Computational Learning Theory

Sommersemester 2017

Mario Holldack, M. Sc.
Prof. Dr. Georg Schnitger
Hannes Seiwert, M. Sc.

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Institut für Informatik
AG Theoretische Informatik

Übungsblatt 2

Ausgabe: 01.05.2017
Abgabe: 08.05.2017

Aufgabe 2.1 *Agnostisches Lernen: Lineare Regression* (12 Punkte)

Sei $X = [0, 1]$ und $Y = \mathbb{R}$. Wir betrachten die Hypothesenklasse $\mathcal{H}$ aller affinen Funktionen $f : X \to Y$ mit $f(x) = a \cdot x + b$, wobei $a, b \in \mathbb{R}$.

Sei $f \in \mathcal{H}$ mit $f(x) = a \cdot x + b$ im Folgenden fest. Wir wollen f anhand von Beispielen lernen. Die Beispiele sind aber durch einen additiven Fehler ξ verrauscht, wobei ξ normalverteilt um den Mittelwert $\mu = 0$ mit Varianz $\sigma^2 > 0$ sei. Der Lernalgorithmus erhält nur die verrauschten Beispiele (x, y) mit $y = f(x) + \xi$.

Die Verteilung D über $X \times Y$ sieht folgendermaßen aus: $D(x)$ ist die Gleichverteilung auf X und für ein gegebenes $x \in X$ ist $D(y|x)$ normalverteilt um den Mittelwert $\mu = f(x)$ mit Varianz σ^2. Als Loss-Funktion wählen wir den quadratischen Loss.

a) Sei die Beispielmenge $S = \{(x_1, y_1), \ldots, (x_s, y_s)\}$ gegeben und ferner seien $x = (x_1, \ldots, x_s)$, $y = (y_1, \ldots, y_s)$ und $\bar{x} = \frac{1}{s} \sum_{i=1}^{s} x_i$ sowie $\bar{y} = \frac{1}{s} \sum_{i=1}^{s} y_i$.

   Zeigen Sie, dass die Hypothese h_S mit

   $$h_S(x) = \hat{a} \cdot x + \hat{b} \quad \text{mit} \quad \hat{a} = \frac{\mathrm{Cov}[x, y]}{\mathrm{Var}[x]} \quad \text{und} \quad \hat{b} = \bar{y} - \hat{a} \cdot \bar{x}$$

   eine ERM-Hypothese ist. Dabei ist $\mathrm{Cov}[x, y] = \frac{1}{s} \sum_i (x_i - \bar{x})(y_i - \bar{y})$ die (empirische) Kovarianz und $\mathrm{Var}[x] = \frac{1}{s} \sum_i (x_i - \bar{x})^2$ die (empirische) Varianz.

   Hinweis: $\mathrm{Loss}^S(h_S)$ ist eine reellwertige Funktion, die von zwei reellen Parametern abhängt, also $\mathrm{Loss}^S(h_S) : \mathbb{R}^2 \to \mathbb{R}$. Wie minimiert man eine solche Funktion?

b) Berechnen Sie den Approximationsfehler von $\mathcal{H}$. Zeigen Sie dann, dass für eine gegebene Hypothese $h \in \mathcal{H}$ mit $h(x) = a' \cdot x + b'$ der Schätzfehler

   $$\tfrac{1}{3}(a'-a)^2 + (a'-a)(b'-b) + (b'-b)^2$$

   beträgt.

Bitte wenden!

Aufgabe 2.2 *Agnostisches PAC-Lernen: Wie viele Beispiele sind notwendig?* (10 Punkte)

In der Vorlesung wurde (für endliche Konzeptklassen und beschränkte Loss-Funktionen) eine obere Schranke für die hinreichende Beispielzahl s für agnostisches PAC-Lernen gezeigt. Wir wollen nun eine *untere* Schranke für die *notwendige* Anzahl an Beispielen herleiten. Betrachte dazu ein Minimalbeispiel[1]:

Gegeben seien die Mengen $X = \{x\}, Y = \{0, 1\}$ und die Hypothesenklasse $\mathcal{H} = \{(x, 0), (x, 1)\} \subseteq X \times Y$. Seien $0 < \varepsilon \leq \sqrt{1/2}$ und $0 < \delta \leq \frac{1}{4}$ fest. Die Verteilung sei D mit $\text{prob}_D[(x, 0)] = \frac{1+\varepsilon}{2}$ und $\text{prob}_D[(x, 1)] = \frac{1-\varepsilon}{2}$. Als Loss-Funktion wählen wir den 0-1-Loss.

Zu lernen ist also das „gute" Konzept $c = (x, 0)$ mit dem geringstem erwartetem Loss.

a) Sei $h = (x, 1)$ das „schlechte" Konzept. Berechnen Sie den Approximationsfehler von $\mathcal{H}$ sowie den Schätzfehler von h.

b) Ein Lernalgorithmus erhalte eine gemäß D zufällig gezogene Beispielmenge S der Größe $s < \frac{\ln(1/4\delta)}{2\varepsilon^2}$ und berechne eine ERM-Hypothese h_S.

   Zeigen Sie: $\text{prob}[\, h_S = (x, 1)\,] > \delta$.

   Hinweis: Sie dürfen folgende Ungleichungen ohne Beweis benutzen:

   - *Sluds Ungleichung:* Sei Z binomialverteilt[2] zu den Parametern n und $p = \frac{1-\varepsilon}{2}$. Dann gilt $\text{prob}[Z \geq n/2] \geq \frac{1}{2}\left(1 - \sqrt{1 - e^{-n\varepsilon^2/(1-\varepsilon^2)}}\right)$
   - $\frac{1}{2}\big(1 - \sqrt{1 - x}\big) \geq x/4$ für $x \in (0, 1]$.

Fazit: Die Beispielzahl s reicht nicht aus, um mit hinreichender Wahrscheinlichkeit das richtige Konzept zu lernen.

Aufgabe 2.3 *VC-Dimension von Kugeln* (10 Punkte)

Für $d \in \mathbb{N}, m \in \mathbb{R}^d$ und $r > 0$ sei $K_r(m) := \{x \in \mathbb{R}^d : \|x - m\|^2 \leq r^2\}$ die d-dimensionale Kugel mit Radius r und Mittelpunkt m. Die Konzeptklasse $\textsc{Kugel}_d = \left\{K_r(m) : m \in \mathbb{R}^d, r > 0\right\}$ besteht aus allen d-dimensionalen Kugeln.

a) Zeigen Sie: $\text{VC}(\textsc{Kugel}_2) = 3$.

b) Zeigen Sie: Für beliebiges $d \in \mathbb{N}$ gilt $\text{VC}(\textsc{Kugel}_d) \leq d + 1$.

   Hinweis: Folgendes Ergebnis aus der Vorlesung könnte hilfreich sein: $\text{VC}(\textsc{Halbraum}_d) \leq d + 1$

[1] Hier eine alltagstaugliche Interpretation: In einem Staat mit einem Zwei-Parteien-System (mit den Parteien 0 und 1) erhalte ein Demoskopie-Institut eine repräsentative Stichprobe der Größe s von Wahlzetteln, um daraus eine Hochrechnung zu erstellen, die vorhersagt, welche Partei die Mehrheit errungen hat. Liegen die (wahren) Ergebnisse der Parteien nah beieinander und ist die Stichprobe zu klein, so ist die Hochrechnung nicht zuverlässig.

[2] D.h. $\text{prob}[Z = k] = \binom{n}{k} p^k (1-p)^{n-k}$, vgl. https://de.wikipedia.org/wiki/Binomialverteilung