Inferencia Estadística y Reconocimiento de Patrones

UNaB, Lic. Cs. de Datos, 2021 cuat. 2
Sebastián Pedersen (sebastian.pedersen (at) unab (punto) edu (punto) ar)

Regresión Logística

(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?
        i. Métricas de evaluación (por ej. accuracy, recall, precision).
    b. ¿De dónde saco datos nuevos para probarlo? Si el dato es nuevo y por lo tanto no clasificado, ¿cómo sé si mi modelo anda bien o mal?
        i. Train/Test split.

Punto de vista Probabilidad/Estadística

- Intento predecir Y dados (X1,...,Xp)
  - (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 | X1,...Xp=x1,...,xp) 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 | X1,...Xp=x1,...,xp) > 0.4 → y=k

  O por ej. elijo el máximo sobre k de P(Y=k | X1,...Xp=x1,...,xp), y clasifico según ese máximo. Para un problema con 3 clases sería:

  - Estim. de P(Y=0 | X1,...,Xp=x1,...,xp) = 0.3
  - Estim. de P(Y=1 | X1,...,Xp=x1,...,xp) = 0.2
  - Estim. de P(Y=2 | X1,...,Xp=x1,...,xp) = 0.5

  } Clasifico como y=2.

Cómo funciona Regresión Logística

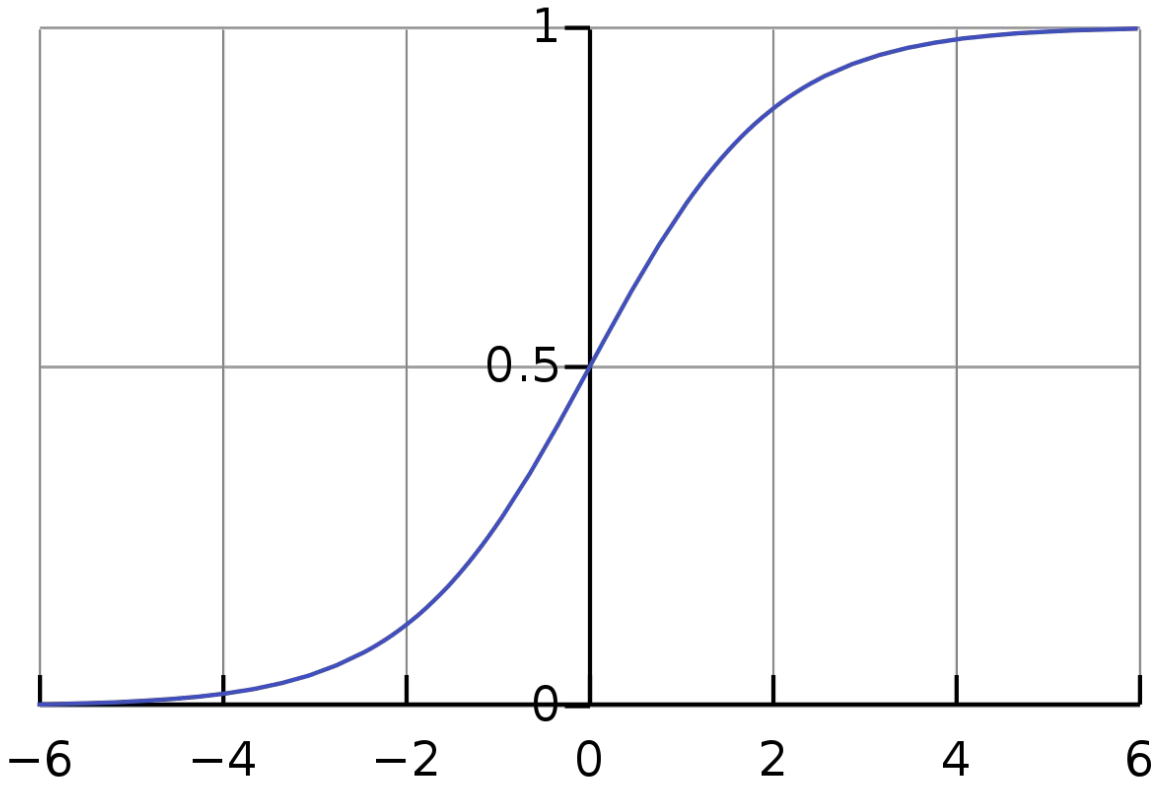
Supone que la probabilidad que quiere estimar P(Y=k | X1,...Xp=x1,...,xp), para k=0 o 1, tiene forma funcional logística:

$$P(Y = 1|X_1, \ldots, X_p = x_1, \ldots, x_p) = \frac{e^{\beta_0 + \beta_1 . (x_1, \ldots, x_p)}}{1 + e^{\beta_0 + \beta_1 . (x_1, \ldots, x_p)}} \qquad \beta_0 \in \mathbb{R},\ \beta_1 \in \mathbb{R}^p$$

- Esto para el caso binario: solamente dos clases. Para el caso con más clases se extiende de forma análoga.
- Observar que para el caso binario, solamente hace falta estimar el caso k=1, pues el otro se calcula por el complemento.

Regresión Logística: ¿por qué una función logística?

$$f(t) = \frac{e^t}{1 + e^t}$$



- Toma valores entre 0 y 1. Por lo tanto se adecúa a estimar una probabilidad.
- Es suave (derivable). Por lo tanto facilita la aplicación de métodos numéricos.
- Entre otras cosas.

Cómo funciona Regresión Logística

Quiere estimar P(Y=1 | X=x), mediante una función logística:

$$P(Y=1|X=x)=\frac{e^{\beta_0+\beta_{11}x_1+\cdots+\beta_{1p}x_p}}{1+e^{\beta_0+\beta_{11}x_1+\cdots+\beta_{1p}x_p}} \qquad \beta_0\in\mathbb{R},\ \beta_1=(\beta_{11},\ldots,\beta_{1p})\in\mathbb{R}^p$$

- (Abrevié X1,...,Xp=x1,...,xp como X=x).
- Los coeficientes se estiman por Máxima Verosimilitud, a partir de los datos de entrenamiento. (Por ahora queda en el tintero).

Regresión Logística: interpretación de los coeficientes

Una vez estimados los coeficientes, tenemos estimada la probabilidad deseada:

$$P(Y=1|X=x)=\frac{e^{\beta_0+\beta_{11}x_1+\cdots+\beta_{1p}x_p}}{1+e^{\beta_0+\beta_{11}x_1+\cdots+\beta_{1p}x_p}} \qquad \beta_0\in\mathbb{R},\ \beta_1=(\beta_{11},\ldots,\beta_{1p})\in\mathbb{R}^p$$

- Cada coeficiente de $\beta 1$ se puede interpretar como un indicador de la importancia de esa variable en la probabilidad. Esto se debe fundamentalmente a la exponencial.

1. X1=peso, X2=altura, X3=presión arterial, Y=1=tiene riesgo cardíaco

$\beta 11$ = 0.02, $\beta 12$ = 0.01, $\beta 13$ = 0.9 --> entonces X3 tiene más influencia en tener riesgo cardíaco.

2. X1=peso, X2=altura, X3=presión arterial, Y=1=tiene riesgo cardíaco

$\beta 11$ = 0.2, $\beta 12$ = -0.85, $\beta 13$ = 0.9 --> entonces X2 tiene más influencia en no tener riesgo cardíaco (y X3 igual que en el ej. anterior).

Regresión Logística:

$$P(Y = 1|X_1, \ldots, X_p = x_1, \ldots, x_p)$$ ← Directamente estima esta probabilidad.

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

- Para k=1 estim. una probabilidad
- Para k=0 estim. una probabilidad, mediante el complemento.

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

Además puedo interpretar a los coeficientes de la logística como el peso relativo que tiene cada variable a la hora de clasificar el dato (X1,..,Xp)

Referencias

- Hastie, Tibshirani, Introduction to Statistical Learning, sección 4.3
- Bishop, Pattern Recognition and Machine Learning, sección 4.3.2
- Chan, Análisis Inteligente de Datos, sección 9.5