Distancia de Fermat en Clasificadores de Densidad Nuclear

Lic. Gonzalo Barrera Borla

Buenos Aires, 02/01/23

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Instituto del Cálculo

Tesis presentada para optar al título de Magíster en Estadística Matemática
de la Universidad de Buenos Aires

Director: Dr. Pablo Groisman

Abstract

TODO

Contents

1 Introduccion

1.1 El problema de clasificacion

Consideremos el problema de clasificación:

Definition 1. (Problema de clasificación). Sea $\boldsymbol{x} = (x_i)_{i=1}^N$ una muestra de N observaciones, repartidas en M clases $C_1, \ldots, C_M$ mutuamente excluyentes y conjntamente exhaustivas (es decir, $\forall\ i \in [N] \equiv \{1, \ldots, N\}, x_i \in C_j \iff x_i \notin C_k, k \in [M], k \neq j$). Asumamos además que la muestra está compuesta de observaciones independientes entre sí, y en particular, cada clase tiene su propia ley: si $\|C_j\| = N_j$ y $x_i^{(j)}$representa la i-ésima observación de la clase j, resulta que $X_i^{(j)} \sim \mathcal{L}_j(X)\ \forall\ j \in [M], i \in [N_j]$.

Dada una nueva observación x_0 cuya clase es desconocida,

1. (clasificación dura) ¿a qué clase deberíamos asignarla?
2. (clasificación suave) ¿qué probabilidad tiene de pertenecer a cada clase $C_j, j \in [M]$?

Todo método o algoritmo que pretenda responder el problema de clasificación, prescribe un modo u otro de combinar toda la información muestral disponible, ponderando las N observaciones de manera relativa a su cercanía o similitud con x_0. Por caso, $k-$vecinos más cercanos ($k-$NN) asignará la nueva observación x_0 a la clase modal entre las k observaciones de entrenamiento más cercanas (es decir, que minimizan la distancia euclídea $\|x_0 - \cdot\|$). $k-$NN no hace ninguna mención explícita de las leyes de clase $\mathcal{L}_j$, lo cual lo mantiene sencillo a costa de ignorar la estructura del problema.

1.2 Estimación de densidad

Una familia bastante genérica de métodos para resolver el problema de calsificación, consisten aproximadamente de los siguientes pasos:

1. Hacer algunos supuestos sobre la forma de las leyes $\mathcal{L}_j$
2. Hallar estimadores $\hat{\mathcal{L}}_j$ de cada ley $\mathcal{L}_j$ usando las muestras de cada clase, $\boldsymbol{x}^{(j)} = \left(x_i^{(j)}\right)_{i=1}^{N_j}$ y algún procedimiento estándar (e.g.: máxima verosimilitud)
3. Definir una regla de decisión $\mathcal{R}\left(\cdot|\hat{\mathcal{L}}_j, j \in [M]\right) : \mathbb{R}^{d_x} \to [M]$ que dados los estimadores de (2), asigne la observación x_0 a la clase $\mathcal{R}(x_0)$.

Esta familia de clasificadores, se distinguen por una explícita *estimación de densidades* que más tarde se utilizarán para la tarea de clasificación en sí. Por

ejemplo, al considerar el problema de clasificación binaria, el análisis de discriminante lineal (LDA) de Fisher[1] queda encuadrado en esta familia de la siguiene manera:

En (1), asumimos que las las leyes $\mathcal{L}_j$

(a) son todas distribuciones normales con media μ_j y

(b) homocedásticas: $\Sigma_j = \Sigma \; \forall \; j \in [M]$).

En (2), estimamos $\hat{\mu_j}, \hat{\Sigma}$ por máxima verosimilitud,
$\hat{\mu_j} = N_j^{-1} \sum_{i=1}^{N_j} x_i^{(j)}$
$\hat{\Sigma} = N^{-1} \sum_{j=1}^{M} \sum_{i=1}^{N_j} (x_i^{(j)} - \hat{\mu_j})(x_i^{(j)} - \hat{\mu_j})$.
Y la regla de (3) es la indicadora $1\,(\cdot)$ del discriminante lineal

$$\begin{aligned} \mathcal{R}\,(x) &= 1\,(w \cdot x > c) \\ w &= \Sigma^{-1}(\mu_1 - \mu_0) \\ c &= w \cdot \frac{1}{2}(\mu_1 + \mu_0) \end{aligned}$$

con los parámetros μ_j, Σ reemplazados por las estimaciones de (2).

Inevitablemente, existe un *trade-off* entre lo restrictivo de los supuestos de (1), y la generalidad del clasificador resultante. En el caso de LDA, los supuestos (leyes normales y homocedasticidad) son inverosímiles en casi cualquier escenario real, pero el clasificador resultante es muy sencillo de computar. En general, este será el caso para todos los métodos *paramétricos* de estimación de densidad, en que de todas las posibles funciones de densidad, quedan acotadas a aquellas que se pueden expresar de forma cerrada con una expresión predefinida (en este caso, la densidad normal), y Q parámetros (aquí, μ y Σ).

Alternativamente, existen métodos en que los supuestos de (1) se obvian del todo, o al menos son lo suficientemente generales como para representar todas salvo las más patológicas leyes (e.g.: asumir que la media y dispersión son finitas). A estos se los conoce, naturalmente, como métodos *no paramétricos* de estimación de densidad.

Estimación de densidad por núcleos

La estimación de densidad por núcleos (o KDE, por sus siglas en inglés), es uno de los métodos mejor estudiados dentro del amplio universo no-paramétrico[2]. Introducidos hacia 1960 (Rosenblatt 1958, Parzen 1962) para variables aleatorias unidimensionales, han sido ampliamente desarrollados y adaptados a espacios mucho más generales. El objetivo es encontrar un estimador *suave* de la densidad poblacional f de una v.a. X a partir de una muestra discreta, usando una función no-negativa K llamada *núcleo* ("kernel") y un parámetro de suavización h, el *ancho de banda* ("bandwith").

[1] https://en.wikipedia.org/wiki/Linear_discriminant_analysis
[2] Algo sobre NNs, otros metodos nopa

Definition 2. (función núcleo) Una función K es un *núcleo* ("kernel"), si

- toma únicamente valores reales no-negativos: $K(x) \geq 0 \forall x$,
- está normalizada: $\int_{-\infty}^{+\infty} K(u)\, du = 1$ y
- es simétrica: $K(u) = K(-u)\, \forall u$

Remark 3. Si K es un núcleo, entonces $K_\lambda(u) = \lambda K(\lambda u)$ también lo es, lo cual permite construir un núcleo adecuadamente escalado a los datos.

Definition 4. (KDE univariado) Sea $(x_1, \ldots, x_N)$ una muestra de elementos i.i.d. tomada de cierta distribución univariada con densidad desconocida f, cuya forma deseamos conocer. Su estimador de densidad por núcleos (su "KDE") es

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big)$$

Dejando por un momento de lado qué par (K, h) usar, podemos derivar un clasificador "duro" de manera bastante directa para la versión univariada del problema 1:

Definition 5. (clasificador KDE univariado). Sea $C : \mathbb{R}^{d_x} \to [M]$ la "función de clase", tal que $\forall x \in \mathbb{R}^{d_x},\ C(x) = j \iff x \in C_j$. Sean además $\hat{f}_h^{(1)}, \ldots, \hat{f}_h^{(M)}$ los estimadores de densidad obtenidos según 4. El "clasificador por estimación de densidad nuclear" correspondiente será:

$$\hat{C}(x) = \arg\max_{\mathrm{j}\in[\mathrm{M}]} \hat{f}_h^{(j)}(x)$$

asignando cada observación a la clase en la que maximiza la densidad estimada.

Cuando las clases de las cuales se compone la población se encuentran muy "separadas" entre sí (es decir, $\exists k \in [M] : f_h^{(k)}(x_0) \gg 0\ ,\ f_h^{(j)} \simeq 0\ \forall\ j \in [M]/k$), la clasificación "dura" de 5 será suficiente. Ahora bien, ¿cómo hacemos para cuantificar la incertidumbre asociada a la clasificación, cuando existe más de una clase con densidad estimada no despreciable? Como las $\hat{f}_h^{(j)}$ estimadas identifican distribuciones, es razonable decir que $p(C(x) = j) \propto f_h^{(j)}(x)$. Usando la regla de Bayes y un *a priori* sobre las probabilidades de clase basado en las proporciones muestrales $\hat{p}(C_j) = N_j/N$, podemos conseguir una regla *suave* de clasificación:

Definition 6. (clasificador KDE univariado suave) Sea el problema 1 y los estimadores de densidad de 4. Por la regla de bayes,

$$p(C(x) = j) = \frac{f^{(j)}(x) \cdot p(C_j)}{p(x)}$$

Reemplazando el a priori $p(C_j)$ por su estimación muestral, las densidades $f^{(j)}$ por sus estimadores y usando la ley de la probabilidad total para expandir $p(x)$, obtenemos:

$$\hat{p}\left(C\left(x\right)=j\right)=\frac{\hat{f}_h^{(j)}\left(x\right)\cdot N_j}{\sum_{i\in[M]}\hat{f}_h^{(i)}\left(x\right)\cdot N_i}$$

1.3 La noción de distancia en KDE

Como mencionamos en un principio, toda propuesta de solución al problema de clasificación 1 para una nueva observación x_0, lo hará ponderando su cercanía a los elementos muestrales de cada clase. En la estimación de densidad nuclear univariada, el peso de cada elemento muestral está determinado por $K_h\left(x_0-x_i\right)$, y como K_h es simétrica respecto del 0, $K_h\left(x_0-x_i\right)=K_h\left(\left|x_0-x_i\right|\right)=K_h\left(\left\|x_0-x_i\right\|\right)$. Es decir, que el operador núcleo pondera directamente la distancia euclídea entre la nueva observación y cada elemento muestral.

Esto se vuelve más evidente cuando hemos de expresar un KDE para elementos aleatorios multivariados.

Definition 7. (KDE multivariado, Wand & Jones 1993) Sea $(x_1,\ldots,x_N)$ una muestra de elementos i.i.d. tomada de cierta distribución $d-$dimensional con densidad desconocida f, cuya forma deseamos conocer. Su estimador de densidad por núcleos (su "KDE") es

$$\widehat{f_h}(x)=\frac{1}{n}\sum_{i=1}^{n}K_{\mathbf{H}}(x-x_i)$$

donde K misma es una función de densidad $d-$variada, $\mathbf{H}$ es una matrix $d\times d$ positiva simétrica definida, y para todo $x\in\mathbb{R}^d$, $K_{\mathbf{H}}\left(x\right)=\det\left(\mathbf{H}\right)^{-1/2}K\left(\mathbf{H}^{-1/2}x\right)$.

Tomemos por caso el núcleo gaussiano, $K\left(x\right)=\left(2\pi\right)^{-d/2}\exp\left(-\frac{1}{2}x^Tx\right)$. Luego,

$$\begin{aligned}K_{\mathbf{H}}(x-x_i)&=K_{\mathbf{H}}\left(x-x_i\right)\\&=\det\left(\mathbf{H}\right)^{-1/2}K\left(\mathbf{H}^{-1/2}\left(x-x_i\right)\right)\\&=\left(2\pi\right)^{-d/2}\det\left(\mathbf{H}\right)^{-1/2}\exp\left(-\frac{1}{2}\left\|\mathbf{H}^{-1/2}\left(x-x_i\right)\right\|^2\right)\end{aligned}$$

es decir, que el peso de la i-ésima observación muestral en la estimación de densidad del punto x, depende directamente de su distancia de Mahalanobis a una distrubución normal centrada en x_i con matrix de covarianza $\mathbf{H}$. Cuando $\mathbf{H}=\mathbf{I_d}$, la distancia de Mahalanobis es igual a la euclídea.

- Distancia dist en kde
  - Por omisión: dist euclidea: OK en low dim
    - (¿Es lo mismo que la geodésica en R^d_x? Creo que sí)

  - Euclidea inutil en alta dim: Malediction de la dimensionalidad d_x (Bengio?)
    * Hipótesis de la variedad mu , d_mu << d_x (Bengio)
- KDE en variedad (pelletier)
- Distancia en la variedad (de Riemann)
  - Si se sabe, geodésica H&R (H & Rodríguez?)
  - Si no se sabe, se estima a partir de los datos
    * Aprendizaje de distancia, de representaciones (Bengio)
- Isomap (shortest-path en grafo completo)
- Distancia de Fermat
- Estimador de fermat en grafo completo
  - Isomap como caso particular (p=1) de estimador de dist Fermat (fde)
- Pendientes:
  - isomap
  - soft clf chen

2 Propuesta

- estimador distancia de fermat del grafo de los datos según Groisman et al,
- usar la la distancia estimada como plug-in en kde pelletier
- Check clf acc & perf

3 Análisis experimental

4 Cuentita

5 Conclusiones