TRATAMIENTO DIGITAL DE SEÑALES
Ingeniería de Telecomunicación (4º, 2º c)

Unidad 8ª: Estimación de Densidades de Probabilidad

Aníbal R. Figueiras Vidal
Jesús Cid Sueiro
Ángel Navia Vázquez

Área de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid

Revisión de tipos de procedimientos

* **Paramétricos**: en los que se acepta una forma analítica para la ddp (p.ej., gaussiana) y se estiman sus parámetros (media y varianza, en el ejemplo)

Procedimientos sencillos, pero poco robustos.

* **Semiparamétricos**: en los que la ddp se modela según una expresión con suficiente capacidad de aproximación (habitualmente una mezcla: p. ej. , de gaussianas), y se estiman sus parámetros (p.ej., vía Algoritmo EM).

Son procedimientos eficaces, pero no exentos de dificultades; como:

- determinación del “orden”
- convergencia

No paramétricos: no se realiza asunción alguna sobre la forma de la ddp, y se recurre a estimarla a partir de la aplicación de propiedades generales de la ddp.

Procedimientos robustos: pero también con sus inconvenientes, como:

- alta necesidad de memoria en diseño
- alta carga computacional en aplicación

Introducción a los procedimientos no paramétricos

El procedimiento clásico por antonomasia, en situaciones uni y bidimensionales (para visualizar el resultado), es el **histograma**: se divide el espacio muestral en segmentos o cuadrados, normalmente de dimensión fija, y se asigna a cada uno una altura uniforme igual al número de muestras que contiene, k (proporciona así la forma de la ddp).

Inconvenientes:

- la aproximación es escalonada
- las regiones básicas utilizadas (segmentos, cuadrados, cubos, ... hipercubos) son inadecuadas en muchas dimensiones: un hipercubo, al crecer las dimensiones, reduce mucho su volumen respecto a su lado, y toma aspecto de “puercoespín”; lo que es, obviamente, inapropiado para asignar los datos (puede incluir datos muy distantes y separar datos muy próximos)

Es importante, en cualquier caso, reflexionar sobre la teoría subyacente en la aplicación de los histogramas: en definitiva, toman una región y le asignan el número de muestras que contiene. Lo que se entiende considerando que, en general, la probabilidad de encontrarse en una región R de una ddp es

$$P = \int_R p(\mathbf{x}')\, d\mathbf{x}'$$

que, de acuerdo con la definición de valor medio de una integral definida, puede expresarse

$$P = p(\mathbf{x})\, V_R$$

siendo V_R el volumen de R y $\mathbf{x}$ un cierto punto interior a R

Aceptando que P se puede aproximar por la “frecuencia relativa” de las muestras en R, resulta

$$p(\mathbf{x})V_R \approx k/K$$

siendo k el número de muestras en R y K el número total de muestras; de donde

$$p(\mathbf{x}) \approx \frac{k}{KV_R}$$

(forma que tomaría el histograma si se fuerza área unitaria para p(**x**)).

Los métodos particulares dependen de la elección de R.

Debe resaltarse que se establece un compromiso:

- se produce un efecto de promediado espacial, al tomar **x** como un valor medio integral: por lo que conviene emplear regiones R pequeñas;

- pero, al mismo tiempo, para que la aproximación de P sea razonable, ha de mantenerse k/K lejos de 0: lo que supone que R no puede hacerse indefinidamente pequeña;

por lo que

- conviene que K sea alto: lo que origina cargas elevadas;

- conviene elegir las regiones para mantener k razonablemente alto: es decir, en función de la localización de los datos (a diferencia de los histogramas)

El procedimiento de las **Ventanas (Núcleos) de Parzen**

Si ϕ es la función indicadora de un hipercubo de lado 1 centrado en el origen

$$\phi(\mathbf{x}) = \begin{cases} 1, & \text{si } |x_n| < 1/2, \quad \forall n \text{ (componentes)} \\ 0, & \text{caso contrario} \end{cases}$$

es obvio que

$$\phi\left(\frac{\mathbf{x}-\mathbf{x}^{(k')}}{h}\right) = \begin{cases} 1, & \text{si } \left|x_n - x_n^{(k')}\right| < h/2, \quad \forall n \\ 0, & \text{caso contrario} \end{cases}$$

de forma que, dado **x**, cuenta 1 si $x^{(k')}$ está en el hipercubo, y 0 si no; así:

$$k = \sum_{k'=1}^{K} \phi\left(\frac{\mathbf{x}-\mathbf{x}^{(k')}}{h}\right)$$

con lo que se tiene el estimador

$$\hat{p}(\mathbf{x}) = \frac{1}{Kh^N} \sum_{k'=1}^{K} \phi\left(\frac{\mathbf{x}-\mathbf{x}^{(k')}}{h}\right)$$

($V_R = h^N$, para el hipercubo)

Esta forma de actuar soluciona el problema de los histogramas relacionado con prefijar regiones, ya que el hipercubo se centra en el **x** considerado; pero no las malas características de los hipercubos. Por ello, se suaviza la forma de éstos, manejando **ventanas** o **núcleos** más "lisos", con las restricciones de

- no negatividad: $\phi(\mathbf{x}) \geq 0$
- volumen unitario: $\int \phi(\mathbf{x}) d\mathbf{x} = 1$

y así se obtienen los **estimadores de Parzen**

$$\hat{p}(\mathbf{x}) = \frac{1}{Kh^N} \sum_{k'=1}^{K} \phi\left(\frac{\mathbf{x} - \mathbf{x}^{(k')}}{h} \right)$$

(la elección de h establece el compromiso entre resolución y consistencia que se ha discutido).

Es muy frecuente utilizar ventanas gaussianas

$$\phi\left(\mathbf{x}-\mathbf{x}^{(k')}\right)=\frac{1}{(2\pi)^{N/2}|V|^{1/2}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{x}^{(k')}\right)^{T}V^{-1}\left(\mathbf{x}-\mathbf{x}^{(k')}\right)\right]$$

(h=1: es V lo que marca la concentración), y sobre todo sus versiones circulares

$$\phi\left(\mathbf{x}-\mathbf{x}^{(k')}\right)=\frac{1}{(2\pi v)^{N/2}}\exp\left[-\frac{1}{2v}\left\|\mathbf{x}-\mathbf{x}^{(k')}\right\|^{2}\right]$$

como ventanas para este procedimiento.

No debe olvidarse que la elección de v marcará el comportamiento del estimador: “picudo” si es pequeña, muy alisado si es grande.

El procedimiento de los **k Vecinos más Próximos** (k-NN) ("Nearest Neighbours")

En este caso, V_R se fija en función de las posiciones de los datos: el volumen (de una región R de forma razonable: hiperesférica, p.ej.) en torno a **x** en que haya k muestras; y así

$$\hat{p}(\mathbf{x}) = \frac{k}{KV_R}$$

Nótese que es V_R lo que varía en esta expresión.

Hay métodos rápidos para seleccionar las k muestras más próximas.

T: Estúdiese un procedimiento rápido para aplicar k-NN.

Estimación no paramétrica de las probabilidades a posteriori

A. Vía k-NN

Si suponemos una situación en que las muestras pertenezcan a clases C_c, y se toma

$$\hat{p}(\mathbf{x}) = \sum_{c'=1}^{C} \hat{p}(\mathbf{x}, C_{c'}) = \frac{k}{KV_R}$$

y, con la misma región, dividiendo proporcionalmente,

$$\hat{p}(\mathbf{x}, C_c) = \frac{k_c}{KV_R}$$

resulta razonable tomar

$$\hat{Pr}(C_c|\mathbf{x}) = \frac{\hat{p}(\mathbf{x}, C_c)}{\hat{p}(\mathbf{x})} = \frac{k_c}{k}$$

B. Vía Ventanas de Parzen

Procediendo análogamente

$$\hat{p}(\mathbf{x}) = \frac{1}{Kh^N} \sum_{k'=1}^{K} \phi\left(\frac{\mathbf{x} - \mathbf{x}^{(k')}}{h} \right)$$

y

$$\hat{p}(\mathbf{x}, C_c) = \frac{1}{Kh^N} \sum_{k'=1}^{K_c} \phi\left(\frac{\mathbf{x} - \mathbf{x}_c^{(k')}}{h} \right)$$

donde $\mathbf{x}_c$ indica pertenencia a C_c, y K_c es el número de muestras de la clase c; a partir de ello

$$\hat{P}r(C_c|\mathbf{x}) = \frac{\sum_{k'=1}^{K_c} \phi\left(\frac{\mathbf{x} - \mathbf{x}_c^{(k')}}{h} \right)}{\sum_{k'=1}^{K} \phi\left(\frac{\mathbf{x} - \mathbf{x}^{(k')}}{h} \right)}$$

C. Vía estimadores semiparamétricos (mezclas)

Resulta paralelo al caso de las ventanas; así, para mezclas de gaussianas,

$$\hat{p}(\mathbf{x}, C_c) = \sum_{j=1}^{J_C} \rho_c^{(j)} \frac{1}{(2\pi)^{N/2} \left|\mathbf{V}_c^{(j)}\right|^{1/2}} \exp\left\{ -\frac{1}{2} \left(\mathbf{x} - \mathbf{m}_c^{(j)}\right)^T \left[\mathbf{V}_c^{j}\right]^{-1} \left(\mathbf{x} - \mathbf{m}_c^{(j)}\right) \right\}$$

y de ello

$$\hat{P}r(C_c|\mathbf{x}) = \frac{\displaystyle\sum_{j=1}^{J_C} \rho_c^{(j)} \frac{1}{(2\pi)^{N/2} \left|\mathbf{V}_c^{(j)}\right|^{1/2}} \exp\left\{ -\frac{1}{2} \left(\mathbf{x} - \mathbf{m}_c^{(j)}\right)^T \left[\mathbf{V}_c^{j}\right]^{-1} \left(\mathbf{x} - \mathbf{m}_c^{(j)}\right) \right\}}{\displaystyle\sum_{c=1}^{C} \sum_{j=1}^{J_C} \rho_c^{(j)} \frac{1}{(2\pi)^{N/2} \left|\mathbf{V}_c^{(j)}\right|^{1/2}} \exp\left\{ -\frac{1}{2} \left(\mathbf{x} - \mathbf{m}_c^{(j)}\right)^T \left[\mathbf{V}_c^{j}\right]^{-1} \left(\mathbf{x} - \mathbf{m}_c^{(j)}\right) \right\}}$$

Ejemplo de Ampliación

A: Decisión vía estimación NP de las ddp

Una vez obtenidas las $\hat{Pr}(C_c|\boldsymbol{x})$ *, puede aplicarse la Teoría Bayesiana para decidir.*

Establezca los modos de hacerlo para situaciones MAP a partir de los tres casos anteriores.

¿Cuál es la principal característica de estos métodos?

Al tratarse de una situación MAP, el proceso a seguir será

$$c^* = arg\left[\max_c\left\{Pr(C_c|\boldsymbol{x})\right\}\right]$$

de modo que se tiene:

A.- k-NN:

$$c^* = arg\left[\max_c\left\{k_c\right\}\right]$$

B.- Parzen:

$$c^* = arg\left[\max_c\left\{\sum_{k'=1}^{K}\phi\left(\frac{\boldsymbol{x}-\boldsymbol{x}_c^{(k')}}{h}\right)\right\}\right]$$

C.- Mezclas gaussianas:

$$c^* = arg\left[\max_c\left\{\sum_{j=1}^{J_c}\rho_c^{(j)}\frac{1}{(2\pi)^{N/2}\left|\boldsymbol{V}_c^{(j)}\right|^{1/2}}exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{m}_c^{(j)}\right)^T\left[\boldsymbol{V}_c^{(j)}\right]^{-1}\left(\boldsymbol{x}-\boldsymbol{m}_c^{(j)}\right)\right\}\right\}\right]$$

La característica de estos métodos es que ninguno se diseña orientado a la decisión, sino a partir de la estimación de las ddp: lo que es especialmente limitativo en el caso semiparamétrico, ya que los $\rho_c^{(j)}$, $\boldsymbol{m}_c^{(j)}$ y $\boldsymbol{V}_c^{(j)}$ se establecen con un objetivo distinto de la propia clasificación (ML, típicamente). No obstante, la decisión k-NN, por su sencillez y fácil interpretación, es un método acreditado.

La estimación de Nadaraya-Watson

Si se aproxima la ddp de (**s**, **x**) (dimensiones N_1, N_2, respectivamente) mediante ventanas de Parzen gaussianas circulares (factorizables) a partir de (pares de) muestras, se tendrá:

$$\hat{p}(\mathbf{s},\mathbf{x}) = \frac{1}{K}\sum_{k'=1}^{K} \frac{1}{(2\pi v)^{(N_1+N_2)/2}} \exp\left(-\frac{1}{2v}\left\|\mathbf{s}-\mathbf{s}^{(k')}\right\|_2^2\right) \exp\left(-\frac{1}{2v}\left\|\mathbf{x}-\mathbf{x}^{(k')}\right\|_2^2\right)$$

y si, a partir de ello, se aplica la estimación MS, que implica elegir

$$E\{\mathbf{s} \mid \mathbf{x}\} = \int \mathbf{s}\, p(\mathbf{s} \mid \mathbf{x})\, d\mathbf{s} = \int \frac{\mathbf{s}\, p(\mathbf{s},\mathbf{x})}{p(\mathbf{x})}\, d\mathbf{s} = \frac{\int \mathbf{s}\, p(\mathbf{s},\mathbf{x})\, d\mathbf{s}}{\int p(\mathbf{s},\mathbf{x})\, d\mathbf{s}}$$

se tiene

$$\hat{\mathbf{s}}_p(\mathbf{x}) = \frac{\int \mathbf{s}\,\hat{p}(\mathbf{s},\mathbf{x})\,d\mathbf{s}}{\int \hat{p}(\mathbf{s},\mathbf{x})\,d\mathbf{s}} = \frac{\sum_{k'=1}^{K} \mathbf{s}^{(k')} \exp\left(-\frac{1}{2v}\left\|\mathbf{x}-\mathbf{x}^{(k')}\right\|_2^2\right)}{\sum_{k'=1}^{K} \exp\left(-\frac{1}{2v}\left\|\mathbf{x}-\mathbf{x}^{(k')}\right\|_2^2\right)}$$

Análogamente se puede proceder con mezclas.

El inconveniente es el mismo que el considerado en decisión: no hay orientación del planteamiento a optimizar la estimación de **s**.