Einführung in Statistisches Lernen

Vapnik, An Overview of Statistical Learning Theory (1999)

Nils Kober

12. Mai 2020

Einführung in Statistisches Lernen

1. Situation des Lernproblems
   - Lernen aus Stichproben
   - Risikominimierung
   - Typische Lernprobleme
     - Mustererkennung
     - Regression
     - Dichteschätzung
   - Allgemeines Lernproblem
   - Minimierung des empirischen Risikos
2. Konsistenz des ERM Prinzips
   - Definition
   - Key Theorem of Learning Theory
   - Bedingungen für gleichmäßige Konvergenz

Lernen aus Stichproben

- Betrachte Räume $\mathcal{X}$ (Eingabe) und $\mathcal{Y}$ (Ausgabe, label set).
- *Generator* von zufälligen Elementen $x \in \mathcal{X}$ gemäß unbekannter Verteilung $\mathbb{P}^X$
- Überwacher (Supervisor); bestimmt für jedes $x \in \mathcal{X}$ eine Ausgabe $y \in \mathcal{Y}$ gemäß einer bedingten Verteilung $\mathbb{P}^{Y|X=x}$
- Lernmaschine; implementiert Menge von Funktionen

$$f(\,\cdot\,, \alpha) : \mathcal{X} \to \mathcal{Y},\ \alpha \in \Lambda$$

- Ziel:
  Finde bestmögliche Funktion $f(\,\cdot\,, \hat{\alpha})$, die das Verhalten des Überwachers vorhersagt, basierend auf einer i.i.d. Stichprobe $(x_1, y_1), \ldots (x_n, y_n)$ gezogen gemäß $\mathbb{P}(x, y) = \mathbb{P}^X(x)\mathbb{P}^{Y|X=x}(y)$.

Risikominimierung

- Verlustfunktion (loss function)

$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

- Risikofunktional

$$\begin{aligned} R : \Lambda &\to \mathbb{R} \\ \alpha &\mapsto \int L(y, f(x, \alpha)) \, \mathrm{d}\mathbb{P}(x, y) \end{aligned}$$

- Ziel: Minimiere $R(\alpha)$ über Λ

Mustererkennung

- $\mathcal{Y} = \{0, 1\}$
- $f(\cdot, \alpha),\ \alpha \in \Lambda$ Indikatorfunktionen
- $L(y, f(x, \alpha)) = \begin{cases} 0 & \text{falls } y = f(x, \alpha) \\ 1 & \text{falls } y \neq f(x, \alpha) \end{cases}$
- Dann gilt:

$$\begin{aligned} R(\alpha) &= \int L(y, f(x, \alpha)) \mathrm{d}\mathbb{P}(x, y) \\ &= \int \mathbb{1}\{y \neq f(x, \alpha)\} \mathrm{d}\mathbb{P}(x, y) \\ &= \mathbb{P}(y \neq f(x, \alpha)) \end{aligned}$$

- $R(\alpha)$ gibt Wahrscheinlichkeit eines Klassifikationsfehlers an.

Regression - Schätzen des bedingten Erwartungswerts

- $\mathcal{Y} = \mathbb{R}^d$
- $f(\cdot, \alpha_0) = \int y \, \mathrm{d}\mathbb{P}(y|x)$
- $L(y, f(x, \alpha)) = (y - f(x, \alpha))^2$
- Dann gilt:

$$R(\alpha_0) \leq R(\alpha) \quad \forall \alpha \in \Lambda$$

Beweis: Regression

- Für $a \in \mathbb{R}$ gilt:

$$\mathbb{E}[X - a]^2 = \mathbb{E}[X - \mathbb{E}X]^2 + (\mathbb{E}X - a)^2 \tag{1}$$

Dichteschätzung

- $\mathcal{Y} = [0, 1]$
- $p(\cdot, \hat{\alpha}),\ \alpha \in \Lambda$ Dichtefunktionen
- Sei $p(x, \alpha_0)$ eine Dichte von $\mathbb{P}^X$
- $L(p(x, \alpha)) = -log(p(x, \alpha))$
- Dann gilt:

$$R(\alpha_0) \leq R(\alpha) \quad \forall \alpha \in \Lambda$$

Allgemeines Lernproblem

- Wahrscheinlichkeitsmaß $\mathbb{P}^Z$ auf Raum $\mathcal{Z}$
- Verlustfunktionen $Q(\,\cdot\,, \alpha),\ \alpha \in \Lambda$
- Risikofunktional

$$R(\alpha) = \int Q(z, \alpha)\, \mathrm{d}\mathbb{P}(z)$$

- Minimiere $R(\alpha)$ basierend auf i.i.d. Stichprobe $z_1, ..., z_n$

Minimierung des empirischen Risikos

- Empirisches Risiko

$$R_{emp}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} Q(z_i, \alpha)$$

- Minimum des (realen) Risikos:

$$R(\alpha_0) = \min_{\alpha \in \Lambda} R(\alpha)$$

- Minimum des empirischen Risikos:

$$R_{emp}(\alpha_n) = \min_{\alpha \in \Lambda} R_{emp}(\alpha)$$

- Approximation der Funktion $Q(\,\cdot\,, \alpha_0)$ durch $Q(\,\cdot\,, \alpha_n)$
- Empirical risk minimazation induction principle (ERM principle)

Konsistenz des ERM Prinzips

- Das ERM heißt konsistent, falls

$$R(\alpha_n) \xrightarrow[n\to\infty]{\mathbb{P}} R(\alpha_0)$$
$$\text{und} \quad R_{emp}(\alpha_n) \xrightarrow[n\to\infty]{\mathbb{P}} R(\alpha_0)$$

- Sei $\Lambda(c) = \{\alpha \in \Lambda | R(\alpha) \geq c\}$
- Das ERM heißt nichttrivial/strikt konsistent, falls für alle $c \in \mathbb{R}$ mit $\Lambda(c) \neq \phi$ gilt:

$$\inf_{\alpha\in\Lambda(c)} R_{emp}(\alpha) \xrightarrow[n\to\infty]{\mathbb{P}} \inf_{\alpha\in\Lambda(c)} R(\alpha)$$

- Aus strikter Konsistenz folgt insbesondere die (normale) Konsistenz des ERM Prinzips (falls $R(\alpha)$ nach unten beschränkt ist).

Beweis: Strikte Konsistenz $\Rightarrow$ Konsistenz

- Sei

$$T := \inf_{\alpha \in \Lambda} R(\alpha)$$

Key Theorem of Learning Theory

- Sei $R(\alpha)$ beschränkt, d.h. es gibt $m, M \in \mathbb{R}, s.d.$

$$m \leq R(\alpha) \leq M \quad \forall \alpha \in \Lambda$$

- Dann ist das ERM Prinzip genau dann strikt konsistent, wenn $R(\alpha)$ einseitig gleichmäßig gegen $R_{emp}(\alpha)$ konvergiert.

$$\inf_{\alpha \in \Lambda(c)} R_{emp}(\alpha) \xrightarrow[n \to \infty]{\mathbb{P}} \inf_{\alpha \in \Lambda(c)} R(\alpha) \quad \forall c \in \mathbb{R}$$

$$\Leftrightarrow$$

$$\mathbb{P}\left(\sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \varepsilon \right) \xrightarrow[n \to \infty]{} 0$$

Beweis: Key Theorem of Learning Theory (1)

- Sei T_k das Ereignis

$$\inf_{\alpha \in \Lambda(a_k)} R_{emp}(\alpha) < \inf_{\alpha \in \Lambda(a_k)} R(\alpha) - \frac{\varepsilon}{2}$$

- Sei $T = \bigcup_{k=1}^{N} T_k$
- A das Ereignis

$$\sup_{\alpha \in \Lambda} \left(R(\alpha) - R_{emp}(\alpha) \right) > \varepsilon$$

Beweis: Key Theorem of Learning Theory (2)

- Sei B das Ereignis

$$|\inf_{\alpha \in \Lambda(c)} R(\alpha) - \inf_{\alpha \in \Lambda(c)} R_{emp}(\alpha)| > \varepsilon$$

- Dann gilt $B = B_1 \cup B_2$ mit:

$$B_1 = \left\{ z \in \mathcal{Z} \middle| \inf_{\alpha \in \Lambda(c)} R(\alpha) + \varepsilon < \inf_{\alpha \in \Lambda(c)} R_{emp}(\alpha) \right\}$$

$$B_2 = \left\{ z \in \mathcal{Z} \middle| \inf_{\alpha \in \Lambda(c)} R(\alpha) - \varepsilon > \inf_{\alpha \in \Lambda(c)} R_{emp}(\alpha) \right\}$$

Entropie für Indikatorfunktionen

- Seien $Q(\cdot, \alpha),\ \alpha \in \Lambda$ Indikatorfunktionen
- Anzahl der verschiedenen Vektoren, die aus der Stichprobe $z_1, ..., z_n$ von $Q(\cdot, \alpha),\ \alpha \in \Lambda$ erzeugt werden können:

$$N^{\Lambda}(z_1, ..., z_n) := \#\{q(\alpha) | \alpha \in \Lambda\}, \quad q(\alpha) = Q(z_1, \alpha), ..., Q(z_n, \alpha)$$

- Zufällige Entropie

$$H^{\Lambda}(z_1, ..., z_n) := \log N^{\Lambda}(z_1, ..., z_n)$$

- Entropie

$$H^{\Lambda}(n) := \mathbb{E} H^{\Lambda}(z_1, ..., z_n)$$

Bedingungen gleichmäßige Konvergenz - Indikatorfunktionen

- Es gilt:

$$\sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| \xrightarrow[n \to \infty]{\mathbb{P}} 0$$

$$\Leftrightarrow$$

$$\lim_{n \to \infty} \frac{H^{\Lambda}(n)}{n} = 0$$

ε-Netze

- Sei (M, d) ein metrischer Raum, $G \subset M$.
- $B \subset M$ heißt ε-Netz, falls für alle $g \in G$ ein $b \in B$ existiert, so dass:

$$d(g, b) < \varepsilon$$

- G hat ein endliches ε-Netz, falls für alle $\varepsilon > 0$ ein ε-Netz B_ε mit endlich vielen Elementen existiert.
- Ein ε-Netz heißt minimal, falls es endlich ist und kein ε-Netz mit weniger Elementen existiert.

Entropie für gleichmäßig beschränkte Funktionen

- Seien $Q(\cdot, \alpha),\ \alpha \in \Lambda$ gleichmäßig beschränkt, d.h. es existiert ein $C \in \mathbb{R}$, s.d.

$$|Q(z, \alpha)| \leq C \quad \forall \alpha \in \Lambda, z \in \mathcal{Z}$$

- Für $\varepsilon > 0$ sei $N^{\Lambda}(\varepsilon; z_1, ..., z_n)$ die Anzahl der Elemente eines minimalen ε-Netzes der Menge $\{q(\alpha) | \alpha \in \Lambda\}$ (in der Maximumsnorm).
- zufällige ε-Entropie

$$H^{\Lambda}(\varepsilon; z_1, ..., z_n) = \log N^{\Lambda}(\varepsilon; z_1, ..., z_n)$$

- ε-Entropie

$$H^{\Lambda}(\varepsilon; n) = \mathbb{E} H^{\Lambda}(\varepsilon; z_1, ..., z_n)$$

Bedingungen gleichmäßige Konvergenz

- Seien $Q(\cdot, \alpha),\ \alpha \in \Lambda$ gleichmäßig beschränkt.
- Dann gilt:

$$\sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| \xrightarrow[n \to \infty]{\mathbb{P}} 0$$

$$\Leftrightarrow$$

$$\lim_{n \to \infty} \frac{H^{\Lambda}(\varepsilon; n)}{n} = 0 \quad \forall \varepsilon > 0$$

Zusammenfassung & Ausblick

- Das Key Theorem und die Einführung der Entropie liefern geeignete Kriterien, um die Konsistenz der ERM Prinzips zu garantieren.
- Diese benötigen kein / kaum Wissen über die unbekannte Wahrscheinlichkeitsverteilung $\mathbb{P}$.
- Es ist nur ein Modell (!)
- Ausblick: Wie schnell konvergiert das ERM Prinzip (Konvergenzrate)?

Danke für eure Aufmerksamkeit!