# تکلیف پنجم درس شناسایی الگو

وحید ملکی
شماره دانشجویی: ۴۰۳۱۳۰۰۴

۲۴ آبان ۱۴۰۴

## سؤال ۷

### داده‌ها

مجموعه داده یک‌بعدی به صورت زیر است:

$$X = \{0, 1, 1, 1, 2, 2, 2, 2, 3, 4, 4, 4, 5\}, \quad N = 13.$$

تکرار هر مقدار به شرح زیر است:

- 0: یک بار
- 1: سه بار
- 2: چهار بار
- 3: یک بار
- 4: سه بار
- 5: یک بار

### الف) رسم هیستوگرام با پهنای بازه ۱

پهنای بازه $h = 1$ و مراکز بازه‌ها در نقاط $\{0, 1, 2, 3, 4, 5\}$ قرار دارند. بنابراین بازه‌ها به صورت زیر تعریف می‌شوند:

(مرکز ۰) $[-0.5, 0.5)$
(مرکز ۱) $[0.5, 1.5)$
(مرکز ۲) $[1.5, 2.5)$
(مرکز ۳) $[2.5, 3.5)$
(مرکز ۴) $[3.5, 4.5)$
(مرکز ۵) $[4.5, 5.5)$

تعداد نمونه‌ها در هر بازه:

$$\begin{aligned}
&\text{مرکز ۰}: \quad (x=0)\\
&\text{مرکز ۱}: \quad ()\\
&\text{مرکز ۲}: \quad ()\\
&\text{مرکز ۳}: \quad (x=3)\\
&\text{مرکز ۴}: \quad ()\\
&\text{مرکز ۵}: \quad (x=5)
\end{aligned}$$

تخمین چگالی هیستوگرام در هر بازه:

$$\hat{f}_{\text{hist}}(x) = \frac{\text{تعداد در بازه}}{N \cdot h}, \quad h = 1, \quad N = 13$$

بنابراین:

$$\begin{aligned}
&\hat{f}_{\text{hist}}(0) = \frac{1}{13}, && \hat{f}_{\text{hist}}(1) = \frac{3}{13}, && \hat{f}_{\text{hist}}(2) = \frac{4}{13},\\
&\hat{f}_{\text{hist}}(3) = \frac{1}{13}, && \hat{f}_{\text{hist}}(4) = \frac{3}{13}, && \hat{f}_{\text{hist}}(5) = \frac{1}{13}.
\end{aligned}$$

هیستوگرام یک تابع پله‌ای است که در هر بازه مقدار ثابت دارد.

## ب) رابطه کلی تخمین پارزن

تخمین چگالی پارزن با کرنل دلخواه $K$ و پهنای پنجره $h$ به صورت زیر است:

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right).$$

## ج) تخمین پارزن با کرنل مثلثی

کرنل مثلثی تعریف شده است:

$$K(u) = (1 - |u|) \cdot \delta(|u| \le 1),$$

که $\delta(|u| \le 1)$ نشانگر است (۱ اگر $|u| \le 1$، در غیر این صورت ۰). همچنین:

$$u = \frac{x - x_i}{h}.$$

با انتخاب $h = 1$ (مانند هیستوگرام):

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} (1 - |x - x_i|) \cdot \mathbf{1}(|x - x_i| \le 1).$$

حال تخمین در نقاط $x \in \{0, 1, 2, 3, 4, 5\}$ محاسبه می‌شود:

- $x = 0$: نمونه‌های در فاصله $\le 1$: $x_i = 0, 1$ وزن‌ها: $1 - |0 - 0| = 1$، $1 - |1 - 0| = 0$ مجموع وزن = $1 \Rightarrow \hat{f}(0) = \frac{1}{13}$

- $x = 1$: نمونه‌های $x_i = 0, 1, 2$ وزن‌ها: 0، 1 (سه بار)، 0 مجموع وزن = 3 $\Rightarrow \hat{f}(1) = \frac{3}{13}$
- $x = 2$: نمونه‌های $x_i = 1, 2, 3$ وزن‌ها: 0 (سه بار)، 1 (چهار بار)، 0 مجموع وزن = 4 $\Rightarrow \hat{f}(2) = \frac{4}{13}$
- $x = 3$: نمونه‌های $x_i = 2, 3, 4$ وزن‌ها: 0 (چهار بار)، 1، 0 (سه بار) مجموع وزن = 1 $\Rightarrow \hat{f}(3) = \frac{1}{13}$
- $x = 4$: نمونه‌های $x_i = 3, 4, 5$ وزن‌ها: 0، 1 (سه بار)، 0 مجموع وزن = 3 $\Rightarrow \hat{f}(4) = \frac{3}{13}$
- $x = 5$: نمونه‌های $x_i = 4, 5$ وزن‌ها: 0 (سه بار)، 1 مجموع وزن = 1 $\Rightarrow \hat{f}(5) = \frac{1}{13}$

بنابراین:

$$\hat{f}(0, 1, 2, 3, 4, 5) = \left( \frac{1}{13}, \frac{3}{13}, \frac{4}{13}, \frac{1}{13}, \frac{3}{13}, \frac{1}{13} \right).$$

این مقادیر دقیقاً با ارتفاع هیستوگرام در مراکز بازه‌ها یکسان هستند.

## د) تخمین MLE توزیع گوسی

فرض می‌کنیم داده‌ها از توزیع نرمال $X \sim \mathcal{N}(\mu, \sigma^2)$ هستند. تخمین‌های بیشینه درست‌نمایی:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i, \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2.$$

مجموع داده‌ها:

$$\sum x_i = 0 + 3 \times 1 + 4 \times 2 + 3 + 3 \times 4 + 5 = 31 \quad \Rightarrow \quad \hat{\mu} = \frac{31}{13}.$$

مجموع مربعات انحراف:

$$\sum (x_i - \hat{\mu})^2 = \frac{352}{13} \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{13} \cdot \frac{352}{13} = \frac{352}{169}.$$

پس توزیع گوسی تخمین‌زده شده:

$$X \sim \mathcal{N}\left( \frac{31}{13}, \frac{352}{169} \right).$$

## ه) مقایسه روش‌ها

- **هیستوگرام**: ساده، مستقیم فرکانس‌های تجربی را نشان می‌دهد. **ضعف**: ناپیوسته، وابستگی زیاد به عرض و موقعیت بازه‌ها، نویز در داده‌های کم. **کاربرد**: تحلیل اکتشافی سریع.
- **پارزن با کرنل مثلثی**: چگالی پیوسته و خطی-تکه‌ای تولید می‌کند. بدون فرض پارامتری. **ضعف**: نیاز به انتخاب $h$، بیش‌برازش با $h$ کوچک، بیش‌صاف کردن با $h$ بزرگ. **کاربرد**: داده‌های کافی، عدم اطمینان به شکل پارامتری.
- **MLE گوسی**: مدل پارامتری بسیار صاف با دو پارامتر. **ضعف**: اگر توزیع واقعی چندوجهی یا غیرنرمال باشد، برازش ضعیف دارد. **کاربرد**: وقتی فرض نرمالی معتبر است و نمونه متوسط است.

در این داده، توزیع تجربی تک‌قله در اطراف ۲ دارد. **نتیجه**: هیستوگرام و پارزن دقیقاً الگوی داده را در نقاط شبکه بازتولید می‌کنند اما صاف نیستند. گوسی یک خلاصه صاف مناسب ارائه می‌دهد اما جزئیات محلی را از دست می‌دهد. در عمل، هیستوگرام برای نمایش سریع، پارزن برای تخمین غیرپارامتری، و گوسی برای مدل‌سازی پارامتری مناسب است.