# Theorie des maschinellen Lernens

Hans U. Simon

18. Juli 2017

## 19 Die Nearest-Neighbor-Lernregel

In diesem Abschnitt setzen wir voraus, dass der Grundbereich $\mathcal{X}$ mit einer Metrik $\rho(\cdot,\cdot)$ versehen ist. Zum Beispiel könnte $\mathcal{X} = \mathbb{R}^d$ und $\rho(x,y) = \|x-y\|_2$ gelten.

Es sei $S = [(x_1,y_1),\ldots,(x_m,y_m)] \in (\mathcal{X}\times\mathcal{Y})^m$ eine Trainingsmenge und $u \in \mathcal{X}$ eine Instanz (ohne vorgegebenes Label). Wir bezeichnen dann mit

$$x_1^u,\ldots,x_m^u$$

eine Umordnung der Instanzen $x_1,\ldots,x_m$, so dass

$$\rho(x_1^u,u) \le \ldots \le \rho(x_m^u,u) \ ,$$

wobei (zum Zwecke der Eindeutigkeit) bei gleicher Distanz zu $u$ Vektoren mit kleinerem Index den Vorrang erhalten. $y_1^u,\ldots,y_m^u$ sei eine entsprechende Umordnung der Labels. Somit wäre $x_1^u \in S_{\mathcal{X}}$ der zu $u$ nächstgelegene Nachbar, $x_2^u \in S_{\mathcal{X}}$ der zu $u$ zweitnächstgelegene Nachbar usw.. Die $k$-Nearest-Neighbor-Lernregel macht den folgenden Vorschlag für das Label von $u$.

**k-NN für binäre Klassifikationsprobleme:** Weise $u$ das Label 1 genau dann zu, wenn mindestens $k/2$ der Labels $y_1^u,\ldots,y_k^u$ identisch zu 1 sind (Majoritätsvotum).

**k-NN für Regressionsprobleme:** Weise $u$ das Label $\frac{1}{k}\sum_{i=1}^k y_i^u$ zu.

Es bezeichne im Folgenden $h_S$ die aus $k$-NN resultierende Hypothese. Offensichtlich gilt $h_S(u) = y_1^u$ im Falle $k=1$, d.h., die Lernregel 1-NN weist einer Instanz $u$ einfach das Label ihres „nearest neighbors“ in $S$ zu.

Es sind natürlich auch Varianten von $k$-NN denkbar. Zum Beispiel könnte man im Falle von binären Klassifikationsproblemen ein gewichtetes Majoritätsvotum durchführen, bei dem $y_i^u$ ein zu $\rho(x_i^u,u)$ umgekehrt proportionales Gewicht erhält. Eine entsprechende Bemerkung gilt für Regressionsprobleme.

## 19.1 Eine Fehlerschranke für 1-NN

Wir betrachten binäre Klassifikationsprobleme auf dem Grundbereich $\mathcal{X} = [0,1]^d$. Somit ist $\mathcal{Y} = \{0,1\}$ und $\ell$ ist die Null-Eins-Verlustfunktion. Es bezeichne $\|\cdot\| = \|\cdot\|_2$ die Euklidische Norm in $\mathbb{R}^d$. Wie üblich sei $\mathcal{D}$ eine Verteilung auf $\mathcal{X} \times \mathcal{Y}$. Weiter sei $\mathcal{D}_\mathcal{X}$ die von $\mathcal{D}$ induzierte Randverteilung auf $\mathcal{X}$ und $\eta(x) = \Pr[y = 1|x]$ bezeichne die bedingte Wahrscheinlichkeit für das 1-Label zu einer gegebenen Instanz $x$. Die Hypothese mit der kleinsten Fehlerrate, genannt die *Bayes-optimale Hypothese*, ist dann gegeben durch

$$h^*(x) = \mathbb{1}_{[\eta(x) \geq 1/2]} \ .$$

Da $\mathcal{D}$ dem Lerner unbekannt ist, sind natürlich $\mathcal{D}_\mathcal{X}$, $\eta$ und $h^*$ dem Lerner ebenfalls unbekannt. Das Ziel ist dennoch, eine Hypothese zu lernen, deren Fehlerrate nicht dramatisch viel größer ist als die Fehlerrate von $h^*$. Wenn die räumliche Nähe zweier Instanzen $x, x' \in \mathbb{R}^d$ nichts über die Gleichartigkeit ihrer Labels aussagt, hat man mit der Lernregel 1-NN keine Chance. Wir setzen deshalb voraus, dass $\eta$ für eine Konstante $c > 0$ $c$-Lipschitz ist, d.h.,

$$\forall x, x' \in \mathcal{X} : \ |\eta(x) - \eta(x')| \leq c \cdot \|x - x'\|$$

und beweisen das

**Lemma 19.1** *Mit obigen Notationen und Voraussetzungen gilt für die aus*

$$S = [(x_1, y_1), \ldots, (x_m, y_m)] \sim \mathcal{D}^m$$

*und Anwendung von 1-NN resultierende Hypothese $h_S$:*

$$\mathbb{E}_{S \sim \mathcal{D}^m}[L_\mathcal{D}(h_S)] \leq 2 \cdot L_\mathcal{D}(h^*) + c \cdot \mathbb{E}_{S_\mathcal{X} \sim \mathcal{D}_\mathcal{X}^m, u \sim \mathcal{D}_\mathcal{X}}[\|u - x_1^u\|] \ . \tag{1}$$

**Beweis** Das Zufallsexperiment

$$S \sim \mathcal{D}^m \ \text{ und } \ (u, y) \sim \mathcal{D}$$

ist äquivalent zu

$$S_\mathcal{X} \sim \mathcal{D}_\mathcal{X}^m, u \sim \mathcal{D}_\mathcal{X}, y_i \sim \eta(x_i) \text{ für } i = 1, \ldots, m \text{ und } y \sim \eta(u) \ ,$$

d.h., wir können zuerst zufällige ungelabelte Instanzen ermitteln und danach die zugehörigen Labels. Aus $S_\mathcal{X}$ und $u$ lässt sich $x_1^u$ ermitteln und freilich ist $y_1^u$ gemäß $y_1^u \sim \eta(x_1^u)$ verteilt. Daher gilt:

$$\begin{aligned}
\mathbb{E}_{S \sim \mathcal{D}^m}[L_\mathcal{D}(h_S)] &= \mathbb{E}_{S \sim \mathcal{D}^m, (u,y) \sim \mathcal{D}}[\mathbb{1}_{[y_1^u \neq y]}] \\
&= \mathbb{E}_{S_\mathcal{X} \sim \mathcal{D}_\mathcal{X}^m, u \sim \mathcal{D}_\mathcal{X}} \mathbb{E}_{y_1^u \sim \eta(x_1^u), y \sim \eta(u)}[\mathbb{1}_{[y_1^u \neq y]}] \\
&= \mathbb{E}_{S_\mathcal{X} \sim \mathcal{D}_\mathcal{X}^m, u \sim \mathcal{D}_\mathcal{X}} \left[ \Pr_{y_1^u \sim \eta(x_1^u), y \sim \eta(u)}[y_1^u \neq y] \right]
\end{aligned}$$

Der Term $\Pr_{y_1^u \sim \eta(x_1^u), y \sim \eta(u)}[y_1^u \neq y]$ lässt sich nach oben abschätzen wie folgt:

$$\begin{aligned}
\Pr_{y_1^u \sim \eta(x_1^u), y \sim \eta(u)}[y_1^u \neq y] &= \eta(x_1^u)(1-\eta(u)) + (1-\eta(x_1^u))\eta(u) \\
&= \eta(x_1^u) + \eta(u) - 2\eta(x_1^u)\eta(u) \\
&= 2\eta(u)(1-\eta(u)) + (2\eta(u)-1)(\eta(u)-\eta(x_1^u)) \\
&\leq 2 \cdot \Pr_{y \sim \eta(u)}[h^*(u) \neq y] + c \cdot \|u - x_1^u\|
\end{aligned}$$

Die letzte Ungleichung gilt, da $\eta$ Werte in $[0,1]$ annimmt und $c$-Lipschitz ist, und weil daher folgendes gilt:

- $\eta(u)(1-\eta(u)) \leq \min\{\eta(u), 1-\eta(u)\} = \Pr_{y \sim \eta(u)}[h^*(u) \neq y]$.
- $|2\eta(u)-1| \leq 1$ und $|\eta(u) - \eta(x_1^u)| \leq c \cdot \|u - x_1^u\|$.

Wenn wir beide Seiten der Ungleichung

$$\Pr_{y_1^u \sim \eta(x_1^u), y \sim \eta(u)}[y_1^u \neq y] \leq 2 \cdot \Pr_{y \sim \eta(u)}[h^*(u) \neq y] + c \cdot \|u - x_1^u\|$$

über $S_{\mathcal{X}} \sim \mathcal{D}_{\mathcal{X}}^m$ und $u \sim \mathcal{D}_{\mathcal{X}}$ mitteln, erhalten wir (1). **qed.**

Wir wollen den Term $\mathbb{E}_{S_{\mathcal{X}} \sim \mathcal{D}_{\mathcal{X}}^m, u \sim \mathcal{D}_{\mathcal{X}}}[\|u - x_1^u\|]$ nach oben abschätzen, benötigen dazu aber erst noch folgendes Hilfsresultat:

**Lemma 19.2** *Es seien $C_1, \ldots, C_r$ (messbare) Teilmengen von $\mathcal{X}$ mit den Wahrscheinlichkeitsmassen $w_i := \mathcal{D}_{\mathcal{X}}(C_i)$ für $i = 1, \ldots, r$. Zu $S_{\mathcal{X}} \sim \mathcal{D}_{\mathcal{X}}^m$ assoziieren wir die Zufallsvariable*

$$W(S_{\mathcal{X}}) := \sum_{i: C_i \cap S_{\mathcal{X}} = \emptyset} w_i = \sum_{i=1}^{r} w_i \cdot \mathbb{1}_{[C_i \cap S_{\mathcal{X}} = \emptyset]} \ ,$$

*die die Wahrscheinlichkeitsmassen der nicht von $S$ getroffenen Teilmengen aufsummiert. Dann gilt*

$$\mathbb{E}_{S_{\mathcal{X}} \sim \mathcal{D}_{\mathcal{X}}^m}[W(S_{\mathcal{X}})] \leq \frac{r}{em} \ .$$

**Beweis** Der Beweis ergibt sich aus folgender Rechnung:

$$\begin{aligned}
\mathbb{E}_{S_{\mathcal{X}} \sim \mathcal{D}_{\mathcal{X}}^m}[W(S_{\mathcal{X}})] &= \mathbb{E}\left[\sum_{i=1}^{r} w_i \cdot \mathbb{1}_{[C_i \cap S_{\mathcal{X}} = \emptyset]}\right] \\
&= \sum_{i=1}^{r} w_i \cdot \mathbb{E}[\mathbb{1}_{[C_i \cap S_{\mathcal{X}} = \emptyset]}] \\
&= \sum_{i=1}^{r} w_i (1-w_i)^m \\
&\leq \sum_{i=1}^{r} w_i \cdot e^{-m w_i} \\
&\leq r \cdot \max_{a \in [0,1]} a \cdot e^{-ma} \leq \frac{r}{em}
\end{aligned}$$

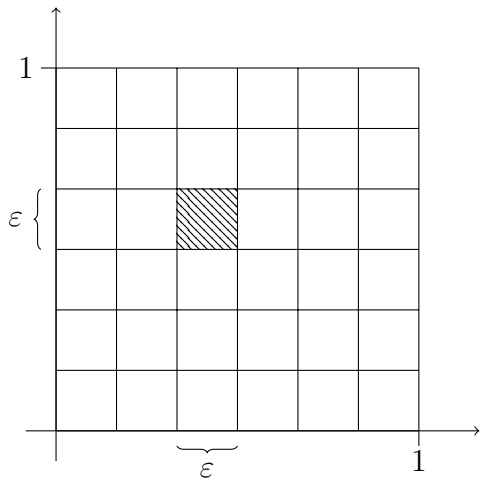


Abbildung 1: Zerlegung des Einheitsquadrats in Unterwürfel mit Kantenlänge $\varepsilon$.

Die letzte Ungleichung gilt, da $\frac{1}{em}$ ein Maximierer von $a \cdot e^{-ma}$ ist. **qed.**

Wir kommen jetzt auf unser eigentliches Ziel zurück: Herleitung einer oberen Schranke für den Term $\mathbb{E}_{S_\mathcal{X} \sim \mathcal{D}_\mathcal{X}^m, u \sim \mathcal{D}_\mathcal{X}}[\|u - x_1^u\|]$.

**Lemma 19.3** *Es sei $\mathcal{X} = [0,1]^d$. Zu gegebenem $S_\mathcal{X} = (x_1, \ldots, x_m) \in \mathcal{X}^m$ und $u \in \mathcal{X}$ bezeichne $x_1^u$ einen Punkt aus $S_\mathcal{X}$ mit der kleinsten Distanz zu $u$. Mit diesen Notationen gilt:*

$$\mathbb{E}_{S_\mathcal{X} \sim \mathcal{D}_\mathcal{X}^m, u \sim \mathcal{D}_\mathcal{X}}[\|u - x_1^u\|] \leq 2\sqrt{d}(em)^{-\frac{1}{d+1}} \quad . \tag{2}$$

**Beweis** Es sei $0 < \varepsilon < 1$ ein Parameter, den wir an einer späteren Stelle des Beweises geschickt wählen. Wir zerlegen den Einheitswürfel $\mathcal{X} = [0,1]^d$ auf die offensichtliche Weise in $r = (1/\varepsilon)^d$ Unterwürfel der Kantenlänge $\varepsilon$. Für $d = 2$ ist dies in Abbildung 1 illustriert. Für $S \sim \mathcal{D}_\mathcal{X}^m$ und $u \sim \mathcal{D}_\mathcal{X}$ unterscheiden wir zwei Fälle:

**Fall 1:** $u$ liegt in einem von $S_\mathcal{X}$ getroffenen Unterwürfel.

**Fall 2:** $u$ liegt in einem von $S_\mathcal{X}$ nicht getroffenen Unterwürfel.

Es sei $W(S)$ die Zufallsvariable aus Lemma 19.2. Offensichtlich tritt Fall 2 (bzw. Fall 1 mit Wahrscheinlichkeit $W(S_\mathcal{X})$ (bzw. $1 - W(S_\mathcal{X})$) auf. Betrachte die Zufallsvariable $X := \|u - x_1^u\|$. Im Fall 1 gilt $X \leq \varepsilon\sqrt{d}$, da $\varepsilon\sqrt{d}$ der Durchmesser eines $d$-dimensionalen Würfels mit Kantenlänge $\varepsilon$ ist. Im Fall 2 gilt die triviale Schranke $X \leq \sqrt{d}$, da $\sqrt{d}$ der Durchmesser von $[0,1]^d$ ist. Weiter geht es mit folgender Rechnung:

$$\begin{aligned}
\mathbb{E}_{S_\mathcal{X} \sim \mathcal{D}_\mathcal{X}^m, u \sim \mathcal{D}_\mathcal{X}}[X] &= \mathbb{E}_{S_\mathcal{X} \sim \mathcal{D}_\mathcal{X}^m}\left[\mathbb{E}_{u \sim \mathcal{D}_\mathcal{X}}[X|\text{Fall 1}] \cdot (1 - W(S)) + \mathbb{E}_{u \sim \mathcal{D}_\mathcal{X}}[X|\text{Fall 2}] \cdot W(S)\right] \\
&\leq \varepsilon\sqrt{d} + \sqrt{d} \cdot \mathbb{E}_{S_\mathcal{X} \sim \mathcal{D}_\mathcal{X}^m} W(S_\mathcal{X}) \\
&\leq \varepsilon\sqrt{d} + \sqrt{d} \cdot \frac{\varepsilon^{-d}}{em} = \sqrt{d} \cdot \left(\varepsilon + \frac{\varepsilon^{-d}}{em}\right)
\end{aligned}$$

Bei der letzten Ungleichung wurde Lemma 19.2 verwendet. Wir können die Bedingung $\varepsilon = \frac{\varepsilon^{-d}}{em}$ erzwingen, indem wir $\varepsilon = (em)^{-\frac{1}{d+1}}$ setzen. Auf diese Weise erhalten wir schließlich (2). **qed.**

Wir merken kurz an, dass der Beweis eine kleine Schwierigkeit unter den Teppich kehrt: mit unserer Wahl von $\varepsilon$ ist $1/\varepsilon$ keine ganze Zahl, d.h., die Unterteilung einer Kante der Länge 1 in Teilkanten der Länge $\varepsilon$ geht nicht exakt auf. Einen exakten Beweis erhält man, wenn man $\varepsilon = 1/T$ für eine geeignete Wahl einer ganzen Zahl $T \geq 1$ setzt. Dann kann man allerdings die Bedingung $\varepsilon = \frac{\varepsilon^{-d}}{em}$ nur noch approximativ erreichen. Ein „sauberer“ Beweis würde das marginal schlechtere Resultat

$$\mathbb{E}_{S_{\mathcal{X}} \sim \mathcal{D}_{\mathcal{X}}^m, u \sim \mathcal{D}_{\mathcal{X}}}[\|u - x_1^u\|] \leq 2\sqrt{d} \frac{1}{\left\lfloor (em)^{\frac{1}{d+1}} \right\rfloor}$$

liefern. Eine entsprechende Bemerkung gilt auch für das folgende Resultat, welches sich direkt aus den Lemmas 19.1 und 19.3 ergibt:

**Theorem 19.4** *Mit den Notationen und Voraussetzungen wie in Lemma 19.1 gilt:*

$$\mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S)] \leq 2 \cdot L_{\mathcal{D}}(h^*) + 2c\sqrt{d}(em)^{-\frac{1}{d+1}} \quad . \tag{3}$$

## 19.2 Fluch der Hochdimensionalität (Curse of Dimensionality)

Um zu erreichen, dass der Term $2c\sqrt{d}(em)^{-\frac{1}{d+1}}$ in (3) einen Beitrag von maximal $\varepsilon$ leistet, müssten wir

$$m \geq \frac{1}{e} \cdot \left( \frac{2c\sqrt{d}}{\varepsilon} \right)^{d+1}$$

setzen. Am meisten schmerzt es, dass die erforderliche Größe $m$ der Trainingsmenge exponentiell mit $d$ wächst, ein Umstand der auch als „Curse of Dimensionality“ bezeichnet wird. Man kann sich natürlich fragen, ob eine verbesserte Analyse oder die Verwendung einer anderen Lernregel nicht vielleicht eine freundlichere Abhängigkeit von $d$ zu Tage fördern könnte. Dies ist i.A. leider nicht der Fall:

**Theorem 19.5** *Zu jeder Wahl von $c, d \geq 1$ und jeder Lernregel $S \mapsto h_S$ existiert eine Verteilung $\mathcal{D}$ auf $[0,1]^d \times \{0,1\}$, so dass folgendes gilt: die von $\mathcal{D}$ induzierte Funktion $\eta : \mathcal{X} \to [0,1]$ ist $c$-Lipschitz und $L_{\mathcal{D}}(h^*) = 0$, aber für alle $m \leq (c+1)^d/2$ gilt $\mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S)] \geq 1/4$.*

**Beweis** Es sei $G_c^d$ das Gitter in $[0,1]^d$ der Maschenweite $1/c$, d.h.,

$$G_c^d = \left\{ \left( \frac{a_1}{c}, \ldots, \frac{a_d}{c} \right) \middle| \, a_1, \ldots, a_d \in \{0, 1, \ldots, c\} \right\} \quad .$$

Das Gitter enthält $|G_c^d| = (c+1)^d$ Gitterpunkte. Da zwei verschiedene Punkte in $G_c^d$ mindestens Abstand $1/c$ voneinander haben, ist jede Funktion von $G_c^d$ nach $[0,1]$ automatisch $c$-Lipschitz. Zu jeder Funktion $f : G_c^d \to \{0,1\}$ assoziieren wir die Verteilung $\mathcal{D}^f$ mit:

1. $\mathcal{D}^f_{\mathcal{X}}$ ist uniform auf $G^d_c$.

2. Die von $\mathcal{D}^f$ induzierte Funktion $\eta$ ist identisch zu $f$, d.h., jedem $x \in G^d_c$ wird (mit Wahrscheinlichkeit 1) das Label $f(x)$ zugewiesen. In diesem Fall ist auch die Bayes-optimale Hypothese $h^*$ identisch zu $f$ und es gilt $L_{\mathcal{D}}(h^*) = 0$.

Wenn wir $f$ uniform zufällig aus der Menge aller Funktionen von $G^d_c$ nach $\{0,1\}$ wählen und $m \leq (c+1)^d/2$, dann sind die Labels auf den mindestens $(c+1)^d/2$ nicht in $S_{\mathcal{X}}$ enthaltenen Punkten aus $G^d_c$ aus der Perspektive des Lerners perfekte Zufallsbits. Daher beträgt die über $\{\mathcal{D}^f |\ f : G^d_c \to \{0,1\}\}$ und $S$ gemittelte Fehlerrate mindestens $1/4$. Nach dem Schubfachprinzip muss ein $f^* \in \{f |\ f : G^d_c \to \{0,1\}\}$ existieren mit:

$$\mathcal{D} = \mathcal{D}^{f^*} \Rightarrow \mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S)] \geq \frac{1}{4} \ ,$$

womit der Beweis des Theorems erbracht wäre. **qed.**

**Effizienzbetrachtungen:** Eine naive Methode zur Ermittlung des „Nearest Neighbor“ würde $m$ Datenpunkte im $\mathbb{R}^d$ durchmustern und daher Laufzeit $O(md)$ benötigen. Durch Einsatz raffinierter Datenstrukturen zur Abspeicherung der $m$ Datenpunkte kann man die Laufzeit auf $o(\text{poly}(d)\log(m))$ verkleinern. Allerdings belegen solche Datenstrukturen Speicherplatz in der Größenordnung $m^{O(d)}$. Es gibt daher Versuche, das „Nearest Neighbor Problem“ approximativ, dafür aber effizienter, zu lösen. Im Rahmen dieser Vorlesung können wir auf Details dazu nicht näher eingehen.