BZQ II: Stochastikpraktikum

Block 2: Nichtparametrische Methoden

Randolf Altmeyer

November 22, 2016

Überblick

1. Monte-Carlo-Methoden, Zufallszahlen, statistische Tests
2. **Nichtparametrische Methoden**
3. Lineares Modell, Klassifikation
4. Markov-Chain-Monte-Carlo-Verfahren
5. Simulation stochastischer Prozesse

Vielleicht: PCA, mehr Datenanalyse

Heute: Nichtparametrische Methoden

1. Dichteschätzung
2. Nichtparametrische Regression
3. Unbiased risk estimation (Kreuzvalidierung)
4. Nichtparametrische Tests

Literatur:

- Alexander Tsybakov, *Introduction to Nonparametric Statistics*
- John Kloke, Joseph McKean, *Nonparametric Statistical Methods in R*

Grundproblem der Statistik

- beobachte $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$, $\theta \in \Theta$ (= Parametermenge)
- **Problem**: Was ist das "wahre" θ?
- **Beispiele**:
  1. $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$, $\theta \in \times = \mathbb{R}$
  2. $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_f$ mit Wahrscheinlichkeitsdichte $f \in \Theta$, z.B. $\Theta = \{f : \int f \, dx = 1, \, f \geq 0, \, f \in C^1\}$
- wichtiges Merkmal in diesen Beispielen:
  1. $\dim(\Theta) = 1 < \infty \Rightarrow$ *parametrisches Problem*
  2. $\dim(\Theta) = \infty \Rightarrow$ *nichtparametrisches Problem*
- parametrisch vs. nichtparametrisch
  - param. Methoden brauchen mehr Annahmen (z.B. an Verteilung), konvergieren dann aber auch i.A. "schnell" ($n^{-1/2}$ = "parametrische Rate")
  - nichtparam. Methoden sind robust, konvergieren langsamer (Rate $> n^{-1/2}$)
  - Annahmen können falsch sein, so dass Methoden falsche Ergebnisse produzieren

Dichteschätzung

- beobachte $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_f$ mit Wahrscheinlichkeitsdichte f
- **Ziel**: approximiere f
- erste (naive) Idee: *Histogrammschätzer*
  - offset/Ursprung x_0, Bandweite $h > 0$
  - disjunkte Intervalle $I_j = (x_0 + j \cdot h, x_0 + (j+1)\, h]$
  - $\hat{f}_{hist}(x) = \frac{\#\{k : X_k \in I_j\}}{nh} = \frac{1}{nh} \sum_{k=1}^{n} \mathbf{1}(X_k \in I_j)$, wenn $x \in I_j$
    denn: $\mathbb{E}\left[\hat{f}_{hist}(x)\right] = \frac{1}{nh} \cdot n \cdot \mathbb{P}(X_1 \in I_j) = \frac{1}{h} \int_{x_0+jh}^{x_0+(j+1)h} f(y)\, dy \approx$
    $\frac{1}{h} \cdot f(x_0) \cdot h = f(x_0)$
- Probleme mit $\hat{f}_{hist}$:
  - nicht stetig (im Gegensatz zu f)
  - hängt von x_0 und h ab
  - selber Wert $\hat{f}_{hist}(x)$ für alle $x \in I_j$

Gleitender Histogramm-Schätzer

- **Idee:** Verteilungsfunktion $F(x) = \mathbb{P}(X_1 \leq x) = \int_{-\infty}^{x} f(y)dy$

$$
\begin{aligned}
f(x) = F'(x) &= \frac{1}{2}\left(F'(x) + F'(x)\right) \\
&= \frac{1}{2}\left(\lim_{h\to 0}\frac{F(x+h) - F(x)}{h} + \lim_{h\to 0}\frac{F(x) - F(x-h)}{h}\right) \\
&= \lim_{h\to 0}\frac{F(x+h) - F(x-h)}{2h} = \lim_{h\to 0}\frac{1}{2h}\mathbb{P}(x-h < X \leq x+h) \\
\Rightarrow \hat{f}_n(x) &= \frac{1}{2nh}\#\{k : X_k \in (x-h, x+h]\} \\
&= \frac{1}{nh}\sum_{k=1}^{n}\left(\frac{1}{2}\mathbf{1}\left(\left|\frac{X_k - x}{h}\right| \leq 1\right)\right)
\end{aligned}
$$

- Vergleich mit Histogramm-Schätzer:
  - $\hat{f}_n(x)$ immer noch unstetig
  - Schätzung sieht "realistischer" aus
  - hängt immer noch von h ab

Kerndichteschätzer

- $\hat{f}_n(x) = \frac{1}{nh} \sum_{k=1}^{n} \left(\frac{1}{2} \mathbf{1} \left(\left| \frac{X_k - x}{h} \right| \leq 1 \right) \right) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{X_k - x}{h} \right)$ mit *Kern* $K(y) = \frac{1}{2} \cdot \mathbf{1}_{[-1,1]}(y)$
- **Kerndichteschätzer**: für allgemeine Funktion $K : \mathbb{R} \to \mathbb{R}$,

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{X_k - x}{h} \right)$$

- K heißt *Kern*, wenn $\int K(x)dx = 1$,
- $\hat{f}_n$ übernimmt Eigenschaften von K:
  - wenn $K \geq 0$, dann ist $\hat{f}_n$ Dichte:
    $\int \hat{f}_n(x)\, dx = \frac{1}{nh} \sum_{k=1}^{n} \int K\left(\frac{X_k - x}{h} \right) dx = \frac{1}{n} \sum_{k=1}^{n} \int K(x)\, dx = 1$
  - $\hat{f}_n$ ist so "glatt" wie K
- **Typische Kerne**:
  - Rechteckskern: $K(y) = \frac{1}{2} \cdot \mathbf{1}_{[-1,1]}(y)$
  - Gaußkern: $K(y) = (2\pi)^{-\frac{1}{2}} e^{-y^2/2}$
  - Epanechnikov-Kern: $K(y) = 3/4 \left(1 - |y|^2\right)_+$

Fehleranalyse

- wähle $h > 0$ so, dass der **mean integrated squared error** minimal wird:

$$\text{MISE} = \mathbb{E}\left[\left\|\hat{f}_n - f(x)\right\|_{L^2}^2\right] = \mathbb{E}\left[\int \left(\hat{f}_n(x) - f(x)\right)^2 dx\right]$$

- **Bias-Varianz-Zerlegung:**

$$\int \mathbb{E}\left[\left(\hat{f}_n(x) - f(x)\right)^2\right] dx = \int \left(\left(\mathbb{E}\left[\hat{f}_n(x)\right] - f(x)\right)^2 + Var\left(\hat{f}_n(x)\right)\right) dx$$
$$=: \int \text{BIAS}^2(x)\, dx + \int \text{VAR}(x)\, dx$$

- **Varianz**:

$$\text{VAR}(x) = Var\left(\frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{X_k - x}{h}\right)\right) = \frac{1}{n^2h^2} \cdot n \cdot Var\left(K\left(\frac{X_1 - x}{h}\right)\right)$$
$$\leq \frac{1}{nh^2}\mathbb{E}\left(K^2\left(\frac{X_1 - x}{h}\right)\right) \leq \frac{1}{nh^2}\int K^2\left(\frac{y - x}{h}\right) f(y)\, dy$$
$$= \frac{1}{nh}\int K^2(y)\, f(x + hy)\, dy$$

$$\Rightarrow \int \text{VAR}(x)\, dx = \frac{1}{nh}\int K^2(y)\, dy$$

Fehleranalyse

- **Bias**:

$$\begin{aligned}\mathrm{BIAS}(x) = \mathbb{E}\left[\hat{f}_n(x)\right] - f(x) &= \frac{1}{nh} \cdot n \cdot \mathbb{E}\left[K\left(\frac{X_1 - x}{h}\right)\right] - f(x) \\ &= \int K(y)\left(f(x+hy) - f(x)\right) dy\end{aligned}$$

- wenn $f \in C^1$ mit kompaktem Träger $\mathrm{supp}(f) \subset A$, dann ist also

$$\int \mathrm{BIAS}^2(x)\, dx \leq \left(\sup_{x \in A} \left|f'(x)\right|\right) h^2 \int_A \left(\int K(y)\,|y|\, dy\right)^2 dx \lesssim h^2 \left(\int K(y)\,|y|\, dy\right)^2$$

- **Bias-Varianz-Dilemma**:

$$\begin{aligned}\mathrm{MISE} = \mathbb{E}\left[\int \left(\hat{f}_n(x) - f(x)\right)^2 dx\right] &= \int \mathrm{BIAS}^2(x)\, dx + \int \mathrm{VAR}(x)\, dx \\ &\lesssim h^2 \left(\int K(y)\,|y|\, dy\right)^2 + \frac{1}{nh} \int K^2(y)\, dy = O\left(h^2 + \frac{1}{nh}\right)\end{aligned}$$

  - h klein: Bias klein, Varianz groß $\Rightarrow$ $\hat{f}_n$ oszilliert stark (Unterglättung)
  - h groß: Bias groß, Varianz klein $\Rightarrow$ starke Abweichung von der "Wahrheit" (Überglättung)

Nichtparametrische Regression

- **beobachte** $(x_1, Y_1), \ldots, (x_n, Y_n)$, wobei

$$Y_i = f(x_i) + \varepsilon_i \in \mathbb{R}^d,$$

  f unbekannte Funktion, ε_i unabhängig und $\mathbb{E}[\varepsilon_i] = 0$
- **Ziel**: approximiere $f(x)$ für $x \neq x_i$
- **Idee:**

$$\hat{f}_n(x) = \operatorname{argmin}_{y \in \mathbb{R}^d} \underbrace{\left(\sum_{k=1}^{n} \| Y_k - y \|^2 w_k(x) \right)}_{= " \mathbb{E}_{w(x)} \left[\| Y - \mathbf{y} \|^2 \right] "}$$

  für *lokale Gewichte* $w_k(x)$ mit $\sum_k w_k(x) = 1$
- **explizite Lösung:** $\hat{f}_n(x) = \sum_{k=1}^{n} Y_k w_k(x)$ (= Erwartungswert von $Y = (Y_1, \ldots, Y_n)$ bezüglich dem Maß $w(x) = (w_1(x), \ldots, w_n(x))$)

Nichtparametrische Regression

- **Beispiele**:
  - $w_k(x) = \frac{1}{n} \Rightarrow$ beste globale Vorhersage von $f(x)$ ist der (globale) Mittelwert $\frac{1}{n}\sum_{k=1}^{n} Y_k$
  - für Kern K mit $\int K(y)dy = 1$ und Bandweite $h > 0$ definiere

$$w_k(x) := \frac{\frac{1}{h}K(\frac{x_k - x}{h})}{\frac{1}{h}\sum_{m=1}^{n} K\left(\frac{x_l - x}{h}\right)}$$

  $\Rightarrow (x_k, Y_k)$ für x_k nahe bei x sind wichtiger für die Vorhersage von $f(x)$
  $\Rightarrow$ "Nadaraya-Watson-Schätzer":

$$\hat{f}_n(x) = \sum_{k=1}^{n} Y_k \frac{\frac{1}{h}K(\frac{x_k - x}{h})}{\frac{1}{h}\sum_{m=1}^{n} K\left(\frac{x_m - x}{h}\right)}$$

  - Fehleranalyse ähnlich wie bei Dichteschätzung, selbes Konvergenzverhalten, abhängig von Bandweite (**Plots**!)

Dichteschätzung vs. Nichtparam. Regression

- **Dichteschätzung**:
  - beobachte $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_f$ für Dichte f
  - Kerndichteschätzer:

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{X_k - x}{h}\right)$$

- **Nichtparametrische Regression**:
  - beobachte $Y_i = f(x_i) + \varepsilon_i$ für unbekannte Funktion f und zufällige Fehler ε_i
  - Nadaraya-Watson-Schätzer:

$$\hat{f}_n(x) = \sum_{k=1}^{n} Y_k \frac{\frac{1}{h} K\left(\frac{x_k - x}{h}\right)}{\frac{1}{h} \sum_{m=1}^{n} K\left(\frac{x_m - x}{h}\right)}$$

- **Frage**: Wie wählt man h?

Bandweitenwahl für Dichteschätzung

- bei Dichteschätzung und C^1-Dichte: $\mathrm{MISE} = O\left(h^2 + \frac{1}{nh}\right) \Rightarrow$ $h^* = O(n^{-1/3}) \Rightarrow \mathrm{MISE} = O(n^{-2/3})$
- Problem für dieses h^*:
  - benötigt Wissen über die Glattheit von f
  - hängt von (eventuell sehr großen) Konstanten ab
- mögliche Lösungen:
  - treffe Annahmen an f (z.B. C^1 mit beschränkter Ableitung)
  - Faustregeln (z.B. Silverman's rule of thumb) $\Rightarrow$ schlechte Idee!
  - besser: datengetriebene Wahl!

Kreuzvalidierung

- suche h so, dass $\text{MISE}(h) = \mathbb{E}\left[\int \left(\hat{f}_n(x) - f(x)\right)^2 dx\right]$ minimial wird
- es gilt:

$$\begin{aligned}\text{argmin}_h \text{MISE}(h) &= \text{argmin}_h \left(\mathbb{E}\left[\int \hat{f}_n^2 - 2\int \hat{f}_n f\right] + \int f^2\right) \\ &= \text{argmin}_h \underbrace{\mathbb{E}\left[\int \hat{f}_n^2 - 2\int \hat{f}_n f\right]}_{J(h)}\end{aligned}$$

- suche einen erwartungstreuen Schätzer $\hat{J}(h)$ von $J(h)$ und wähle $h^* = \text{argmin}_h \hat{J}(h)$
- erwartungstreuer Schätzer von $\mathbb{E}\left[\int \hat{f}_n^2(x)dx\right]$ ist $\int \hat{f}_n^2(x)dx$
- erwartungstreuer Schätzer von $\mathbb{E}\left[\int \hat{f}_n(x)f(x)dx\right]$ ist (check!)

$$\hat{G} = \frac{1}{n}\sum_{j=1}^{n} \hat{f}_{n,-j}(X_j), \quad \hat{f}_{n,-j}(x) = \frac{1}{(n-1)h}\sum_{k\neq j} K\left(\frac{X_k - x}{h}\right)$$

- also ist $\hat{J}(h) = \int \hat{f}_n^2(x)dx - 2\hat{G}$, $\hat{J}(\cdot)$ heißt *Kreuzvalidierungskriterium*

Statistische Tests: Wiederholung

- *Parametermenge* $\Theta = H_0 \uplus H_1$ für *Nullhypothese* H_0 und *Alternative* H_1
- Zufallsvariable $\varphi_n : \Omega \to [0,1]$ heißt *Test* von H_0 gegen H_1, wenn $\varphi_n(X_1, \ldots, X_n) = 1$ bedeutet, dass H_0 verworfen wird (d.h. $\theta \in H_1$) und $\varphi_n(X_1, \ldots, X_n) = 0$ bedeutet, dass H_0 nicht verworfen wird (d.h. $\theta \in H_0$).
- wähle φ_n so, dass
  - $\mathbb{E}_\theta(\varphi_n(X_1, \ldots, X_n)) \leq \alpha$ für kleines $\alpha > 0$ und $\theta \in H_0$ (d.h. φ_n hat Signifikanzniveau/Level α)
  - $\mathbb{E}_\theta(\varphi_n(X_1, \ldots, X_n))$ ist maximal, wenn $\theta \in H_1$ (Power von φ_n)
- oft: $\varphi_n(X_1, \ldots, X_n) = \mathbf{1}(T(X_1, \ldots, X_n) \geq c_\alpha)$ für *kritische* Werte c_α und Statistik $T(X_1, \ldots, X_n)$, d.h. lehne ab wenn $T(X_1, \ldots, X_n)$ zu groß wird

Statistische Tests: Wiederholung

- in vielen Modellen, und wenn $H_0 = \{\theta_0\}$ einelementig, gilt (siehe *VL Mathematische Statistik*):
  - der "beste Test" hat die Form
    $\varphi_n(T(X_1, \ldots, X_n)) = \mathbf{1}(p(X_1, \ldots, X_n) < \alpha)$
  - mit $p(x) = \mathbb{P}_{\theta_0}(T(X_1, \ldots, X_n) \geq T(x))$
- **Nichtparametrischer Test** = die Verteilung $\mathbb{P}_\theta^{\varphi_n}$ für $\theta \in H_0$ ist unabhängig von $\mathbb{P}_\theta^{(X_1,\ldots,X_n)}$, d.h. der Test ist *verteilungsfrei*
  $\Rightarrow$ Test ist robuster da keine Annahmen, aber oft kleine Power

Vergleich von zwei Samples

- beobachte Paare $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P_\theta$, wobei $X_i - Y_i = \theta + \varepsilon_i$, ε_i symmetrisch
- Beispiel: (X_i, Y_i) = Intelligenzquotienten von Zwillingen, wobei *zufällig* einer der beiden Zwillinge eine "spezielle" Behandlung erhält (z.B. Hausaufgabenbetreuung)
- Teste

$$H_0 : \theta = 0 \text{ "Behandlung hat keinen Effekt"}$$
$$H_1 : \theta \neq 0 \text{ "Behandlung hat Effekt"}$$

- **1. Lösung:** T-Test
  - nimm an, dass $X_i - Y_i \sim N(\theta, \sigma^2)$
  - dann gilt für $\theta = 0$: $t = \frac{\bar{X} - \bar{Y}}{s/\sqrt{n}} \overset{d}{\sim} t(n-1)$ ($t(n-1)$ = Student-t-Verteilung mit $n-1$ Freiheitsgraden, s = Standardabweichung von $(X_i - Y_i)_{i=1,\ldots,n}$)
  - lehne ab, wenn t zu groß wird (einseitiger Test) bzw. wenn $|t|$ zu groß wird (zweiseitiger Test)

Vergleich von zwei Samples

- **2. Lösung:** Wilcoxon-Vorzeichen-Test
  - keine Verteilungsannahme
  - definiere die Vorzeichen $S_i = \text{sign}(X_i - Y_i)$ und Rangstatistiken R_i (Position von $|X_i - Y_i|$ wenn aufsteigend angeordnet)
  - $W = \sum_{k=1}^{n} S_i R_i$, wegen Symmetrie und Unabhängigkeit der $X_i - Y_i$ kann Verteilung von W explizit bestimmt werden (aufwändig!)
  - lehne ab, wenn W zu groß wird (einseitiger Test) bzw. wenn $|W|$ zu groß wird (zweiseitiger Test)