

Einführung in Statistisches Lernen

Vapnik, An Overview of Statistical Learning Theory (1999)

1 Situation des Lernproblems

1.1 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

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

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

1.2 Risikominimierung

- Verlustfunktion (loss function)

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

- Risikofunktional

$$R : \Lambda \rightarrow \mathbb{R}$$
$$\alpha \mapsto \int L(y, f(x, \alpha)) d\mathbb{P}(x, y)$$

- Ziel: Minimiere $R(\alpha)$ über Λ

1.3 Typische Lernprobleme

1.3.1 Mustererkennung

- $\mathcal{Y} = \{0, 1\}$
- $f(\cdot, \alpha)$, $\alpha \in \Lambda$ Indikatorfunktionen
- $L(y, f(x, \alpha)) = \begin{cases} 0 & \text{falls } y = f(x, \alpha) \\ 1 & \text{falls } y \neq f(x, \alpha) \end{cases}$
- $R(\alpha)$ gibt Wahrscheinlichkeit eines Klassifikationsfehlers an.

1.3.2 Regression

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

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

1.3.3 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:

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

1.4 Allgemeines Lernproblem

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

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

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

1.5 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 minimization induction principle (ERM principle)

2 Konsistenz des ERM Prinzips

2.1 Definition

- Das ERM heißt konsistent, falls

$$R(\alpha_n) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} R(\alpha_0)$$

und

$$R_{emp}(\alpha_n) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} R(\alpha_0)$$

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

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

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

2.2 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.

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

\Leftrightarrow

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

2.3 Bedingungen für gleichmäßige Konvergenz

2.3.1 Indikatorfunktionen

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

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

- Zufällige Entropie

$$H^\Lambda(z_1, \dots, z_n) := \log N^\Lambda(z_1, \dots, z_n)$$

- Entropie

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

- Es gilt:

$$\begin{aligned} \sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| &\xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0 \\ &\Leftrightarrow \\ \lim_{n \rightarrow \infty} \frac{H^\Lambda(n)}{n} &= 0 \end{aligned}$$

2.3.2 Beschränkte Funktionen

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

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

$$H^\Lambda(\varepsilon; z_1, \dots, z_n) = \log N^\Lambda(\varepsilon; z_1, \dots, z_n)$$

- ε -Entropie

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

- Dann gilt:

$$\begin{aligned} \sup_{\alpha \in \Lambda} |R(\alpha) - R_{emp}(\alpha)| &\xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0 \\ &\Leftrightarrow \\ \lim_{n \rightarrow \infty} \frac{H^\Lambda(\varepsilon; n)}{n} &= 0 \quad \forall \varepsilon > 0 \end{aligned}$$

3 Quellenverzeichnis

Vapnik, V. (1999). *An Overview of Statistical Learning Theory*. IEEE Transactions on Neural Networks 10, 988-999.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons Inc.