eine $n \times d$ Matrix mit Zeilenvektoren $a_1, \ldots a_n$ mit $a_i = (x_i - \overline{x})$ mit $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ für alle $1 \leq i \leq n$.

Wir betrachten zunächst den Fall $k = 1$. In diesem Fall hat die Matrix V nur einen Spaltenvektor v_1. Dieser Spaltenvektor spannt einen 1-dimensionalen Unterraum, also eine Gerade durch den Ursprung, und wir betrachten die Projektionen der Eingabemenge auf diese Gerade.

Betrachte das Dreieck mit den Eckpunkten a_i, der Projektion $y_i = v_1 \langle v_1, a_i \rangle$, und dem Nullpunkt. Seien $\beta_i = \| v_1 \langle v_1, a_i \rangle \|$ und $\alpha_i = \| a_i - v_1 \langle v_1, a_i \rangle \|$, und $\|a_i\|$ die Seitenlängen dieses Dreiecks.



Es folgt aus dem Satz von Pythagoras, dass

$$\beta_i^2 = \|a_i\|^2 + \alpha_i^2$$

Damit ist $\alpha_i^2 = \|a_i\|^2 - \beta_i^2$. Wir suchen nach einem Vektor v_1 mit $\|v_1\| = 1$, sodass $\phi(v_1) = \sum_{i=1}^n \alpha_i^2$ minimiert wird. Durch Einsetzen der obigen Beobachtung erhalten wir

$$\underset{\substack{v_1 \in \mathbb{R}^d \\ \|v_1\|=1}}{\arg\min} \sum_{i=1}^n \alpha_i^2 = \underset{\substack{v_1 \in \mathbb{R}^d \\ \|v_1\|=1}}{\arg\min} \sum_{i=1}^n \|a_i\|^2 - \beta_i^2 = \underset{\substack{v_1 \in \mathbb{R}^d \\ \|v_1\|=1}}{\arg\max} \sum_{i=1}^n \beta_i^2$$

Um die Summe auf der rechten Seite noch weiter zu vereinfachen, beobachten wir, dass

$$\beta_i = \| v_1 \langle v_1, a_i \rangle \| = | \langle v_1, a_i \rangle |,$$

da $\|v_1\| = 1$ ist. Also ist

$$\sum_{i=1}^n \beta_i^2 = \sum_{i=1}^n | \langle v_1, a_i \rangle |^2 = \|Av_1\|^2$$

Das heißt, ϕ zu minimieren ist äquivalent dazu, $\|Av_1\|$ zu maximieren.

Angenommen, wir könnten v_1 bestimmen. Betrachte den folgenden Greedy-Algorithmus, der weitere Spaltenvektoren $v_2, \ldots, v_k$ der Matrix V unter dieser Annahme bestimmt.

```
Greedy-Algorithmus(n × d Matrix A)
  1. v_1 = arg max_{||v_1||=1} ||Av_1||
  2. σ_1 = ||Av_1||
  3. while σ_i ≠ 0 do
  4.     i = i + 1
  5.     v_i = arg max_{||v_i||=1, v_i ⊥ v_1,...v_{i-1}} ||Av_i||
  6.     σ_i = ||Av_i||
  7. Return v_1, ..., v_i
```

Man kann zeigen, dass der Algorithmus eine sogenannte Singulärwertzerlegung der Matrix A bestimmt. Allgemein besteht die Singulärwertzerlegung einer reellen Matrix A aus drei Matrizen U, D, V, mit

$$A = U \cdot D \cdot V^T$$

und mit den folgenden Eigenschaften der Matrizen

- U ist eine $n \times r$ Matrix mit orthonormalen Spaltenvektoren $u_1, \ldots, u_r$,
- V ist eine $d \times r$ Matrix mit orthonormalen Spaltenvektoren $v_1, \ldots, v_r$,
- D ist eine $r \times r$ Diagonalmatrix mit Einträgen $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$,

wobei r den Rang der Matrix A bezeichnet, das heißt r ist die maximale Anzahl linear unabhängiger Zeilenvektoren von A.

Wir nennen die Spaltenvektoren von V die *rechten Singulärvektoren*, die Spaltenvektoren von U die *linken Singulärvektoren* und die Werte $\sigma_1, \ldots, \sigma_r$ die *Singulärwerte*. Wir können die obige Gleichung schreiben als

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T,$$

Betrachten wir nur die Summe der ersten k Terme, dann erhalten wir eine Matrix

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

Die Zeilenvektoren von A_k entsprechen den Vektoren y_i in dem von V aufgespannten k-dimensionalen Unterraum, welche unsere Datenpunkte a_i approximieren sollen. Dadurch, dass die Singulärwerte ihrer Größe nach geordnet sind, wählen wir mit A_k genau die Terme aus, die am stärksten in die Summe eingehen.

Alternativ können die Vektoren $v_1, \ldots, v_k$ durch eine Eigendekomposition der Matrix $A^T A$ bestimmt werden. Dort würden wir die k Eigenvektoren mit den größten Eigenwerten auswählen. Die Darstellung der Datenpunkte im Unterraum der ersten k Eigenvektoren, bzw. Singulärvektoren, wird auch als Eigenkomponentenanalyse bezeichnet.

4 Potenzmethode

Wie kann man nun den Singulärvektor $\arg\max_{\|v_1\|=1} \|Av_1\|$ bestimmen? Dafür betrachten wir die sogenannte Potenzmethode. Die Methode hat ihren Namen daher, dass sie das Ergebnis bestimmt indem sie eine Matrix immer wieder mit sich selbst multipliziert, um eine hohe Potenz dieser Matrix zu berechnen.

Betrachte die Matrix $B = A^T \cdot A$. Sei $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ die Singulärwertzerlegung, wie oben definiert. Dann ist

$$A^T = \sum_{i=1}^{r} \sigma_i (u_i v_i^T)^T = \sum_{i=1}^{r} \sigma_i v_i u_i^T.$$

Also erhalten wir für B

$$\begin{aligned} B &= \left(\sum_{i=1}^{r} \sigma_i v_i u_i^T \right) \left(\sum_{j=1}^{r} \sigma_j u_j v_j^T \right) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{r} \sigma_i \sigma_j (v_i u_i^T)(u_j v_j^T) \\ &= \sum_{i=1}^{r} \sigma_i^2 v_i (u_i^T u_i) v_i^T + \sum_{i=1}^{r} \sum_{\substack{j=1 \\ i \neq j}}^{r} \sigma_i \sigma_j v_i (u_i^T u_j) v_j^T \end{aligned}$$

Da die Vektoren $u_1, \ldots, u_r$ orthonormal sind, gilt $u_i^T u_i = 1$ für $1 \leq i$ und $u_i^T u_j = 0$ für $i \neq j$. Daher folgt

$$B = \sum_{i=1}^{r} \sigma_i^2 v_i v_i^T$$

Betrachte nun die Matrix $B^2 = B \cdot B$.

$$\begin{aligned} B^2 &= \left(\sum_{i=1}^{r} \sigma_i^2 v_i v_i^T \right) \left(\sum_{j=1}^{r} \sigma_j^2 v_j v_j^T \right) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{r} \sigma_i \sigma_j (v_i v_i^T)(v_j v_j^T) \\ &= \sum_{i=1}^{r} \sigma_i^2 v_i (v_i^T v_i) v_i^T + \sum_{i=1}^{r} \sum_{\substack{j=1 \\ i \neq j}}^{r} \sigma_i \sigma_j v_i (v_i^T v_j) v_j^T \end{aligned}$$

Da die Vektoren $v_1, \ldots, v_r$ orthonormal sind, gilt $v_i^T v_i = 1$ für $1 \leq i$ und $v_i^T v_j = 0$ für $i \neq j$. Daher erhalten wir

$$B^2 = \sum_{i=1}^{r} \sigma_i^4 v_i v_i^T$$

Allgemein können wir damit für die $k$te Potenz von B herleiten, dass

$$B^k = \sum_{i=1}^{r} \sigma_i^{2k} v_i v_i^T$$

da der Term $(v_i^T v_i)$ immer gleich 1 ist und bei der Multiplikation stets wegfällt. Wenn $\sigma_1 > \sigma_2$, dann konvergiert B^k für große Werte von k gegen den ersten Term der Summe,

$$B^k \to \sigma_1^{2k} v_1 v_1^T$$

Das heißt, wir können v_1 bestimmen, indem wir einen Spaltenvektor von B^k normieren.

Referenzen

- Foundations of Machine Learning, Kapitel 15.1 und 15.3.1
- Understanding Machine Learning, Kapitel 23.1
- Avrim Blum, John Hopcroft, Ravindran Khannan, Foundations of Data Science, Kapitel 3
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, Elements of Statistical Learning