AGML, Sommersemester 2020 Vorlesung 17 (7 Seiten)

Nächste Nachbarn

Anne Driemel Letzte Aktualisierung: 25. Juni 2020

Ein grundlegender Lernalgorithmus im Maschinellen Lernen ist der Nächste-Nachbarn-Algorithmus. Die Idee ist sehr einfach. Um einen Punkt $q \in X$ auf Basis einer Trainingsmenge $S \subseteq X \times \{-1, +1\}$ zu klassifizieren, berechnen wir den Punkt in S, der q am ähnlichsten ist und geben das entsprechende Label zurück. Dafür müssen wir die Ähnlichkeit zunächst definieren. Einfacher ist es meist, den Punkt zu betrachten, der den geringsten Abstand unter einem bestimmten Distanzmaß hat. Wir betrachten hier zunächst den Euklidischen Abstand. Unsere Hypothese ist also die folgende Funktion $h_S : X \to \{+1, -1\}$ definiert durch

$$h_S(x) = y_i \qquad \text{mit} \qquad i = \arg\min_{1 \le i \le m} \|x - x_i\|$$

In diesem Kontext bezeichnen wir x_i als den nächsten Nachbarn von x in S.

Diese einfache Variante der Nächste-Nachbarn Hypothese leidet unter dem Problem des Overfittings. Um dem entgegen zu wirken, werden oft die Labels der k nächsten Nachbarn betrachtet, wobei $k \in \mathbb{N}$ ein Parameter ist. Formal können wir die resultierende Hypothese wie folgt definieren. Für ein $x \in X$ sei $\pi_x : \{1, \dots, m\} \to \{1, \dots, m\}$ eine Bijektion der Menge S auf sich selbst, sodass für alle $i, j \in \{1, \dots, m\}$ gilt

$$\pi_x(i) \le \pi_x(j) \qquad \Longrightarrow \qquad \|x - x_i\| \le \|x - x_j\|$$

Das heißt, π_x stellt eine Permutation der Menge S dar, welche einer aufsteigend sortierten Reihenfolge bezüglich des Abstands zu x entspricht. [1]

Sei $\mathcal{N}_k(x)$ die Indexmenge der k nächsten Nachbarn von x in S. Formal,

$$\mathcal{N}_k(x) = \left\{ \ \pi_x^{-1}(i) \ \middle| \ 1 \le i \le k \ \right\}$$

Die k-NN Hypothese ist die Funktion $h_{S,k} : X \to \{+1, -1\}$ definiert durch

$$h_{S,k}(x) = \arg\max_{\ell \in \{+1,-1\}} \left| \{ \ j \in \mathcal{N}_k(x) \ \mid \ y_j = \ell \ \} \right|$$

Wir bezeichnen das algorithmische Problem, die k nächsten Nachbarn in einer Menge zu finden als das k-NN Problem.

Obwohl wir immer noch von Hypothesen sprechen, macht es hier keinen Sinn, die VC-Dimension der entsprechenden Hypothesenklasse zu betrachten. Wir würden dann feststellen, dass die VC-Dimension von der Größe von S abhängt und hätten dann keine Möglichkeit mehr, im Rahmen des PAC-Frameworks, die minimale Größe von S anhand der VC-Dimension festzulegen. Nichtsdestotrotz bildet die Klasse der Lernalgorithmen, die auf dem Prinzip der nächsten Nachbarn basiert, eine grundlegende Methode im Maschinellen Lernen.

[1] Beachte, dass π_x dadurch noch nicht eindeutig definiert ist, da es nicht für jedes x eine eindeutige Permutation der nächsten Nachbarn gibt. Wir definieren deshalb ausserdem die folgende Bedingung, welche die Permutation eindeutig macht.

$$\pi_x(i) < \pi_x(j) \text{ und } \|x - x_i\| = \|x - x_j\| \implies i < j$$

1 Voronoi-Diagramme

Für eine feste Menge S lässt sich die Hypothese h_S (bzw. $h_{S,k}$) durch ein sogenanntes Voronoi-Diagram darstellen. Bei der Hypothese $h_{S,k}$ sprechen wir dann von einem Voronoi-Diagramm der k-ten Ordnung.

Definition 17.1. *Sei $S \subseteq \mathbb{R}^d$ mit $|S| = m$. Sei $k \leq m$ eine natürliche Zahl. Die Voronoi-Region einer Menge $A \subseteq \{1, \ldots, m\}$ mit $|A| = k$ ist die Menge*

$$\mathcal{V}_k(A) = \left\{ x \in \mathbb{R}^d \;\middle|\; \mathcal{N}_k(x) = A \right\}$$

Das Voronoi-Diagramm ist die Unterteilung des Raumes $\mathbb{R}^d$ in die Voronoi-Regionen für alle $A \subseteq \{1, \ldots, m\}$ mit $|A| = k$.

Das Voronoi-Diagram ist also die Unterteilung der Grundmenge in genau die Regionen, für die die Ausgabe des k-NN Problems gleich ist. Jede Strukturierung der Trainingsmenge, die einen effiziente Beantwortung der Frage nach den k nächsten Nachbarn von einem Anfragepunkt x erlaubt, beantwortet implizit die Frage, in welcher Voronoi-Region sich x befindet. Wir interessieren uns deshalb für die Struktur des Voronoi-Diagramms und insbesondere die Komplexität des Diagramms. Wir werden feststellen, dass das Voronoi-Diagram für $k = 1$ und $d = 2$ eine überraschend einfache Struktur hat.

1.1 k-NN auf der Geraden

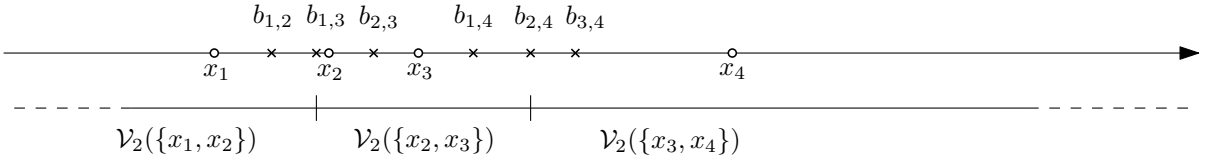
Für $d = 1$ betrachten wir das arithmetische Mittel zwischen zwei Punkten der Trainingsmenge, $b_{i,j} = \frac{x_i + x_j}{2}$. Der Wert $b_{i,j}$ unterteilt die Grundmenge in zwei disjunkte Intervalle

$$I_- = (-\infty, b_{i,j}) \qquad \text{und} \qquad I_+ = (b_{i,j}, \infty)$$

Dabei gilt für ein beliebiges Paar von Punkten $x, x' \in \mathbb{R} \setminus \{b_{i,j}\}$, dass sie genau dann demselben Intervall angehören, wenn sie in der Menge $\{x_i, x_j\}$ denselben nächsten Nachbarn haben.

Allgemeiner, können wir die Werte $b_{i,j}$ der Menge $\binom{S}{2}$ betrachten, also der Menge aller Punktepaare aus S. Diese unterteilen die Grundmenge $\mathbb{R}$ in eine beschränkte Anzahl von Intervallen, sodass in jedem Intervall die Permutation π_x für alle Punkte x in dem Intervall gleich ist. Im Voronoi-Diagram der k-ten Ordnung fassen wir all jene Intervalle zu einer Menge zusammen, bei der die k nächsten Nachbarn, also die Menge $\mathcal{N}_k(x)$, gleich sind.

Beispiel 17.2. *Sei $k = 2$ und seien $x_1, x_2, x_3, x_4 \in \mathbb{R}$ wie folgt*



Für $k = 2$ haben wir in diesem Beispiel die folgenden nicht-leeren Voronoi-Regionen:
$\mathcal{V}_2(\{x_1, x_2\}) = (-\infty, b_{1,3}]$ $\mathcal{V}_2(\{x_2, x_3\}) = (b_{1,3}, b_{2,4}]$ $\mathcal{V}_2(\{x_3, x_4\}) = (b_{2,4}, \infty)$.

Man kann zeigen, dass das Voronoi-Diagramm von m Punkten in $\mathbb{R}$ aus genau $m - k + 1$ nicht-leeren Voronoi-Regionen besteht, die jeweils ein zusammenhängendes Intervall bilden. Es hat also höchstens lineare Komplexität. Für $d = 2$ kann man allerdings Punktmengen finden, für

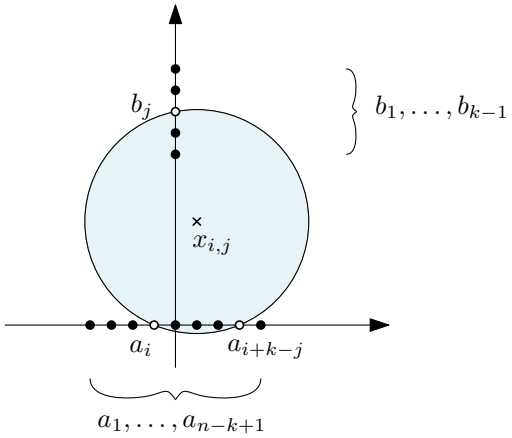


Abbildung 1: Es gibt Punktmengen mit mindestens $(k-1)(m-2k)$ nicht-leeren Voronoi-Regionen in der Ebene.

die das Voronoi-Diagramm der k-ten Ordnung mindestens $(m-2k)(k-1)$ nicht-leere Voronoi-Regionen enthält. Es hat also im schlimmsten Fall mindestens quadratische Komplexität. Im Beispiel in Abbildung 1 gibt es Punkte $a_1, \ldots, a_{m-k+1}$ auf der x-Achse und $k-1$ Punkte auf der y-Achse, die so gewählt sind, dass für jede Koembination von Indizes $(i,j) \in \{1, \ldots, m-2k\} \times \{1, \ldots, k-1\}$ ein Kreis existiert, der genau die Punkte $A_{i,j} = \{b_1, \ldots, b_j\} \cup \{a_i, \ldots, a_{i+k-j}\}$ enthält. Der Mittelpunkt dieses Kreises ist also enthalten in der Voronoi-Region $\mathcal{V}_k(A_{i,j})$. Das bedeutet, dass diese Voronoi-Region nicht leer ist. Also gibt es mindestens $(m-2k)(k-1)$ nicht-leere Voronoi-Regionen.

1.2 1-NN in der Ebene

Für den Fall $k = 1$ hat das Voronoi-Diagram eine überraschend einfache geometrische Struktur. Die Punkte $b_{i,j}$, an denen sich die Permutation der nächsten Nachbarn für $d = 1$ ändert, können wir verallgemeinern zu dem Bisektor, der wie folgt definiert ist.

Definition 17.3. *Der Bisektor $B(p,q)$ zwischen zwei Punkten $p \in \mathbb{R}^d$ und $q \in \mathbb{R}^d$ ist die Menge*

$$B(p,q) = \left\{ x \in \mathbb{R}^d \;\middle|\; \|p-x\| = \|q-x\| \right\}$$

Der Bisektor enthält alle Punkte, für die der Abstand zum Punkt p und der Abstand zum Punkt q genau gleich ist. Für feste p und q ist der Bisektor eine Hyperebene, wie sich leicht überprüfen lässt:

$$\begin{aligned}
& \|p-x\| = \|q-x\| \\
\Leftrightarrow \quad & \|p-x\|^2 = \|q-x\|^2 \\
\Leftrightarrow \quad & \langle p-x, p-x\rangle = \langle q-x, q-x\rangle \\
\Leftrightarrow \quad & \langle p,p\rangle + \langle x,x\rangle - 2\langle p,x\rangle = \langle q,q\rangle + \langle x,x\rangle - 2\langle q,x\rangle \\
\Leftrightarrow \quad & \langle p,p\rangle - 2\langle p,x\rangle = \langle q,q\rangle - 2\langle q,x\rangle \\
\Leftrightarrow \quad & 2\langle q,x\rangle - 2\langle p,x\rangle = \langle q,q\rangle - \langle p,p\rangle \\
\Leftrightarrow \quad & \langle 2(q-p), x\rangle = \langle q,q\rangle - \langle p,p\rangle \\
\Leftrightarrow \quad & \langle w_{p,q}, x\rangle = u_{p,q}
\end{aligned}$$

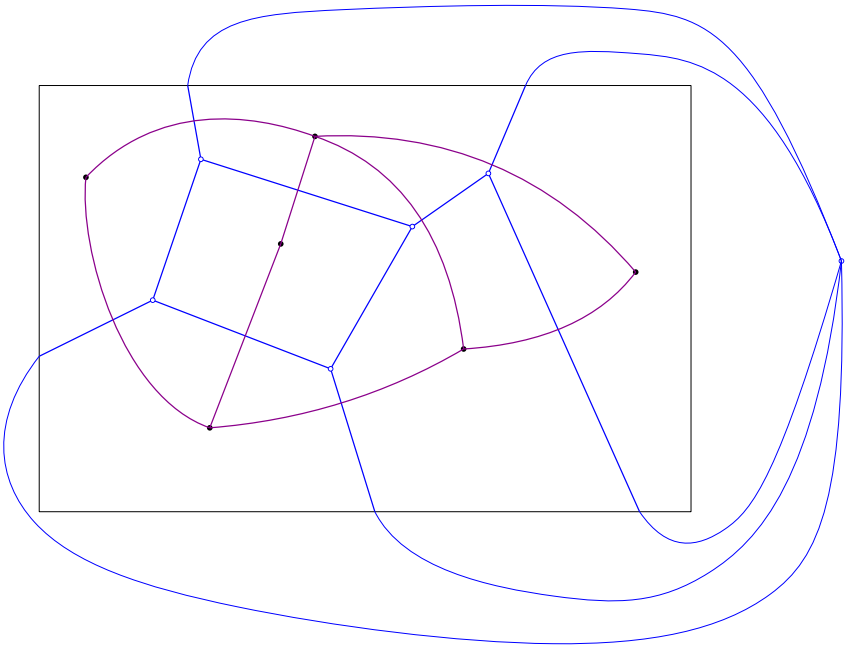


Abbildung 2: Im Kasten sieht man ein Aussschnitt des Voronoi-Diagramms der schwarzen Punkte ($k = 1$). Der blaue Knoten ist der virtuelle Knoten, der alle unbeschränkten Kanten verbindet. Die lila Kanten sind Kanten des dualen Graphen.

mit $w_{p,q} = 2(q - p) \in \mathbb{R}^d$ und $u_{p,q} = \langle q, q\rangle - \langle p, p\rangle \in \mathbb{R}$.

Der Bisektor unterteilt die Grundmenge in zwei offene Halbräume.

$$H_-(p,q) = \left\{ x \in \mathbb{R}^d \;\middle|\; \langle w_{p,q}, x\rangle < u_{p,q} \right\} \text{ und } H_+(p,q) = \left\{ x \in \mathbb{R}^d \;\middle|\; \langle w_{p,q}, x\rangle > u_{p,q} \right\}$$

Dabei gilt für ein beliebiges Paar von Punkten $x, x' \in \mathbb{R}^d \backslash B(p,q)$, dass sie genau dann demselben Halbraum angehören, wenn sie in der Menge $\{p, q\}$ denselben eindeutigen nächsten Nachbarn haben.

Die Voronoi-Region eines Punktes x_i in der Menge $S = \{x_1, \ldots, x_m\}$ ist die Menge der Punkte, für die x_i der eindeutige nächste Nachbar ist.[2]

$$\mathcal{V}_1(x_i) = \bigcap_{\substack{1 \le j \le m \\ i \neq j}} H_-(x_i, x_j)$$

Die Voronoi-Region ist also eine zusammenhängende Menge. Das folgt direkt aus der Konvexität der Halbräume und daraus, dass die Konvexität von Mengen unter endlichen Schnitten abgeschlossen ist.

Die Grenzen der Voronoi-Regionen bestehen aus Teilen der Bisektoren. In der Ebene formen diese zusammen einen Graphen mit Knoten und Kanten. Jeder Punkt auf einer Kante hat dabei den gleichen Abstand zu seinen zwei nächste Nachbarn. Jeder Punkt auf einem Knoten hat den gleichen Abstand zu seinen drei nächsten Nachbarn. Wir können die Anzahl der Knoten und Kanten im Voronoi-Diagramm wie folgt beschränken.

Satz 17.4. *Das Voronoi-Diagramm von m Punkten in $\mathbb{R}^2$ hat $O(m)$ Knoten und Kanten.*

[2]Mathematisch ist das nicht ganz korrekt, da wir die Voronoi-Regionen etwas anders definiert haben. Die Mengen unterscheiden sich aber nur am Rand. Wir sehen darüber um einer einfacheren Definition willen hinweg.

Beweis. Wir nutzen Eulers Formel für planare Graphen. Für einen Graphen G mit v Knoten, e Kanten und f Flächen besagt sie, dass

$$v - e + f = 2$$

Wir wollen diese Formel auf den Graphen der die Voronoi-Regionen begrenzt anwenden. Dafür müssen wir einen virtuellen Knoten hinzufügen, der mit allen unbeschränkten Kanten verbunden ist.[3] Wir wissen, dass $f = m$, da f die Flächen des Graphen mit den Voronoi-Regionen korrespondieren. Sei d_i die Anzahl der Kanten, die inzident zum $i$ten Voronoi-Knoten sind. Wir können die Summe der Knotengrade auf zwei Arten begrenzten,

$$2e = \sum_{i=1}^{v} d_i \geq 3v$$

da jede Voronoi-Kante zu genau 2 Voronoi-Knoten inzident ist, und da jeder Voronoi-Knoten zu mindestens zu 3 Voronoi-Kanten inzident ist. Wir nehmen hier an, dass $m > 2$, sonst ist die Aussage im Satz trivial erfüllt. Daraus folgt $v \leq \frac{2}{3}e$ und daher folgt aus Eulers Formel

$$e = f + v - 2 \leq m + \frac{2}{3}e - 2$$

Dies können wir umformen zu

$$e \leq 3(m - 2)$$

Also ist $e \in O(m)$. Daraus folgt auch, da $v \leq \frac{2}{3}e$, dass $v \in O(m)$. □

1.3 k-NN in der Ebene

Für $k > 1$ können wir und das Voronoi Diagram höherer Ordnung wie folgt vorstellen. Für jede Region $\mathcal{V}_1(x_i)$ im Voronoi-Diagram von S betrachten wir das Voronoi-Diagram von $S \setminus \{x_i\}$ beschränkt auf die Region $\mathcal{V}_1(x_i)$. Das gibt uns die Regionen $\mathcal{V}_2(\{x_i, x_j\}) \cap \mathcal{V}_1(x_i)$ für alle $i \neq j$. Das können wir rekursiv fortführen um weitere Voronoi-Diagramme höherer Ordnung für $k > 2$ zu finden. Allgemein kann man beobachten, dass die Voronoi-Regionen höherer Ordnung immer von Teilen der Bisektoren der Menge $\binom{S}{2}$ begrenzt werden. Insbesondere teilen die Bisektoren die Ebene in Regionen, sodass in jeder Region die Permutation der nächsten Nachbarn gleich ist.

1.4 Voronoi-Diagramme in höheren Dimensionen

In höheren Dimension steigt die Komplexität des Voronoi-Diagramms exponentiell mit der Dimension. Für $d = 3$ kann das Voronoi-Diagram schon quadratische Größe haben. Dafür konstruieren wir zwei windschiefe Geraden g_A und g_B, also zwei Geraden die nicht in derselben Ebene liegen. Sei $A = \{a_1, \ldots, a_n\}$ eine Menge von $n = \lceil \frac{m}{2} \rceil$ Punkten auf g_A und sei $B = \{b_1, \ldots, b'_n\}$ eine Menge von $n' = \lfloor \frac{m}{2} \rfloor$ Punkten auf g_B. Wir nehmen an, dass zwischen zwei Punkten a_i und a_{i+1} kein weiterer Punkt aus A auf g_A liegt, und ähnlich nehmen wir an, dass zwischen zwei Punkten b_i und b_{i+1} kein weiterer Punkt aus B auf g_B liegt. Nun können wir für jedes Tupel $(i, j) \in \{1, \ldots, n-1\} \times \{1, \ldots, n'-1\}$ die Kugel betrachten, die a_i, a_{i+1}, b_j und b_{j+1} auf dem Rand hat. Da die beiden Geraden windschief sind, liegen die vier Punkte nicht in einer Ebene und bestimmen somit eindeutig eine Kugel. Die Kugel enthält keine weiteren Punkte aus $A \cup B$. Daher ist das Zentrum der Kugel ein Knoten im Voronoi-Diagram von $A \cup B$. Daraus folgt, dass das Voronoi-Diagramm mindestens $(n-1)(n'-1) \in \Omega(m^2)$ Knoten hat.

[3] Wir könnten stattdessen auch den dualen Graphen betrachten, welcher auch ein planarer Graph ist. Dieser ist in Abbildung 2 abgebildet. Der virtuelle Knoten entspricht dann der äußeren Fläche.

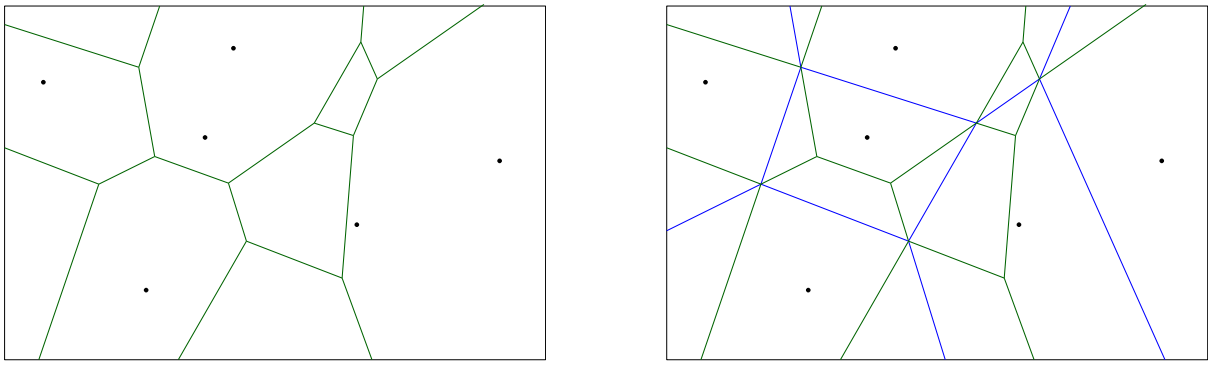


Abbildung 3: Links: Voronoi-Diagramm der zweiten Ordnung für die Punktmenge aus Abbildung 2; Rechts: Voronoi-Diagramme für $k = 1$ und $k = 2$ übereinander gezeichnet.

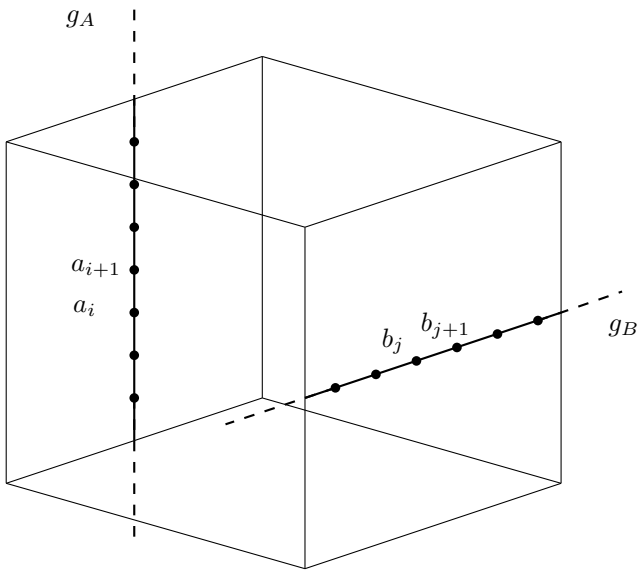


Abbildung 4: Beispiel einer Konstruktion einer Menge von m Punkten in $\mathbb{R}^3$ mit mindestens $\Omega(n^2)$ vielen Voronoi-Knoten.

Allgemein, im $\mathbb{R}^d$ ist die Anzahl der Knoten des Voronoi-Diagramms von m Punkten in $\Theta(m^{\lceil \frac{d}{2} \rceil})$ im schlimmsten Fall. Die Komplexität von Voronoi-Diagrammen höherer Ordnung im $\mathbb{R}^d$ ist nicht genau bekannt. Es ist aber zu vermuten, dass diese noch höher ist, als für $k = 1$.

Aus diesem Grund werden in höheren Dimensionen die k nächsten Nachbarn nicht durch die explizite Berechnung und Vorverarbeitung des Voronoi-Diagramms bestimmt. Alternativ können alle Abstände zu der Menge S explizit berechnet werden, was eine lange Klassifizierungslaufzeit hat. Eine andere Möglichkeit ist es, die nächsten Nachbarn approximativ zu bestimmen. Damit werden wir uns in der nächsten Vorlesung beschäftigen.

Referenzen

- Understanding Machine Learning, Kapitel 19.
- Rolf Klein, Algorithmische Geometrie, Springer, 1996, (Kapitel 5).