Schriftliche Ausarbeitung zum Seminarvortrag

Einführung in die Kerndichteschätzung

Yavuzâlp Dal

Stand: 14. Dezember 2025

1 Einleitung & Motivation

Inhaltsverzeichnis

Warum nichtparametrisch?

Klassische parametrische Verfahren (z. B. unter Annahme einer Normalverteilung) sind oft zu starr. Wenn die wahre Verteilung unbekannt ist, liefert die nichtparametrische Schätzung der Dichtefunktion f oft bessere Einblicke in die Datenstruktur (z. B. Schiefe, Multimodalität) als die reine Verteilungsfunktion F.
Ferner können wir eine Verteilung normalerweise nicht beweisen. Wir nehmen sie lediglich an. Folglich erzeugen wir hier eine extreme Kopplung zwischen (potentiell fehlerhaften) Interpretationen und unserer Schätzung. Offensichtlich ist es folglich für eine produktive Arbeit sinnvoll, diese Kopplung möglichst abzuschwächen; wobei wir im Nachfolgenden gesehen haben werden, dass sich bestimmte (schwache) Annahmen wie die Stetigkeit nicht zwingend vermeiden lassen, jedoch deren Einfluss i.d.R. nicht zu groß ist

Zielsetzung

Wir suchen eine Annäherung an den Wert $f(x)$ an einer Stelle x, ohne eine bestimmte parametrische Familie vorauszusetzen.

2 Der Rosenblatt-Schätzer (Die theoretische Basis)

Der Ausgangspunkt für moderne Dichteschätzer ist der Ansatz von Rosenblatt (1956).

- **Idee:** Da $f(x) = F'(x)$, nutzt man den Differenzenquotienten der empirischen Verteilungsfunktion F_n.
- **Der Schätzer:**

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} \tag{1}$$

  Dies entspricht der relativen Häufigkeit von Datenpunkten im Intervall $(x-h, x+h]$, geteilt durch die Länge $2h$.
- **Konsistenz:** Damit der Schätzer gegen die wahre Dichte konvergiert (MSE $\to 0$), muss für den Stichprobenumfang $n \to \infty$ gelten:
  1. Die Bandbreite $h \to 0$ (um den Bias zu reduzieren).
  2. Das Produkt $nh \to \infty$ (um die Varianz zu reduzieren).

3 Das Histogramm (Der Klassiker)

Das Histogramm ist die älteste Methode, hat aber methodische Schwächen für präzise Analysen.

- **Funktionsweise:** Zerlegung des Datenraums in Boxen (Bins) der Breite h. Die Höhe der Box entspricht der relativen Häufigkeit $\frac{n_k}{n}$ normiert durch h.
- **Probleme:**
  1. **Unstetigkeit:** Die geschätzte Dichte ist eine Treppenfunktion, obwohl die wahre Natur oft glatt (stetig differenzierbar) ist.
  2. **Subjektivität:** Form und Aussage hängen stark vom Startpunkt x_0 und der Klassenbreite h ab.
- **Optimale Bandbreite (h):** Die Wahl von h ist ein Balanceakt (Bias-Varianz-Tradeoff). Ein zu kleines h erzeugt eine „zackige“ Kurve (hohe Varianz), ein zu großes h glättet wichtige Details weg (hoher Bias).
  - *Faustregel (Scott):* $h \approx 3.49sn^{-1/3}$ (basiert auf Normalverteilungsannahme).
  - *Robuste Regel (Freedman-Diaconis):* Nutzt den Interquartilsabstand, um Ausreißer weniger zu gewichten:

$$h^* = 2(x_{0.75} - x_{0.25})n^{-1/3}$$

4 Kernschätzer (Kernel Density Estimation - KDE)

Um die Glattheit zu garantieren, verallgemeinert man den Rosenblatt-Ansatz durch sogenannte Kerne.

- **Der Schätzer:**

$$\hat{f}n(x) = \frac{1}{nh} \sum i = 1^n K\left(\frac{x - X_i}{h}\right) \tag{2}$$

Jeder Datenpunkt X_i bekommt eine kleine „Glockenkurve“ (Kern K), und die Summe dieser Kurven ergibt die Gesamtdichte.

- **Wahl des Kerns (K):** Der Kern muss zu 1 integrieren ($\int K(x)dx = 1$). Gängige Kerne sind:
  - *Rechteck-Kern:* Entspricht dem naiven Rosenblatt-Schätzer.
  - *Gauß-Kern:* Standardnormalverteilung (glatt und differenzierbar).
  - *Epanechnikov-Kern:* Parabolisch; dieser Kern ist theoretisch optimal, da er den integrierten mittleren quadratischen Fehler (IMSE) minimiert.
- **Bandbreitenwahl (Silverman's Rule):** Die Bandbreite h ist viel entscheidender als die Kernform. Eine Standard-Faustformel (für Gauß-Kerne) ist:

$$h_{opt} \approx 1.06\sigma n^{-1/5} \tag{3}$$

  Oft wird σ robust durch den Quartilsabstand geschätzt ($\hat{h}_{opt} \approx 0.79Qn^{-1/5}$), um Oversmoothing zu vermeiden.

5 Nichtparametrische Regression

Hier verlassen wir die Dichteschätzung und betrachten den Zusammenhang zwischen zwei Variablen X und Y: $Y = m(X) + \epsilon$.

5.1 Watson-Nadaraya Schätzer (Allgemeine Regression)

Wenn wir keine Formel für $m(x)$ kennen, schätzen wir den Wert als gewichteten Mittelwert der umliegenden Y-Werte:

$$\hat{m}WN(x) = \frac{\sum i = 1^n K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)} \tag{4}$$

Datenpunkte X_i, die nahe an x liegen, erhalten durch den Kern K ein hohes Gewicht.

5.2 Robuste Lineare Regression (Theil-Schätzer)

Selbst wenn wir einen linearen Trend ($Y = \alpha + \beta X$) vermuten, ist die klassische Methode der kleinsten Quadrate (OLS) anfällig für Ausreißer. Das Buch stellt robuste Alternativen vor:

1. **Theil-Methode I:** Man teilt die Daten in zwei Hälften und berechnet Steigungen zwischen Paaren $(i, i + n/2)$. Der Schätzer ist der Median dieser Steigungen.
2. **Theil-Sen-Schätzer (Methode II):** Dies ist die präzisere Variante. Man berechnet die Steigung zwischen **allen** möglichen Paaren $i < j$:

$$H_{ij} = \frac{Y_j - Y_i}{X_j - X_i} \tag{5}$$

   Der Schätzer für den Anstieg β ist der **Median** all dieser Steigungen (H_{ij}). Das macht ihn extrem robust.
3. **Hypothesentests:** Zum Testen von β wird auf rangbasierte Verfahren wie Kendalls τ (Tau) zurückgegriffen, da diese verteilungsfrei sind.

6 Zusammenfassung für das Seminar

- **Flexibilität:** Nichtparametrische Methoden (KDE, Kernel-Regression) passen sich den Daten an („let the data speak“) und zwingen ihnen keine Form auf.
- **Parameter h:** Die Wahl der Bandbreite ist der kritischste Schritt. Es ist ein Kompromiss zwischen Rauschen (zu kleines h) und Informationsverlust (zu großes h).
- **Robustheit:** Verfahren wie der Theil-Sen-Schätzer bieten mächtige Alternativen zur klassischen Regression, besonders wenn Daten Ausreißer enthalten oder nicht normalverteilt sind.