1 Dimensionsreduktion

Die Dimensionsreduktion ist ein effektives Verfahren zur Verminderung der Komplexität des Modells. Das Ziel ist also Merkmale herauszufiltern, die den kleinsten oder keinen Zusammenhang mit dem Endergebnis haben.. Dies kann durch verschiedene Verfahren passieren.

1.1 Hauptkomponentenanalyse

In der Hauptkomponentenanalyse werden die k-dimensionalen Daten x in einen d-dimensionalen Raum zu Daten y transformiert.
Das Ziel der Hauptkomponentenanalyse besteht darin die Varianz der Messdaten untereinander, sowie die Kovarianz auszuwerten. Die Varianz $\mathrm{Var}(x_i)$ gibt dabei Aufschluss über die Streuung der Daten, während die Kovarianz $\mathrm{Cov}(x_i, x_j)$ die lineare Abhängigkeit zwischen zwei Variablen misst. Eine hohe Varianz bedeutet eine große Streuung der Daten und damit einen höheren Einfluss dieses Features auf die Klassifizierung, während ein positiver Wert der Kovarianz einen linearen, ein negativer einen entgegengesetzten linearen und eine Null keinen statistischen linearen Zusammenhang aufweißt.

$$\mathrm{Var}(x^{(i)}) = \sum_{k=0}^{n} (x_k^{(i)} - \mu^{(i)})^2 \tag{1}$$

$$\mathrm{Cov}(x^{(i)}, x^j) = \sum_{k=0}^{n} (x_k^{(i)} - \mu^{(i)})(x_k^{(j)} - \mu^{(j)}) \tag{2}$$

Aus den Definitionen geht hervor:

$$\mathrm{Cov}(x^{(i)}, x^i) = \mathrm{Var}(x^{(i)}) \tag{3}$$

Ebenfalls gilt die Symmetrie:

$$\mathrm{Cov}(x^{(j)}, x^i) = \mathrm{Cov}(x^{(i)}, x^{(j)}) \tag{4}$$

Nun wollen wir die Komponente mit der maximalen Varianz finden. Dazu definieren wir einen normierten Vektor a, für den gilt:

$$\sum_{i=1}^{k} a_i^2 = 1 \tag{5}$$

Die Bedeutung dieses Vektors wird später deutlich. Nun wollen wir die Varianz der Daten maximieren und schauen, welche Vektorkomponenten von a zur maximalen Varianz führen. Die Funktion

$$\mathrm{Var}(a \cdot x) = \mathrm{Var}(\sum_{i=1}^{k} a_i x_i) \tag{6}$$

wird unter der Nebenbedingung:

$$\sum_{i=1}^{k} a_i^2 = 1 \tag{7}$$

optimiert. Dafür wird das Verfahren des Lagrange-Multiplikators angewandt. Die folgenden Bedingungen müssen dabei erfüllt sein:

$$\forall j : \frac{\partial}{\partial a_j} \mathrm{Var}(a \cdot x) = \lambda \frac{\partial}{\partial a_j} \sum_{i=1}^{k} a_i^2 \tag{8}$$

Durch die Ableitungsregel für Polynome gilt für die linke Seite:

$$\forall j : \frac{\partial}{\partial a_j} \mathrm{Var}(a \cdot x) = 2\lambda a_j \tag{9}$$

Für die rechte Seite gilt nach Einsetzen der Regel unter der Annahme bereits standartisierter Daten:

$$\forall j : \frac{\partial}{\partial a_j} \sum_{k=0}^{n} (a \cdot x)^2 = 2\lambda a_j \tag{10}$$

Durch die Kettenregel gilt also:

$$\forall j : 2x_j \sum_{k=0}^{n} (a \cdot x) = 2\lambda a_j \tag{11}$$

Dies kann man in die Klammer ziehen und auf beide Seiten durch 2 teilen.

$$\forall j : \sum_{k=0}^{n} x_j (a \cdot x) = \lambda a_j \tag{12}$$

Durch das Ausmultiplzieren der Klammer gilt:

$$\forall j : \sum_{k=0}^{n} \sum_{i=1}^{k} a_i x_i x_j = \lambda a_j \tag{13}$$

Dies lässt sich mit der Kovarianz zusammenfassen:

$$\forall j : \sum_{k=0}^{n} \sum_{i=1}^{k} a_i \operatorname{Var}(i x_i, x_j) = \lambda a_j \tag{14}$$

Diese Bedingungen entsprechen einer Eigenwertgleichung mit der Kovarianzmatrix C. Damit lassen sich die Gleichungen so schreiben:

$$C \cdot a = \lambda a \tag{15}$$

Dies hat zur Folge, dass keiner der Eigenvektoren a_i in die gleiche Richtung wie ein anderer Eigenvektor a_j zeigt, da alle Eigenvektoren orthogonal zueinander sind. Nun gilt ebenfalls für die Varianz mit dem Ausmultiplizieren der Klammer:

$$\operatorname{Var}(a \cdot x) = a^T \cdot C \cdot a \tag{16}$$

Durch die obige Gleichung gilt für die Kovarianzmatrix:

$$\operatorname{Var}(a \cdot x) = a^T \lambda a \tag{17}$$

Aus der Vektormultipliaktion gilt:

$$\operatorname{Var}(a \cdot x) = \lambda \sum_{i=1}^{k} a_i^2 \tag{18}$$

Durch die Randbedingung des Vektors kann dies vereinfacht werden.

$$\operatorname{Var}(a_n \cdot x) = \lambda_n \tag{19}$$

Somit sind die Eigenwerte dieser Vektoren also auch die Varianz der Daten in der Richtung des Vektors a. Damit können also durch die obige Eigenwertgleichung bestimmt werden, welche Richtung der Daten am meisten Varianz hat und damit die meisten Informationen birgt. Der Gesamtanteil der Varianz einer Richtung ergibt sich durch

$$p = \frac{\lambda_n}{\sum_{i=1}^{k} \lambda_i} \tag{20}$$

Nün können die d Eigenvektoren mit der größten Varianz zur Kosntruktion einer Matrix genutzt werden, die die Daten in einen kleineren Raum transformiert, indem aus den d Vektorena_i mit den größten Eigenwerten eine Matrix A der Dimensionen $d \times k$ zu machen. Für diese Matrix gilt:

$$X' = A \cdot X \tag{21}$$

Dabei ist A die Zusammensetzung aus den Eigenvektoren mit dem größtem Eigenwert. Für jeden Eintrag gilt also, wenn l die Ordnung der Eigenwerte darstellt und P der Eigenvektor.

$$A_{l,n} = a_n \in P_l \tag{22}$$

Diese Transformation der Daten optimiert also die Informationen nach Informationsgehalt, also der Varianz.

2 Lineare Diskriminanzanalyse

Die lineare Diskriminanzanalyse funktioniert nach dem ähnlichen Prinzip, wie die Hauptkomponentenanalyse, allerdings mit einer anderen Matrix und dem Ziel die separierendsten Merkmale herauszufiltern. Dabei werden die Punkte ebenfalls auf diese Hyperebene transformiert um die Varianz zu minimieren. Die Matrizen sind hierbei die Streumatrix S_m der einzelnen Klassem und die Streumatrix S_w der Klassen untereinander, anstatt der Kovarianzmatrix. Das Herausfinden der Achsen mit dem höchsten Informationsgehalt wird wieder durch die Eigenwertgleichung

$$v_n \cdot S_w \cdot S_m^{-1} = \lambda \cdot v_n \tag{23}$$

erreicht. Die Maximierung der Funktion ergibt sich durch die Maximierung der Streuung zwische den Klassen und die Minimierung der Streuung der Daten in den Klassen. Die Streuung der Daten ist ein Maß für de Informationsgehalt der Daten. Die Maximierung der Streuung zwischen den Datenklassen und die Minimierung der Einzelsteuung maximieren dabei die Abstände zwischen den Klassen und machen sie somit spearierbarer, indem gleiche Klassen eher zusammenrücken und unterschiedliche Klassen asueinander rücken, was sich in der Streuung niederschlägt.
Durch das selbige Verfahren der Optimierung durch das Verfahren von Lagrange wird hier die Eigenwertgleichung hergeleitet. Dabei ist S_m^-1 die inverse Matrix der Streuung unter den Klassen.

3 Kernel-Hauptkomponente für nichtlineare Zuordnungen

Die Kernel-Hauptkomponente funktioniert ebenfalls wie die PCA, transformiert aber die Merkmale in einen höheren Merkmalsraum um dort die PCA durchzuführen. Um nun die Ähnlichkeit zweier Merkmalsvektoren in einem höherdimensionalen Raum festzustellen wird hier der Kernel-Trick angewandt.
Der Kerneltrick ersetzt das Skalarprodukt zweier Merkmalsvektoren durch eine Kernelfunktion, die sich die Transformation spart.

$$k(x^{(i)}, x^{(j)}) = \phi(x^{(i)})\phi(x^{(j)}) \tag{24}$$

Dadurch lässt sich die Varianz neu berechnen.

$$\mathrm{Cov}(x^{(i)}, x^{(j)}) = \sum_{n=1}^{k} \phi(x_n^{(i)})\phi(x_n^{(j)}) \tag{25}$$

Durch Einsetzen in die Kovarianzmatrix und das Einsetzen ergibt sich:

$$\frac{1}{k}\phi(X)\phi(x)^T \cdot a = \lambda a \tag{26}$$

Dies ergibt die Kernelmatrix K:

$$\frac{1}{k}K \cdot a = \lambda a \tag{27}$$

Es gibt viele verschiedene Kernelfunktionen, allerdings ist eine sehr geläufige die des gaußschen Kernels, da die Funktion radialsymmetrisch ist und als Ähnlichkeitsmaß nur Werte zwischen 0 und 1 annehmen kann.

$$k(x^{(i)}, x^{(j)}) = e^{-\frac{||x^{(i)} - x^{(j)}||^2}{2\sigma^2}} \tag{28}$$

Die Kovarianz- oder die Streuungsmatrix sind hier Werte der Zusammenhänge zwischen zwei konkreten Merkmalsdatensätzen. Durch den Kerneltrick lässt sich dieses Ähnlichkeitsmaß direkt mit der Kernelfunktion erstellen. Dabei wird die Ähnlichkeitsmatrix oder Kernelmatrix K mit dem jeweiligen Wert der Kernelfunktion belegt.

$$K_{i,j} = k(x^{(i)}, x^{(j)}) \tag{29}$$

Nun wird das gleiche Verfahren wie bei der PCA angewandt. Dabei kommen, wie mit den größten Eigenvektoren, die Werte der größten Ähnlichkeit

hervor. Dabei geben die Eigenvektoren wieder die Achsen an, allerdings nicht die Achsen der Originaldaten, sondern die Achsen der projezierten Daten. Nun kann die PCA wieder angewandt werden. Dieses Verfahren wird auch oft angewandt, wenn die ursprüngliche Testdatenmenge keine lineare Separabilität hat, nach einer zusätzlichen Dimension allerdings schon. Die Transformation neuer Daten erfolgt mit der Formel:

$$\phi(X)^t a = v \tag{30}$$

Dies lässt sich durch den Kerneltrick weiter vereinfachen.

$$v = \sum_{i=1}^{k} a^{(i)} \phi(X')^T \phi(X) \tag{31}$$

$$v = \sum_{i=1}^{k} a^{(i)} k(X, X^T) \tag{32}$$