Universität Salzburg | Florian Graf

Machine Learning

Übungsblatt **14** | 14 Punkte

Aufgabe 1. *Nächste-Nachbarn-Klassifikation* 14 P.

In dieser Aufgabe untersuchen wir das asymptotische Verhalten des 1-NN Klassifikators.
Wir betrachten zunächst folgende Situation. Es sei $\mathbf{x} \in \mathbb{R}^d$ und $\mathbf{x}_{\text{NN}} \in \mathbb{R}^d$ sein nächster Nachbar in der Menge $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, also $\mathbf{x}_{\text{NN}} := \arg\min_i \operatorname{dist}(\mathbf{x}, \mathbf{x}_i)$. Wir nehmen an, dass die $\mathbf{x}_i$ unabhängig von der gleichen Verteilung gezogen wurden. Des Weiteren nennen wir $B_r(\mathbf{x}) = \{\mathbf{z} \in \mathbb{R}^d \colon \operatorname{dist}(\mathbf{x}, \mathbf{z}) < r\}$ für $r > 0$.

(a) Zeigen Sie, dass $\mathbb{P}[\operatorname{dist}(\mathbf{x}, \mathbf{x}_{\text{NN}}) < r] = 1 - (1 - \mathbb{P}[\mathbf{x}_1 \in B_r(\mathbf{x})])^n$. Hierbei ist $\mathbf{x}$ fest (deterministisch).

(b) Folgern Sie, dass dann für große Stichprobenumfänge n, $\mathbf{x}_{\text{NN}}$ mit hoher Wahrscheinlichkeit nahe bei $\mathbf{x}$ liegt, sofern für alle $r > 0$ gilt, dass $\mathbb{P}[\mathbf{x}_1 \in B_r(\mathbf{x})] > 0$. Erklären Sie außerdem, warum diese Annahme notwendig ist und was sie bedeutet.

(c) Erläutern Sie dieses Resultat im Kontext der Nächste-Nachbarn Klassifikation.

Wir betrachten nun den Euklidischen Abstand, also $\operatorname{dist}_{\text{E}}(\mathbf{y}, \mathbf{z}) := \sqrt{\sum_{k=1}^d (y_k - z_k)^2}$ und nehmen an, dass unsere Wahrscheinlichkeitsverteilung derart ist, dass die einzelnen Koordinaten von $\mathbf{y}$ unabhängig voneinander sind.

(d) Zeigen Sie, dass $\mathbb{P}[\mathbf{y} \in B_r(\mathbf{z})] \leq \left(\max_{k \in \{1,\ldots,d\}} \mathbb{P}[|y_k - z_k| < r]\right)^d$. Hierbei ist $\mathbf{z}$ fest (deterministisch).

(e) Diskutieren Sie dieses Ergebnis im Kontext der Nächste-Nachbarn Klassifikation.

(f) Berechnen Sie die rechte Seite der Ungleichung aus (d) im Falle, dass $\mathbf{y}$ von einer Gleichverteilung auf $[0,1]^d$ gezogen wird und r klein genug ist, sodass $B_r(\mathbf{z}) \subset [0,1]^d$.

(g) Oftmals benutzt man auch den Mahalonobis-Abstand für die Nächste-Nachbarn-Klassifikation, er ist definiert als $\operatorname{dist}_{\text{M}}(\mathbf{y}, \mathbf{z}) := \sqrt{(\mathbf{y} - \mathbf{z})^\top \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{z})}$, wobei $\mathbf{\Sigma}$ (eine Schätzung der) Kovarianzmatrix der zugrundeliegenden Verteilung ist. Verbessert sich dadurch das Verhalten der Nächste-Nachbarn-Klassifikation bezüglich der Dimension d der Merkmale? Es genügt hierbei, den Fall, $\mathbf{z} = \mathbb{E}[\mathbf{y}]$ zu betrachten. Sie dürfen annehmen, dass Σ invertierbar ist.