Inferencia Estadística y Reconocimiento de Patrones

UNaB, Lic. Cs. de Datos, 2021 cuat. 2

Sebastián Pedersen (sebastian.pedersen (at) unab (punto) edu (punto) ar)

Clasificadores bayesianos, Discriminante lineal/cuadrático

(Clasificación supervisada)

Recordatorio del problema de clasificación supervisada

1. Tengo datos de entrenamiento: (X1,X2, …, Xp) → Y
    a. (X1, …, Xp) son las variables, características o predictores (lo que mido)
    b. Y es la clasificación, target, label o etiqueta.
    c. Tengo muchos de estos datos (o la cantidad que pueda).
2. Con los datos de entrenamiento construyo mi modelo predictor/clasificador.
3. Con el modelo predictor clasifico nuevos datos (X1,...,Xp)
    a. Ej.: (peso, altura,presión, cant. infartos) → SÍ/No riesgo cardíaco.
4. Queda en el tintero: ¿cómo evalúo a mi modelo predictor?
    a. ¿Cómo mido qué tan bien está clasificando?
    b. ¿De dónde saco datos nuevos para probarlo?
    c. Si el dato es nuevo y por lo tanto no clasificado, ¿cómo sé si mi modelo anda bien o mal?

Punto de vista Probabilidad/Estadística

- Intento predecir Y dados $(X_1,...,X_p)$
  - (peso, altura, presión, cant. infartos) → SÍ/NO con prob. p.
  - (80, 170, 150, 2) → SÍ con prob. 0.6
  - (60, 155, 140, 1) → NO con prob. 0.3
  - Etc.
- Es decir intento estimar $P(Y=k \mid X_1,...X_p=x_1,...,x_p)$ para $k=0$ o 1 (o la cantidad de clases que haya).
- Una vez que tengo estimada esa probabilidad, pongo una regla para clasificar a los datos nuevos:

  Por ejemplo si la estimación de $P(Y=k \mid X_1,...X_p=x_1,...,x_p) > 0.4 \rightarrow y=k$

  O elijo el máximo sobre k de $P(Y=k \mid X_1,...X_p=x_1,...,x_p)$, y clasifico según ese máximo. Por ejemplo para un problema con 3 clases:

  - Estim. de $P(Y=0 \mid X_1,...,X_p=x_1,...,x_p) = 0.3$
  - Estim. de $P(Y=1 \mid X_1,...,X_p=x_1,...,x_p) = 0.2$
  - Estim. de $P(Y=2 \mid X_1,...,X_p=x_1,...,x_p) = 0.5$

  Clasifico como $y=2$.

Cómo funciona un clasificador bayesiano

Primero transforma la probabilidad que quiere estimar de la siguiente forma:

$$P(Y = k|X_1, \ldots, X_p = x_1, \ldots, x_p) = \frac{P(X_1, \ldots, X_p = x_1, \ldots, x_p|Y = k)P(Y = k)}{P(X_1, \ldots, X_p = x_1, \ldots, x_p)}$$

La anterior igualdad viene de aplicar esto:

$$P(Y|X) = \frac{P(Y \cap X)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \quad \rightarrow \quad P(X \cap Y) = P(X|Y)P(Y)$$

Cómo funciona un clasificador bayesiano

Ahora usando esa transformación:

$$P(Y = k|X_1, \ldots, X_p = x_1, \ldots, x_p) = \frac{P(X_1, \ldots, X_p = x_1, \ldots, x_p|Y = k)P(Y = k)}{P(X_1, \ldots, X_p = x_1, \ldots, x_p)}$$

Estima los valores del miembro derecho (para estimar el miembro izquierdo, que es el objetivo final). Es decir hay que estimar tres cosas:

- P(Y=k)
- P(X1,...,Xp=x1,...,xp)
- P(X1,...,Xp=x1,...,xp | Y=k)

$$P(Y=k|X_1,\ldots,X_p=x_1,\ldots,x_p)=\frac{P(X_1,\ldots,X_p=x_1,\ldots,x_p|Y=k)P(Y=k)}{P(X_1,\ldots,X_p=x_1,\ldots,x_p)}$$

Cómo funciona un clasificador bayesiano

1. P(Y=k) la estima por proporciones de los datos de entrenamiento: exactamente igual que para Bayes ingenuo.

2. P(X1,...,Xp=x1,...,xp) se estima exactamente igual que para Bayes ingenuo, aunque para Discriminante Lineal/Cuadrático queda incluso más claro que no es necesaria esta estimación.

3. Para estimar P(X1,...,Xp=x1,...,xp | Y=k) hay varias opciones:
    a. Bayes ingenuo (naive Bayes)
    b. Discriminante Lineal o Cuadrático
    c. Otras.

$$P(Y=k|X_1,\ldots,X_p=x_1,\ldots,x_p)=\frac{P(X_1,\ldots,X_p=x_1,\ldots,x_p|Y=k)P(Y=k)}{P(X_1,\ldots,X_p=x_1,\ldots,x_p)}$$

Discriminante lineal / cuadrático

Para cada clase k, asume que el condicional resp. a la clase tiene distribución gaussiana. Pongo X = X1,...,Xp para simplificar la notación:

$$P(X=x|Y=k)\sim N(\mu_k,\Sigma_k) \qquad X|Y=k\sim N(\mu_k,\Sigma_k)$$

Por ejemplo para el caso de dos clases habría dos:

$$P(X=x|Y=0)\sim N(\mu_0,\Sigma_0) \qquad X|Y=0\sim N(\mu_0,\Sigma_0)$$

$$P(X=x|Y=1)\sim N(\mu_1,\Sigma_1) \qquad X|Y=1\sim N(\mu_1,\Sigma_1)$$

$$P(Y=k|X_1,\ldots,X_p=x_1,\ldots,x_p)=\frac{P(X_1,\ldots,X_p=x_1,\ldots,x_p|Y=k)P(Y=k)}{P(X_1,\ldots,X_p=x_1,\ldots,x_p)}$$

Discriminante lineal / cuadrático

Es decir para cada clase k, asume que puede calcular la prob. del condicional resp. a la clase por una gaussiana:

$$P(X=x|Y=k)=\frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}\mathrm{e}^{-\frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)}$$

Para el caso de dos clases sería:

$$P(X=x|Y=0)=\frac{1}{(2\pi)^{p/2}|\Sigma_0|^{1/2}}\mathrm{e}^{-\frac{1}{2}(x-\mu_0)^t\Sigma_0^{-1}(x-\mu_0)}$$

$$P(X=x|Y=1)=\frac{1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}}\mathrm{e}^{-\frac{1}{2}(x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)}$$

$$P(Y=k|X_1,\ldots,X_p=x_1,\ldots,x_p)=\frac{P(X_1,\ldots,X_p=x_1,\ldots,x_p|Y=k)P(Y=k)}{P(X_1,\ldots,X_p=x_1,\ldots,x_p)}$$

Discriminante lineal / cuadrático

Y luego para cada clase k, estima μ y Σ a partir de los datos de entrenamiento (análogo a Bayes ingenuo), y aproxima la probabilidad anterior:

$$P(X=x|Y=k)=\frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}\mathrm{e}^{-\frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)}$$

Por lo tanto para estimar $P(Y=k|X=x)=\frac{P(X=x|Y=k)P(Y=k)}{P(X=x)}$

- P(X=x | Y=k) la estima por la gaussiana, y a P(Y=k) y P(X=x) igual que en Bayes ingenuo.

$$P(Y=k|X_1,\ldots,X_p=x_1,\ldots,x_p)=\frac{P(X_1,\ldots,X_p=x_1,\ldots,x_p|Y=k)P(Y=k)}{P(X_1,\ldots,X_p=x_1,\ldots,x_p)}$$

Discriminante lineal / cuadrático

Para cada clase k, estima μ y Σ a partir de los datos de entrenamiento (análogo a Bayes ingenuo), y aproxima la probabilidad anterior:

$$P(X=x|Y=k)=\frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}\mathrm{e}^{-\frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)}$$

- Discriminante lineal: asume Σ igual para todas las clases (k=0, 1, 2, …)
- Discriminante cuadrático: no asume que las Σ sean iguales.

Discriminante lineal / cuadrático

$$P(Y = k|X = x) = \underbrace{\frac{P(X = x|Y = k)P(Y = k)}{P(X = x)}}$$

Para estimar esta ← Estima estas tres cosas.

Ahora dado un dato (X1,...,Xp) puedo estimar la probabilidad de pertenecer a cada clase:

- Para k=0 estim. una probabilidad
- Para k=1 estim. una probabilidad, etc.

Y por ejemplo clasificar según la probabilidad más alta, o algún punto de corte adecuado.

Referencias

- Hastie, Tibshirani, Introduction to Statistical Learning, secciones 4.4.1 a 4.4.3
- Bishop, Pattern Recognition and Machine Learning, sección 4.1
- Chan, Análisis Inteligente de Datos, secciones 9.1 y 9.2