Métodos no paramétricos de clasificación

Lección 13

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

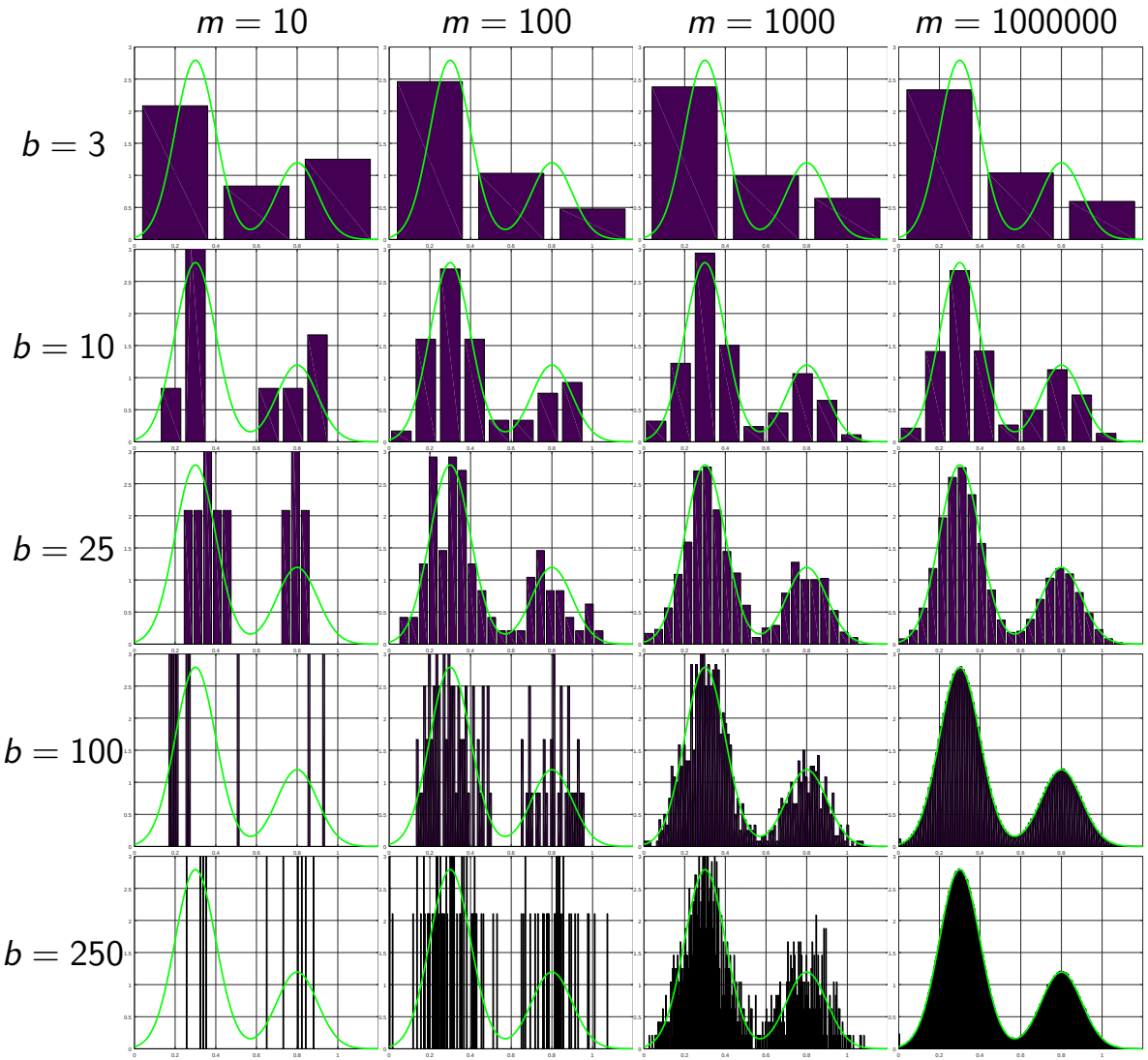
Métodos generativos no paramétricos de clasificación

- En general, los métodos generativos no paramétricos buscan formas de estimar $p(\underline{\mathbf{x}}|y)$ a partir de los datos.
- La clasificación selecciona entonces la mayor probabilidad $p(y|\underline{\mathbf{x}}) = p(\underline{\mathbf{x}}|y)p(y)/p(\underline{\mathbf{x}})$
- Ya revisamos un método generativo, no paramétrico, de clasificación: Modelos de probabilidad de color

Histogramas como estimadores no paramétricos

- En este método, utilizamos histogramas para estimar $p(\underline{\mathbf{x}}|y)$: uno para $y = 0$ (no-piel) y otro para $y = 1$ (piel).
- Podemos llamar al método **semi**paramétrico porque una vez calculado el histograma no necesitamos los datos.
- Supongamos que cada dimensión se divide en b celdas.
- Con n dimensiones, histograma tiene entonces b^n celdas
- Incremento exponencial de celdas con dimensión del espacio de entrada hace difícil "llenar" el histograma para $b \gg 1$
- División discreta de las celdas conduce a discontinuidades problemáticas

Dependencia de celdas y número de datos



Maldición de la dimensión

- Fenómenos en 1D empeoran en n-D
- Caso de probabilidades de color (3D) es manejable, pues $b = 32$ y $n = 3$, por lo que el número total de celdas ($32^3 = 2^{15} = 32768$) es fácilmente manejable con los millones de píxeles disponibles para estimarlo, y solo se requieren dos modelos (p. ej. piel y no-piel).
- Un simple cambio que considerara posición en la estimación y pasara a $n = 5$ produce un histograma de $32^5 = 2^{25} = 33\,554\,432$ celdas, que requerirá muchos más datos para que llegue a representar fielmente la distribución subyacente.
- El número de datos requerido crece exponencialmente con la dimensión del problema: este efecto se conoce como la **maldición de la dimensión** (*curse of dimensionality*)

Distribución de densidad binomial (1)

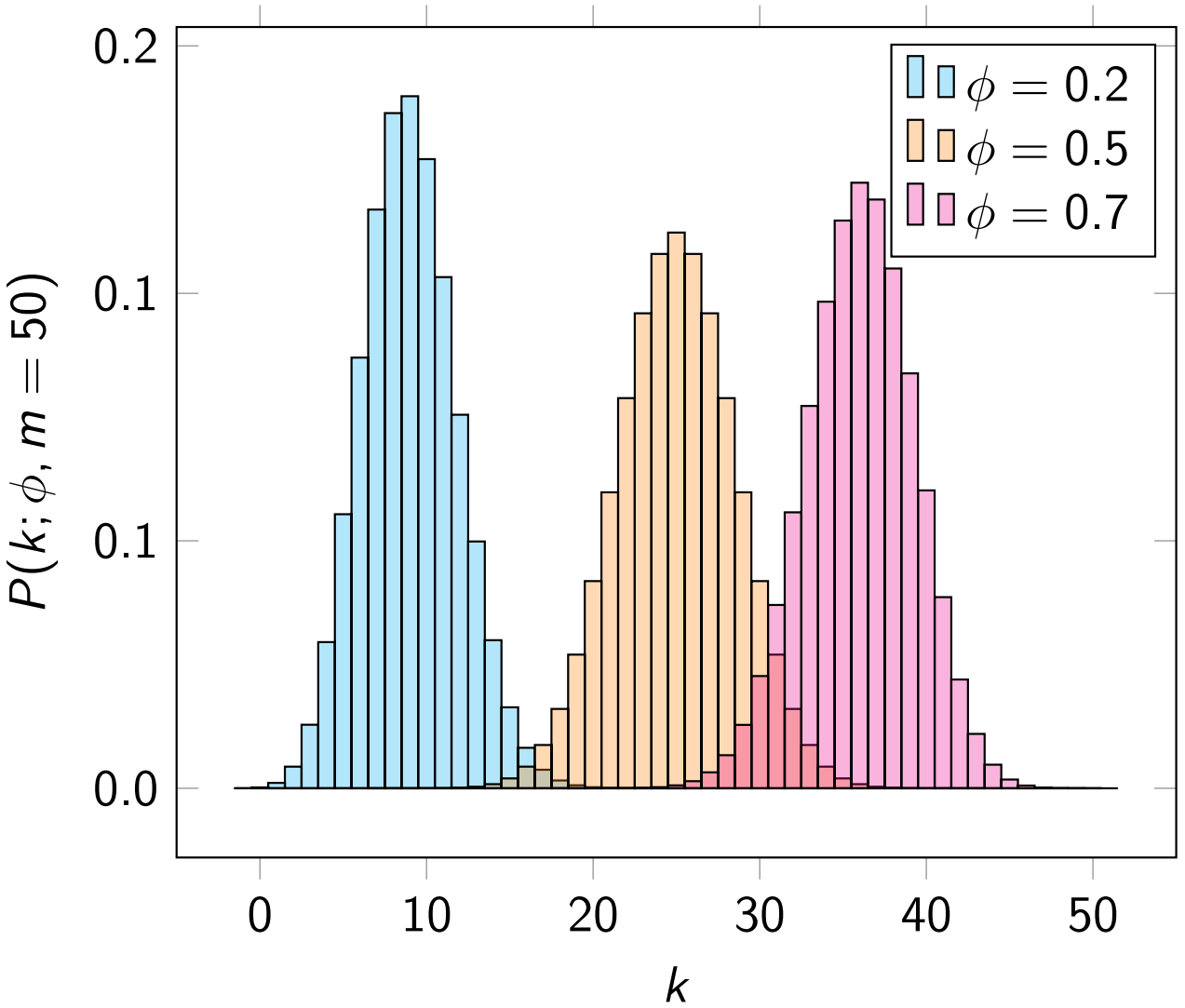
- Necesitaremos la distribución de densidad **binomial**
- Es una generalización de la distribución de Bernoulli
- Es una distribución de probabilidad de masa **discreta** que indica el número k de aciertos en una secuencia de m experimentos independientes, con cada experimento individual siguiendo la misma distribución de Bernoulli con parámetro ϕ
- La distribución de probabilidad de masa binomial es:

$$\mathcal{B}(m,\phi) = P(k; m,\phi) = \binom{m}{k} \phi^k (1-\phi)^{m-k}$$

$$= \frac{m!}{k!(m-k)!} \phi^k (1-\phi)^{m-k}$$

- La distribución de Bernoulli equivale a $m = 1$ y $k = y$.

Distribución de densidad binomial (2)



Distribución de densidad binomial (3)

- Sea $K \sim \mathcal{B}(m,\phi)$ una variable aleatoria.
- La esperanza de K es $\mathrm{E}[K] = m\phi$
- La varianza de K es $m\phi(1-\phi)$

Estimacion de densidad (1)

- Si los vectores $\underline{\mathbf{x}}$ se toman de un proceso con densidad de probabilidad $p(\underline{\mathbf{x}})$, entonces, la probabilidad de que $\underline{\mathbf{x}}$ esté dentro de una región $\mathcal{R}$ es:

$$\phi = \int_{\mathcal{R}} p(\underline{\mathbf{x}}')\, d\underline{\mathbf{x}}'$$

- Si tenemos m puntos tomados del mismo proceso con densidad $p(\underline{\mathbf{x}})$, entonces la probabilidad de que k de ellos estén dentro de la región $\mathcal{R}$ sigue la distribución binomial $K \sim \mathcal{B}(m,\phi)$
- La media de la fracción de puntos que caen en la región es

$$\mathrm{E}\left[\frac{K}{m}\right] = \frac{1}{m}\mathrm{E}[K] = \frac{m\phi}{m} = \phi$$

Estimacion de densidad (2)

- La varianza de esta fracción de puntos es

$$\text{Var}\left[\frac{K}{m}\right] = \text{E}\left[\left(\frac{K}{m} - \phi\right)^2\right] = \frac{\phi(1-\phi)}{m}$$

que se hace cero si $m \to \infty$ (varianza cero $\Rightarrow$ pico en media)

- Como $\text{Var}[K/m] \to 0$ para $m \to \infty$ podemos suponer que

$$\phi \approx \frac{k}{m}$$

- Si $p(\underline{\mathbf{x}})$ es aproximadamente constante en $\mathcal{R}$, entonces

$$\phi = \int_{\mathcal{R}} p(\underline{\mathbf{x}}')\, d\underline{\mathbf{x}}' \approx p(\underline{\mathbf{x}}) \underbrace{\int_{\mathcal{R}} d\underline{\mathbf{x}}'}_{V} = p(\underline{\mathbf{x}})V, \qquad \underline{\mathbf{x}} \in \mathcal{R}$$

Estimacion de densidad (3)

- Combinando los dos valores de ϕ obtenemos

$$\frac{k}{m} \approx \phi \approx p(\underline{\mathbf{x}})V \quad \Rightarrow \quad p(\underline{\mathbf{x}}) \approx \frac{k}{mV}$$

- Exactitud de aproximaciones aumenta mientras mayor sea m
- Tenemos dos estrategias para estimar $p(\underline{\mathbf{x}})$:
  1. V fijo y se determina k (estimación de densidad con kernels)
  2. k fijo y se determina entonces V (k-NN)

Estimación de densidad con kernels (1)

- Tomemos $\mathcal{R}$ como un hipercubo con lados de tamaño h centrados en $\underline{\mathbf{x}} \in \mathbb{R}^n$
- El volumen del hipercubo es entonces $V = h^n$
- Sea $H(\underline{\mathbf{u}})$ una ventana de Parzen:

$$H(\underline{\mathbf{u}}) = \begin{cases} 1 & \|\underline{\mathbf{u}}\|_\infty < 1/2 \\ 0 & \text{en otro caso} \end{cases}$$

(hipercubo de lado 1 centrado en el origen)

- $H\left(\frac{\underline{\mathbf{x}}-\underline{\mathbf{x}}^{(i)}}{h}\right)$ es 1 si $\underline{\mathbf{x}}^{(i)}$ está dentro del hipercubo de lado h centrado en $\underline{\mathbf{x}}$

Estimación de densidad con kernels (2)

- El número total de puntos en el hipercubo es entonces:

$$k = \sum_{i=1}^{m} H\left(\frac{\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(i)}}{h}\right)$$

- Similar a histograma pero celda de lado h se centra en $\underline{\mathbf{x}}$
- Introduciendo esto y $V = h^n$ en $p(\underline{\mathbf{x}}) \approx k/mV$ obtenemos la densidad del modelo:

$$\tilde{p}(\underline{\mathbf{x}}) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{h^n} H\left(\frac{\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(i)}}{h}\right)$$

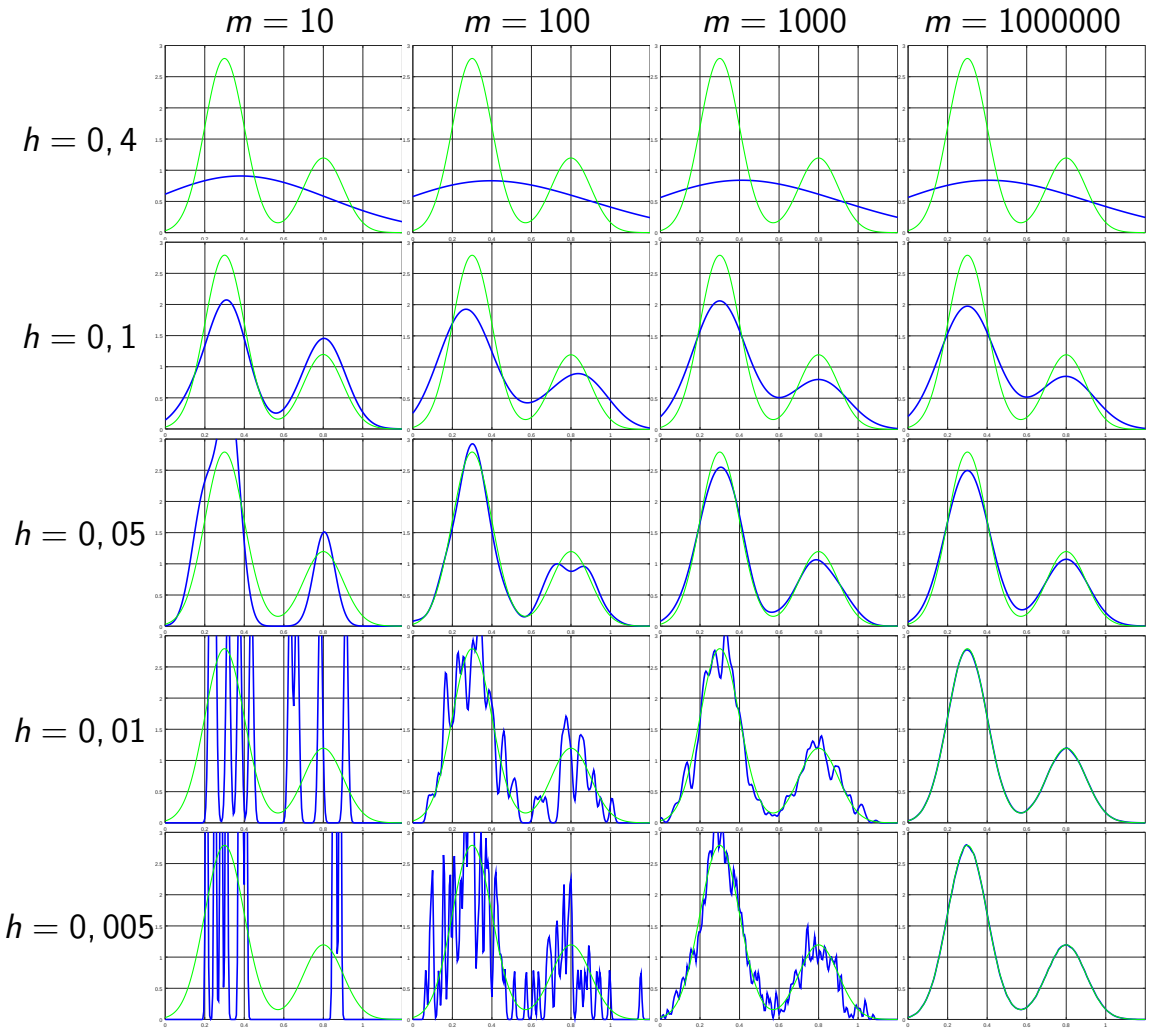
- El término h se denomina **ancho de banda**.
- Las discontinuidades causadas por las cajas H se pueden eliminar usando otros kernels.

Estimación de densidad con kernels (3)

- Basta con satisfacer $H(\underline{\mathbf{u}}) \geq 0$ y $\int H(\underline{\mathbf{u}})\, d\underline{\mathbf{u}} = 1$ para que $\tilde{p}(\underline{\mathbf{x}})$ satisfaga las condiciones para una distribución de densidad probabilística.
- Se usa con frecuencia el kernel gaussiano:

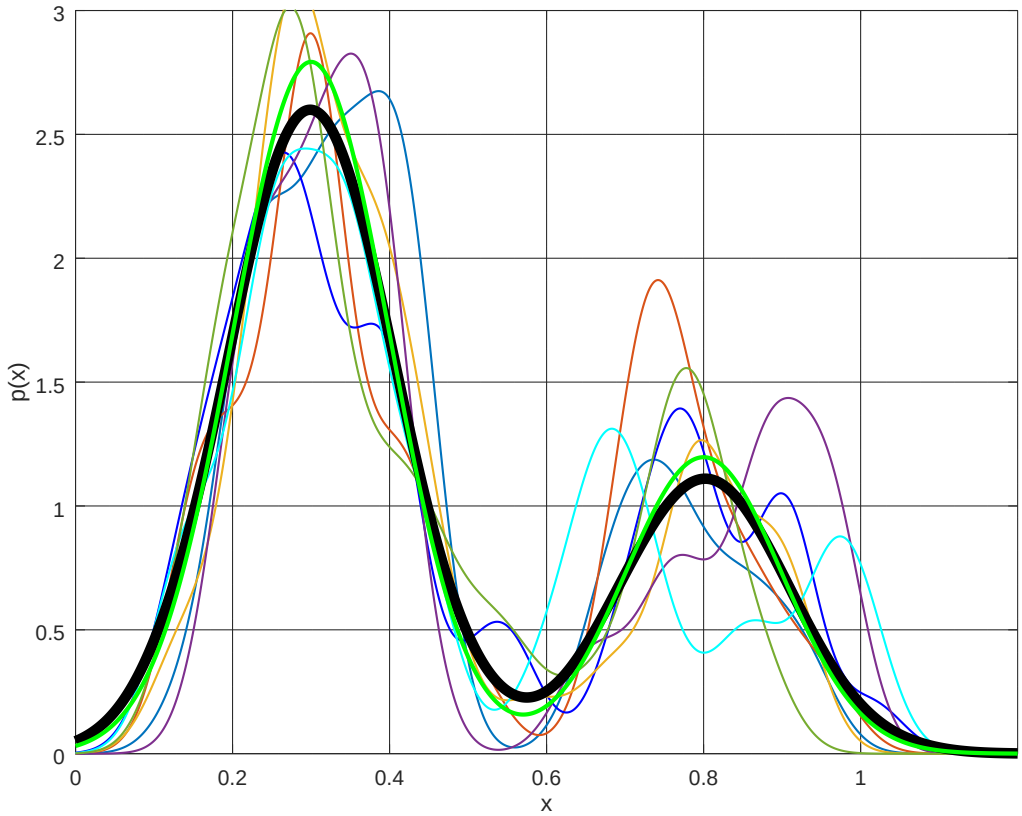
$$\tilde{p}(\underline{\mathbf{x}}) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{(2\pi h^2)^{n/2}} \exp\left(-\frac{\|\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(i)}\|^2}{2h^2} \right)$$

Estimación de densidad con kernels (4)



Esperanza de densidad estimada (1)

- Si calculamos la media de la densidad estimada (media del valor estimado en $\underline{\mathbf{x}}$) sobre distintas selecciones de puntos



Esperanza de densidad estimada (2)

¿A qué converge esa media?

- Aplicando la esperanza a $\tilde{p}(\underline{\mathbf{x}})$:

$$\mathrm{E}[\tilde{p}(\underline{\mathbf{x}})] = \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{1}{h^n}H\left(\frac{\underline{\mathbf{x}}-\underline{\mathbf{x}}'}{h}\right)\right] = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{h^n}\mathrm{E}\left[H\left(\frac{\underline{\mathbf{x}}-\underline{\mathbf{x}}'}{h}\right)\right]$$
$$= \frac{1}{h^n}\mathrm{E}\left[H\left(\frac{\underline{\mathbf{x}}-\underline{\mathbf{x}}'}{h}\right)\right] = \frac{1}{h^n}\int H\left(\frac{\underline{\mathbf{x}}-\underline{\mathbf{x}}'}{h}\right)p(\underline{\mathbf{x}}')\,d\underline{\mathbf{x}}'$$

que ¡es la **convolución** de la densidad $p(\underline{\mathbf{x}})$ con kernel H!

- La densidad es *suavizada* con H
- Si H se hace angosta (tiende a $\delta(\underline{\mathbf{x}})$) entonces $\mathrm{E}[\tilde{p}(\underline{\mathbf{x}})]$ tiende a $p(\underline{\mathbf{x}})$, pero requiere $m \to \infty$, o tendrá mucho ruido
- ¡Se requiere un compromiso para h! (¿recuerdan la regresión ponderada localmente?)

Esperanza de densidad estimada (3)

- El mayor problema de la estimación con kernels es que requieren el almacenamiento de todos los puntos $\underline{\mathbf{x}}^{(i)}$, y el costo de la estimación de densidad depende de ese número de puntos.

Dependencia de h

- Tomar h constante en estimación con kernels conduce a problemas:
  - Si h es muy grande, regiones sobrepobladas de $\underline{\mathbf{x}}$ se sobre suavizan
  - Si h es muy pequeño, regiones poco pobladas de $\underline{\mathbf{x}}$ tendrán estimaciones ruidosas
- Valor óptimo de h depende entonces de la posición de $\underline{\mathbf{x}}$
- La estimación de k vecinos más cercanos pretende solucionar esto

Estimación con kNN (1)

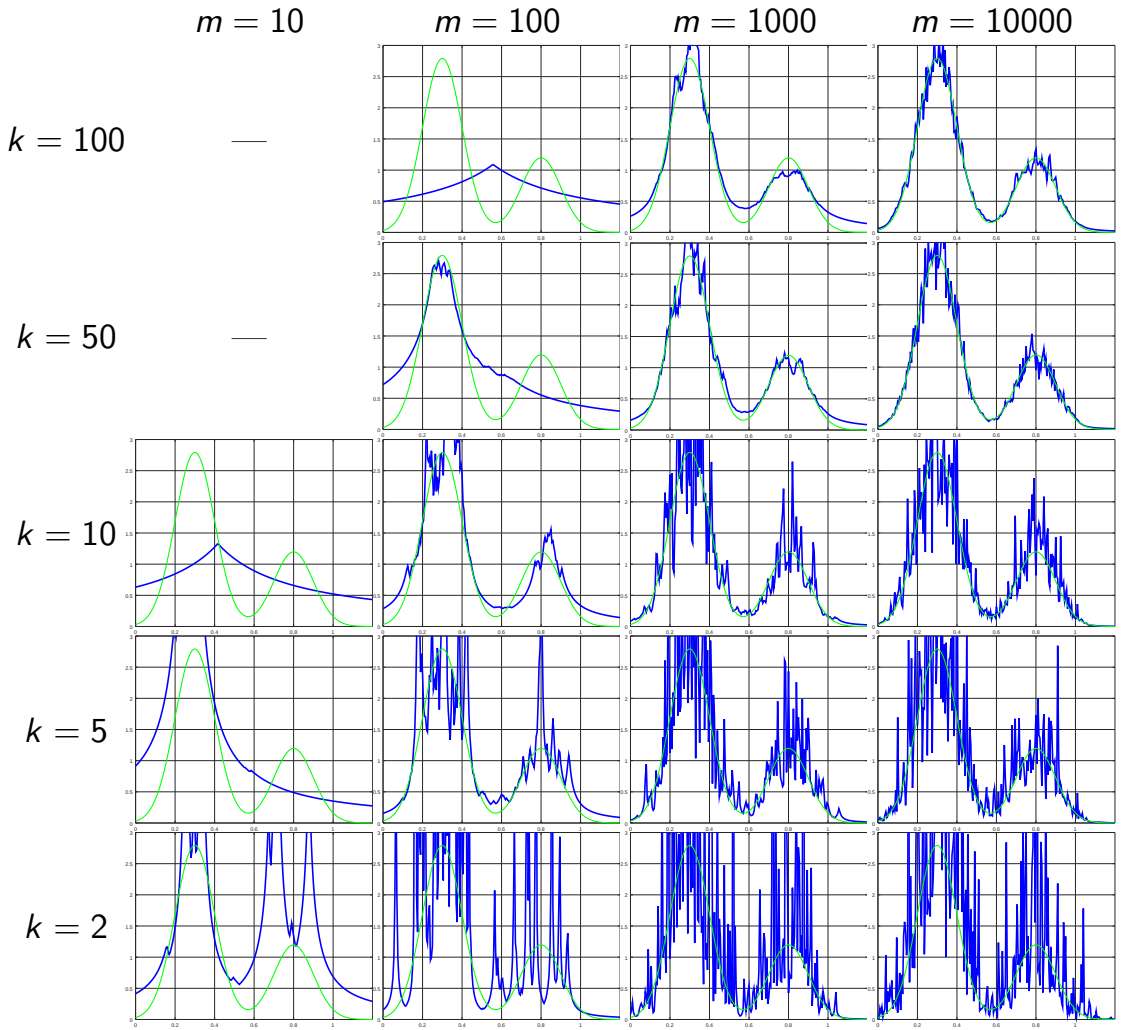
- Retomando

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

mantenemos fijo k e incrementamos el volumen V de una hiperesfera centrada en $\underline{\mathbf{x}}$ hasta que incluya exactamente k puntos.

- Se utilizan estructuras de datos para almacenar puntos, que sean eficientes para búsqueda de vecinos más cercanos (kd-trees, biblioteca FLANN, . . .)

Estimación con kNN (2)



Desventajas de la estimación con kNN

- Desventaja del kNN es que lo estimado no es una densidad probabilística (integral no suma 1)
- Desventaja del kNN y la estimación con kernels es que requerimos mantener todo el conjunto de puntos de entrenamiento.
- Por eso requerimos *kd*-trees u otras estructuras para búsqueda eficiente

Clasificación con k vecinos más cercanos (1)

- Como método generativo, con kNN estimamos $p(\underline{\mathbf{x}}|y = c)$, donde $y = c$ indica que $\underline{\mathbf{x}}$ pertenece a la clase c:

$$p(\underline{\mathbf{x}}|y = c) = \frac{k_c}{m_c V}$$

con

  - m_c el número de puntos en el conjunto de entrenamiento que pertenecen a la clase c,
  - $m = \sum_c m_c$.
  - k es el número total de puntos en el volumen V, y
  - k_c es el número de puntos de la clase c que están dentro de ese volumen V.
  - $k = \sum_c k_c$

Clasificación con k vecinos más cercanos (2)

- Con la probabilidad a priori

$$P(y = c) = \frac{m_c}{m}$$

obtenemos la densidad no condicional

$$p(\underline{\mathbf{x}}) = \sum_c p(\underline{\mathbf{x}}|y = c)P(y = c)$$
$$= \sum_c \frac{k_c}{m_c V} \cdot \frac{m_c}{m} = \sum_c \frac{k_c}{mV} = \frac{k}{mV}$$

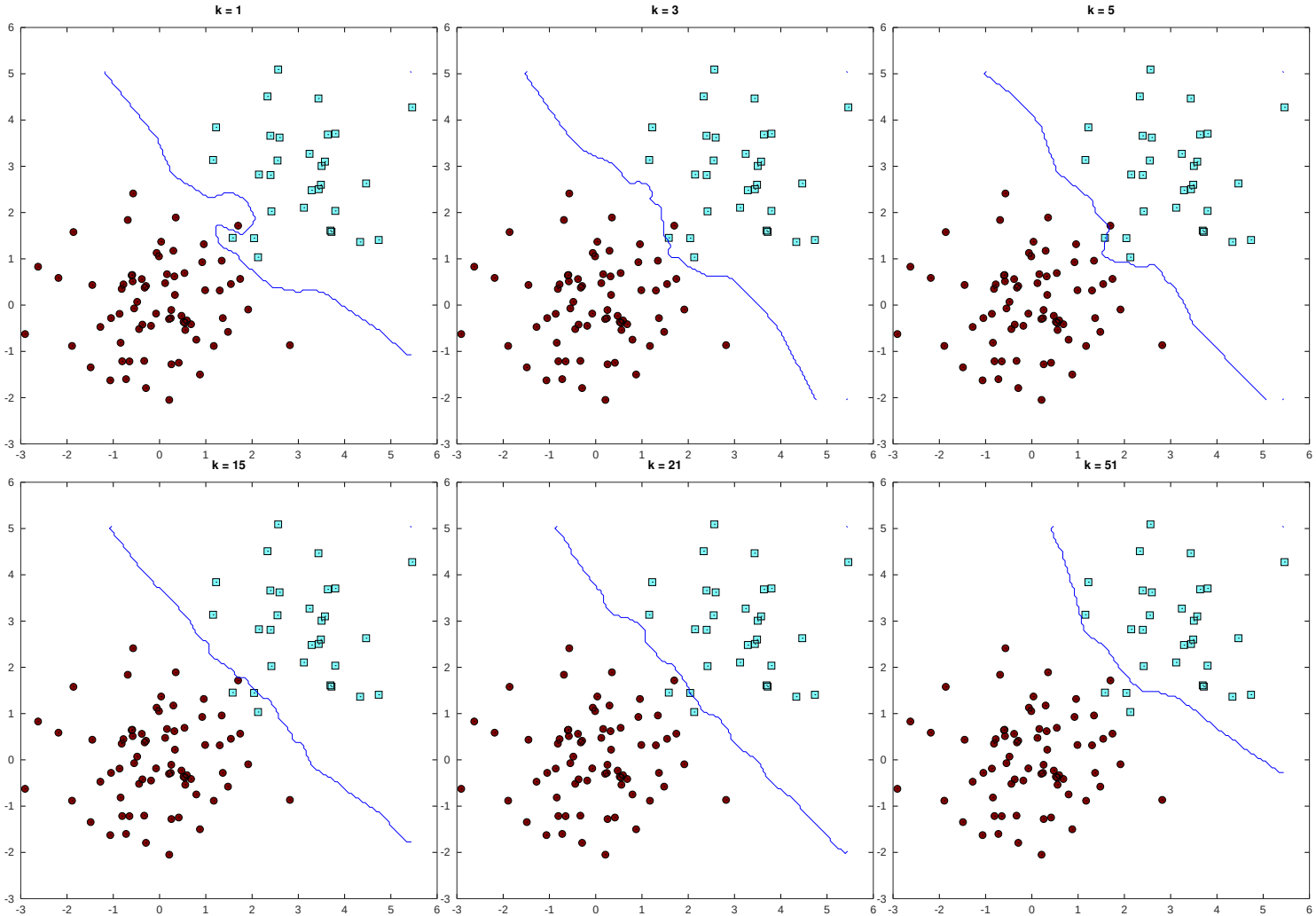
- Finalmente, con la regla de Bayes

$$P(y = c|\underline{\mathbf{x}}) = \frac{p(\underline{\mathbf{x}}|y = c)P(y = c)}{p(\underline{\mathbf{x}})} = \frac{k_c}{k}$$

Clasificación con k vecinos más cercanos (3)

- Para minimizar la probabilidad de error de clasificación, $\underline{\mathbf{x}}$ debe ser asignado a la clase con más representantes en los k vecinos más cercanos.
- Nótese que este clasificador tiene un origen generativo (un modelo para cada clase), pero la regla final se aplica de forma discriminativa

Clasificación con k vecinos más cercanos (4)



Resumen

1. Métodos generativos
   - Histogramas
   - Estimación de densidad
   - k vecinos más cercanos
2. Métodos discriminativos
   - Clasificación con k vecinos más cercanos

Este documento ha sido elaborado con software libre incluyendo LaTeX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux