AGML, Sommersemester 2020 Vorlesung 18 (6 Seiten)

Approximative Nächste Nachbarn

Anne Driemel Letzte Aktualisierung: 1. Juli 2020

In der letzten Vorlesung ging es um eine Klasse von Lernalgorithmen, die auf dem Prinzip der nächsten Nachbarn basiert. Zur Erinnerung, die grundlegende Variante dieser Lernalgorithms nutzt die folgende Hypothese $h_S : X \to \{+1, -1\}$ definiert für eine Trainingsmenge $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in \mathbb{R}^d \times \{-1, +1\}$ durch

$$h_S(x) = y_i \qquad \text{mit} \qquad i = \operatorname*{arg\,min}_{1 \le i \le m} d(x, x_i),$$

wobei $d : X \times X \to \mathbb{R}_{\ge 0}$ eine Abstandsfunktion definiert. Wir bezeichnen x_i als den nächsten Nachbarn von x in S. Für den Euklidischen Abstand hatten wir das algorithmische Problem, den nächsten Nachbarn zu bestimmen, mithilfe der Voronoi-Diagramme analysiert. Das Voronoi-Diagramm beschreibt im Grunde die inverse Funktion der Hypothese h_S. Da die Komplexität eines Voronoi-Diagramms im schlimmsten Fall exponentiell mit der Dimension d wächst bieten sie leider keine effiziente algorithmische Lösung des Problems an. Es bleibt uns scheinbar nur die Möglichkeit, die Abstände zu allen m Elementen der Trainingsmenge zu berechnen, um die Funktion h_S an einem Punkt x zu evaluieren. Dies wird auch als *lineare Suche* bezeichnet, da die Laufzeit in $O(dm)$ ist.

Wir wollen heute eine approximative Variante dieses Problems betrachten. Das Ziel ist es, auf der Menge S eine Datenstruktur zu berechnen, welche eine effizientere Klassifizierung zulässt, also eine Klassifizierungslaufzeit besser als die der linearen Suche, wobei wir trotzdem noch dem Prinzip der nächsten Nachbarn treu bleiben wollen.

1 Lokalitätssensitive Funktionen

Definition 18.1. *Sei $\mathcal{F}$ eine Klasse von Funktionen der Form $h : X \to U$, wobei auf U eine Ordnungsrelation $\le$ definiert ist, und sei $d : X \times X \to \mathbb{R}_{\ge 0}$ eine Abstandsfunktion. $\mathcal{F}$ ist (r, R, α, β)-lokalitätssensitiv bezüglich der Funktion d, wenn für $x, y \in X$*

$$\mathbf{Pr}_{h \in \mathcal{F}}[h(x) = h(y)] > \alpha \quad \textit{falls} \quad d(x, y) < r \tag{1}$$

$$\mathbf{Pr}_{h \in \mathcal{F}}[h(x) = h(y)] < \beta \quad \textit{falls} \quad d(x, y) > R \tag{2}$$

Wir sagen, dass eine Klasse von Funktionen lokalitätssensitiv ist, wenn sie (r, R, α, β)-lokalitätssensitiv ist für ein $\alpha > 0$, ein $\beta < 1$ und $r, R > 0$ mit $r \le R$.

Idealerweise wollen wir, dass α möglichst groß ist, dass β möglichst klein ist und dass R/r möglichst nah bei 1 ist. Die Intuition dahinter ist, dass zwei Punkte, die nah beieinander liegen dann eine hohe Wahrscheinlichkeit haben, durch ein zufälliges h auf denselben Schlüssel abgebildet zu werden, während Punkte, die weit entfernt voneinander entfernt liegen eine niedrige Wahrscheinlichkeit haben, durch ein zufälliges h auf denselben Schlüssel abgebildet zu werden.

Lokalitätssensitive Funktionen erlauben es uns, bekannte Suchstrukturen, wie zum Beispiel Suchbäume, oder Hashing, auf das Nächste-Nachbarn-Problem in höheren Dimensionen anzuwenden. Sei D solch eine Suchstruktur. Wir können dann eine lokalitätssensitive Funktion h zufällig aus der Klasse $\mathcal{F}$ wählen und die Schlüssel $z_i = h(x_i)$ für jeden Punkt $(x_i, y_i) \in S$ aus der Trainingsmenge erzeugen. Die Datensätze (x_i, y_i) speichern wir dann mit dem zugehörigen Schlüssel in der Suchstruktur D. Um den nächsten Nachbarn eines Punktes $y \in X$ in S zu

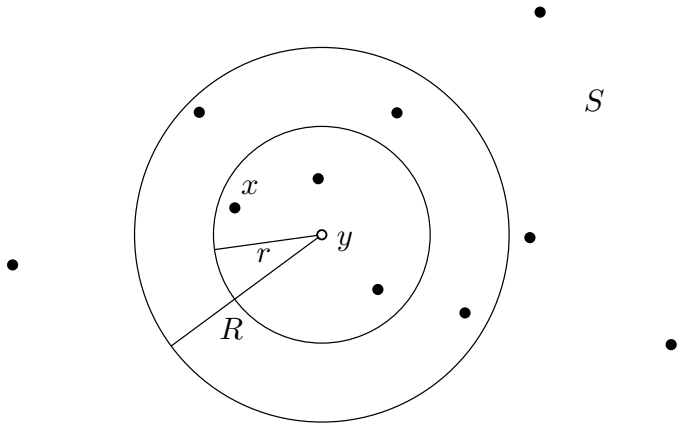

Abbildung 1: Schematische Darstellung der Unterteilung der Punktmenge S anhand der Abstände zu y. Für die Punkte, deren Abstand zwischen r und R liegt, wird in Definition 18.1 kein Aussage gemacht. Sie fallen in den Bereich des Approximationsfehlers.

finden, erzeugen wir den Schlüssel $z = h(y)$ und suchen in D nach z. Dafür ist es wichtig, dass auf der Schlüsselmenge eine Ordnungsrelation existiert, die es erlaubt, die Schlüssel zu sortieren. Beachte, dass die lokalitätssensitiven Funktionen hier nur eine Entscheidungsvariante des Nächste-Nachbarn-Problems lösen, da die Abstandsparameter r und R fest sind. Für den Euklidischen Abstand lässt sich dies durch Skalierung der Punktmenge auf alle anderen Abstandsparameter erweitern.

Insgesamt erinnern lokalitätssensitive Funktionen stark an das Hashing. Sie sollten aber nicht damit verwechselt werden. Beim Hashing geht es darum, ein Universum U auf eine kleine Indexmenge $\{1, \dots, m\}$ abzubilden. Das Ziel ist, einen Datensatz (Teilmenge des Universums) in einem Array der Größe m abzuspeichern und konstante Zugriffszeit auf die Elemente des Datensatzes zu erreichen. Eine sogenannte Hash-Kollision tritt dann auf, wenn zwei verschiedene Elemente im Datensatz auf denselben Schlüssel abgebildet werden. Um mehrere Elemente unter demselben Schlüssel zu speichern, können zusätzliche verkettete Listen verwendet werden. Beim Hashing ist $\mathcal{H}$ also eine Menge von Funktionen $h : U \to \{1, \dots, m\}$. Beim sogenannten uniformen Hashing gilt die folgende Annahme für jede zwei $x, y \in U$: $\mathbf{Pr}_{h \in \mathcal{F}}[h(x) = h(y)] = \frac{1}{m}$. Diese Annahme erlaubt es, die Auswirkungen von Hash-Kollisionen auf die Zugriffszeit zu beschränken. Hash-Kollisionen von nicht-identischen Elementen sollen vermieden werden, da sie die Zugriffszeit verlängern. Bei lokalitätssensitiven Funktionen hingegen sind Hash-Kollisionen sogar erwünscht, sofern sie vorrangig unter den nächsten Nachbarn auftreten. Oft werden beide miteinander kombiniert, indem man erst eine lokalitätssensitive Funktion anwendet und dann auf den so berechneten Schlüssel, eine Hashfunktion anwendet, um die Schlüsselmenge effizient speichern zu können und darin effizient suchen zu können. Wir ignorieren diesen Aspekt hier und beschränken uns auf die Analyse der lokalitätssensitiven Funktionen.

Definition 18.2. *Sei $X = \mathbb{R}$. Sei $\mathcal{F}$ die Klasse von Funktionen $h_\eta : X \to \mathbb{Z}$ mit $h_\eta(x) = \lceil x + \eta \rceil$, und mit $\eta \in [0, 1)$. Betrachte die Wahrscheinlichkeitsverteilung über $\mathcal{F}$, bei der η gleichverteilt im Intervall $[0, 1)$ gewählt wird.*

Lemma 18.3. *Die Klasse $\mathcal{F}$ aus Definition 18.2 ist lokalitätssensitiv bezüglich des Euklidischen Abstandes. Insbesondere gilt für jedes $x, y \in X$*

$$\mathbf{Pr}_{h_\eta \in \mathcal{F}}[h_\eta(x) = h_\eta(y)] = \max(0, 1 - |x - y|)$$

Beweis. Wenn $x = y$, dann ist die Wahrscheinlichkeit dass x und y auf denselben Funktionswert abgebildet werden gleich 1. Für $|x - y| \geq 1$ werden x und y immer auf unterschiedliche Funktionswerte abgebildet, egal welchen Wert η annimmt. Wir betrachten also den Fall $|x - y| < 1$.

Sei ohne Beschränkung der Allgemeinheit $x \leq y$. Die Werte x und y werden genau dann *nicht* auf denselben Funktionswert abgebildet, wenn in dem Intervall zwischen den Werten $(x + \eta)$ und $(y + \eta)$ eine ganze Zahl liegt. Insbesondere gilt

$$h_\eta(x) \neq h_\eta(y) \quad \Leftrightarrow \quad \lceil x + \eta \rceil \in [x + \eta, y + \eta)$$

Betrachten wir die Zufallsvariable τ definiert durch

$$\tau = \lceil x + \eta \rceil - (x + \eta)$$

Dann gilt nach obiger Betrachtung

$$h_\eta(x) \neq h_\eta(y) \quad \Leftrightarrow \quad \tau \in [0, y - x) \tag{3}$$

Welche Verteilung hat also die Zufallsvariable τ?

Eine wichtige Beobachtung ist, dass $\lceil x + \eta \rceil$ mit $\eta \in [0, 1)$ nur zwei verschiedene Werte annehmen kann, nämlich $\lceil x + \eta \rceil \in \{\lceil x \rceil, \lceil x + 1 \rceil\}$.

Wir betrachten eine weitere Zufallsvariable $\tau' = x + \eta$ in den zwei Fällen.

(Fall 1) $\lceil \tau' \rceil = \lceil x \rceil \quad \Rightarrow \quad \tau' \in [x, \lceil x \rceil]$

(Fall 2) $\lceil \tau' \rceil = \lceil x + 1 \rceil \quad \Rightarrow \quad \tau' \in (\lceil x \rceil, x + 1)$

Daraus ergibt sich für $\tau = \lceil \tau' \rceil - \tau'$

(Fall 1) $\tau = \lceil x \rceil - \tau'$

(Fall 2) $\tau = \lceil x + 1 \rceil - \tau'$.

Es ergeben sich die folgenden Intervalle für Werte von τ in den beiden Fällen.

(Fall 1) $\tau \in [\lceil \lceil x \rceil \rceil - \lceil x \rceil, \lceil x \rceil - x] = [0, \lceil x \rceil - x]$

(Fall 2) $\tau \in (\lceil x + 1 \rceil - (x + 1), \lceil x + 1 \rceil - \lceil x \rceil) = (\lceil x \rceil - x, 1)$

Da η in $[0, 1)$ gleichverteilt ist und daher τ' in $[x, x + 1)$ gleichverteilt ist, schließen wir daraus, dass $\tau \in [0, 1)$ gleichverteilt ist. Nun folgt aus (3), dass wenn $|x - y| < 1$ ist,

$$\mathbf{Pr}_{h_\eta \in \mathcal{F}} [h_\eta(x) \neq h_\eta(y)] = |x - y|$$

Daraus folgt

$$\mathbf{Pr}_{h_\eta \in \mathcal{F}} [h_\eta(x) = h_\eta(y)] = \max(0, 1 - |x - y|)$$

Sei $t \in (0, 1)$ ein Parameter. Es folgt nun, dass die Klasse $\mathcal{F}$ aus Definition 18.2 (r, R, α, β)-lokalitätssensitiv ist mit $\alpha = \beta = 1 - t$ und $r = R = t$. □

Wir wollen die Definition auf höhere Dimensionen erweitern. Dafür wählen wir zufällig eine Gerade durch den Ursprung und projizieren die Punkte auf den eindimensionalen Unterraum und wenden die Funktion aus Definition 18.2 auf den Unterraum an. Das geht am einfachsten indem man einen Einheitsvektor zufällig gleichverteilt auf dem Einheitskreis wählt. Der Einheitskreis $\mathbb{S}^1$ ist die Menge der Einheitsvektoren in $\mathbb{R}^2$. Formal, ist $\mathbb{S}^1 = \{x \in \mathbb{R}^2 \mid \|x\| = 1\}$ definiert. Wir können einen Vektor u zufällig gleichverteilt aus $\mathbb{S}^1$ auswählen indem wir einen Winkel ϕ zufällig gleichverteilt im Interval $[0, 2\pi)$ auswählen und $u = (\cos\phi, \sin\phi)$ definieren.

Definition 18.4. *Sei $X = \mathbb{R}^2$. Sei $\mathcal{F}$ die Klasse von Funktionen $h_{u,\eta} : X \to \mathbb{Z}$ mit $h_{u,\eta}(x) = \lceil \langle x, u \rangle + \eta \rceil$, und mit $\eta \in [0, 1)$ und $u \in \mathbb{S}^1$. Betrachte die Wahrscheinlichkeitsverteilung über $\mathcal{F}$, bei der η gleichverteilt im Intervall $[0, 1)$ gewählt wird und $u = (\cos(\phi), \sin(\phi))$, wobei ϕ gleichverteilt im Intervall $[0, 2\pi)$ gewählt wird.*

Lemma 18.5. *Die Klasse $\mathcal{F}$ aus Definition 18.4 ist (r, R, α, β)-lokalitätssensitiv bezüglich des Euklidischen Abstandes mit $r = \frac{1}{2}$,$R = 2$,$\alpha = \frac{1}{2}$,$\beta = \frac{1}{3}$.*

Beweis. Zunächst stellen wir fest, dass

$$| \langle x, u \rangle - \langle y, u \rangle | = | \langle x - y, u \rangle | = \|x - y\| \cdot \|u\| \cdot | \cos \theta | = \|x - y\| \cdot | \cos \theta |,$$

wobei wir mit θ den Winkel zwischen den Vektoren u und $(x - y)$ bezeichnen.

Wir betrachten beide Fälle aus der Definition der lokalitätssensitiven Funktionen. Sei $0 \leq \|x - y\| < \frac{1}{2}$. In diesem Fall, gilt für die Wahrscheinlichkeit, dass x und y auf verschiedene Funktionswerte abgebildet werden

$$\mathbf{Pr}_{h_{u,\eta} \in \mathcal{F}} [h_{u,\eta}(x) \neq h_{u,\eta}(y)] = | \langle x, u \rangle - \langle y, u \rangle | = \|x - y\| \cdot | \cos \theta | < \frac{1}{2}$$

Also ist

$$\mathbf{Pr}_{h_{u,\eta} \in \mathcal{F}} [h_{u,\eta}(x) = h_{u,\eta}(y)] > \frac{1}{2}$$

Im zweiten Fall betrachten wir $\|x - y\| > 2$. Im Ereignis, dass x und y auf denselben Funktionswert abgebildet werden, muss gelten

$$| \langle x, u \rangle - \langle x, u \rangle | < 1$$

Wir setzen ein und formen um und erhalten

$$1 > \|x - y\| | \cos \theta | > 2 | \cos \theta |$$

Daraus folgern wir, dass $| \cos \theta | < \frac{1}{2}$ gelten muss, in dem Ereignis, dass x und y auf denselben Funktionswert abgebildet werden. Der Vektor $(x - y)$ ist fest und unabhängig von der Wahl der lokalitätssensitiven Funktion $h_{\eta,u}$ mit $u = (\cos \phi, \sin \phi)$. Insbesondere muss der Winkel θ gleichverteilt in $[0, \pi)$ sein, da ϕ gleichverteilt in $[0, 2\pi)$ ist. Also ist

$$\mathbf{Pr}_{h_{u,\eta} \in \mathcal{F}} [h_{u,\eta}(x) = h_{u,\eta}(y)] \leq \mathbf{Pr} \left[| \cos \theta | \leq \frac{1}{2} \right] = \mathbf{Pr} \left[\theta \in \left[\frac{\pi}{3}, \frac{2\pi}{3} \right] \right] = \frac{1}{3}$$

□

Die Analyse aus obigem Beweis funktioniert unter der Annahme, dass $X = \mathbb{R}^2$. Tatsächlich kann man aber zeigen, dass eine ähnliche Klasse von Funktionen auch in höheren Dimensionen lokalitätssensitiv bezüglich des Euklidischen Abstandes ist.

2 Verstärkung durch Komposition

In den obigen Funktionsklassen für den Euklidischen Abstand ist die Erfolgswahrscheinlichkeit für zwei Punkte, die nah beieinander liegen, auf denselben Funktionswert abgebildet zu werden noch nicht hoch genug für praktische Anwendungen. Es ist daher sinnvoll, die Wahrscheinlichkeiten zu verstärken indem man eine Komposition von mehreren Funktionen benutzt.

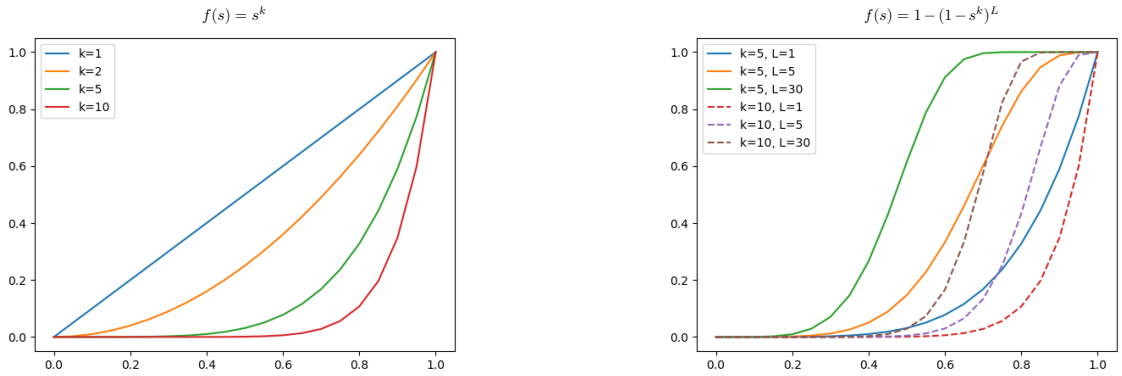

Abbildung 2: Wahrscheinlichkeit für Hashkollisionen für feste $x, y \in X$ bei Komposition, in Abhängigkeit von $s = \mathbf{Pr}_{h \in \mathcal{F}}[h(x) = h(y)]$. Links: k-fache UND-Komposition; Rechts: k-fache UND-Komposition gefolgt von L-facher ODER-Komposition.

Sei k eine natürliche Zahl. Sei $\mathcal{F}$ eine Klasse von (r, R, α, β)-lokalitätssensitiven Funktionen. Eine k-fache UND-Komposition ist eine Funktion $g : X \to U^k$ definiert durch $g(x) = (h_1(x), \ldots, h_k(x))$ mit $h_1, \ldots, h_k \in \mathcal{F}$. Beachte, dass auf U^k eine Ordnungsrelation existiert, sofern auf U eine Ordnungsrelation existiert, zum Beispiel können wir die lexikographische Ordnung annehmen. Dies wird eine UND-Komposition genannt, da $g(x) = g(y)$ voraussetzt, dass $h_i(x) = h_i(x)$ für alle $1 \leq i \leq k$. Wir bezeichnen die resultierende Klasse von Funktionen mit $\mathcal{F}^k$.

Zusätzlich können wir eine ODER-Komposition betrachten. Dies ist eine Komposition der resultierenden Datenstrukturen. Sei L eine natürliche Zahl. Seien $g_1, \ldots, g_L$ zufällig aus $\mathcal{F}^k$ gewählt. Wir berechnen für jede Funktion g_i den Schlüssel $g_i(x)$ für jedes $x \in S$ der Trainingsmenge und fügen $g_i(x)$ in eine separate Datenstruktur D_i ein. Bei einer Anfrage mit einem Element $y \in X$ berechnen wir den Schlüssel $g_i(y)$ und suchen mit diesem Schlüssel in den Datenstrukturen $D_1, \ldots, D_L$. Die Suche ist erfolgreich, wenn wir ein $x \in S$ finden, mit $d(x, y) \leq R$. Angenommen es existiert ein $x \in S$ mit $d(x, y) < r$. Was ist dann die Wahrscheinlichkeit, dass $g_i(x) = g_i(y)$ für mindestens eines der $i \in \{1, \ldots, L\}$?

Lemma 18.6. *Seien $k, L \in \mathbb{N}$. Sei $\mathcal{F}$ eine Klasse von lokalitätssensitiven Funktionen auf einer Grundmenge X. Sei $(g_1, \ldots, g_L)$ eine k-fache UND-Komposition gefolgt von einer L-fachen ODER-Komposition mit $k \cdot L$ Funktionen unabhängig zufällig gewählt aus $\mathcal{F}$. Dann gilt für jedes $x, y \in X$*

$$\mathbf{Pr}\left[\exists i \in \{1, \ldots, L\} : g_i(x) = g_i(y)\right] = 1 - (1 - (\mathbf{Pr}_{h \in \mathcal{F}}[h(x) = h(y)])^k)^L$$

Beweis. Sei $x, y \in X$ fest und sei $s = \mathbf{Pr}_{h \in \mathcal{F}}[h(x) = h(y)]$. Sei $i \in \{1, \ldots, L\}$ fest. Die Wahrscheinlichkeit, dass $g_i(x) = g_i(y)$ ist s^k, da die Funktionswerte von x und y für alle k Funktionen gleich sein müssen. Betrachten wir nun das Ereignis, dass $g_i(x) \neq g_i(y)$ für alle $i \in \{1, \ldots, L\}$. Die Wahrscheinlichkeit dafür ist $(1 - s^k)^L$. Die Wahrscheinlichkeit im Satz ist die Gegenwahrscheinlichkeit dazu. □

Beispiel 18.7. *Betrachten wir die Klasse von Funktionen aus Definition 18.2 für den Euklidischen Abstand in $\mathbb{R}$. Seien $x, y \in \mathbb{R}$ fest und sei $s = \max(0, 1 - |x - y|)$. Laut Lemma 18.3 ist s ist die Wahrscheinlichkeit, dass x und y auf denselben Funktionswert abgebildet werden. Laut Lemma 18.6 ist die Wahrscheinlichkeit, dass $g_i(x) = g_i(y)$ für mindestens eines der*

$i \in \{1, \ldots, L\}$ gleich $1 - (1 - s^k)^L$. Abbildung 2 zeigt Beispiele von Funktionengraphen dieser Funktion für verschiedene Werte von k und L.

Referenzen

- Jeff M. Phillips, Mathematical Foundations of Data Science, Kapitel 4.6, `http://www.cs.utah.edu/~jeffp/M4D/M4D-v0.6.pdf`
- Sariel Har-Peled, Geometric Approximation Algorithms, Springer, Kapitel 15.2, `https://sarielhp.org/book/` (Preprint)