Modelos de clasificación lineales

Martín Onetto

I. MODELOS DE CLASIFICACIÓN

El objetivo de los modelos de clasificación es construir una función $G(X)$ que tome valores en un espacio discreto $\mathcal{G}$. La función G construye fronteras en el espacio de partida que determinan a qué categoría corresponde cada punto del dominio. Estas fronteras pueden ser suaves o no. Para un gran conjunto de modelos, estas "decision boundaries"son lineales y se conocen como métodos lineales de clasificación.

Nos vamos a concentrar además en modelos que emplean una "discriminant function" $\delta_k(x)$ para cada clase k. Luego, a cada x se le asignará la clase k donde tome el valor más grande. Un ejemplo de estos modelos son aquellos que utilizan como función discriminante las probabilidades posteriores $P(G = k|X = x)$, y en caso de que sea una función lineal en x, también sus fronteras de decisión serán lineales.

Estrictamente hablando, para que las fronteras de decisión sean lineales, basta con que haya una transformación monótona de δ_k que resulte lineal, para que las fronteras sean lineales.

Ejemplo

Imaginemos que solo tenemos dos clases posibles. Un modelo común para las probabilidades posteriores es:

$$P(G = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}$$

$$P(G = 0|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}$$

(¿Te suena de algo, tal vez Fermi?)

Una transformación que lleva estas probabilidades a una función lineal del dominio es el "logit"($\log[p/(1-p)]$):

Si $p = P(G = 1|X = x)$, como solo hay dos clases, entonces $1 - p = P(G = 2|X = x)$. Al aplicar el "logit", obtenemos un "bayes factor"de ambas opciones:

$$\log \frac{P(G = 1|X = x)}{P(G = 2|X = x)} = \beta_0 + \beta^T x$$

la cual es lineal. En la zona donde los "log odds"son cero, es donde coinciden las probabilidades y donde hay una frontera definida por un hiperplano lineal.

CLASIFICADOR DE REGRESIÓN LINEAL

Comenzamos definiendo las categorías target como variables indicadoras, por lo que si el espacio de llegada $\mathcal{G}$ tiene K clases, habrá K variables respuesta Y_k, $k = 1, \ldots, K$, que tendrán $Y_k = 1$ si $G_k = 1$ y en otro caso serán 0. Si estructuramos las Y_k en un vector, tenemos $Y = (Y_1, \ldots, Y_k)$, y dadas N observaciones de entrenamiento, obtenemos una matriz $\mathbf{Y}$ de $N \times K$ donde en cada fila hay un solo 1 y el resto son 0's.

Si procedemos con un ajuste lineal como hicimos en el caso de la regresión, sabemos que la solución para β es $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Aquí, X representa los datos de dimensión p, donde además agregamos un 1 para tomar como parámetro la ordenada al origen. Por lo tanto, cuando tenemos una nueva observación x, la clasificamos como:

$$\mathbf{f}(x)^T = (1, x^T)\hat{\mathbf{B}} \tag{1}$$

Esto resulta en un vector de tamaño K.

Identificamos la componente k con mayor valor y clasificamos acorde:

$$G(x) = \arg \max_{k \in \mathcal{G}} f_k(x) \tag{2}$$

¿Por qué funciona esto?

De la misma manera que en la regresión lineal, uno podría enfocarse en encontrar el B tal que se minimice la diferencia entre lo que predice el modelo y el resultado y_k clasificado, es decir:

$$\min_{\mathbf{B}} \sum_{i=1}^{N} \left\| y_i - \left[(1, x_i^T)\mathbf{B}\right]^T \right\|^2$$

Luego, definiendo las categorías target como vectores columna t_k que valen 1 cuando nos referimos a la categoría k y 0 en los otros casos, el proceso de clasificación consiste en:

$$\hat{G}(x) = \arg\min_k \|f(x) - t_k\|$$

lo cual es matemáticamente equivalente a lo anterior.

Este enfoque sencillo tiene problemas cuando hay 3 o más clases en $\mathcal{G}$ dado que hay clases que no son bien representadas y resultan orientadas hacia las otras dos.

MÉTODO DE DISCRIMINACIÓN LINEAL (LDA)

Como vimos anteriormente, una forma de función discriminante es la probabilidad posterior $P(G = k|X = x)$. Esta probabilidad posterior se calcula como:

$$P(G = k|X = x) = \frac{P(X = x|G = k)P(G = k)}{\sum_{l=1}^{K} P(X = x|G = l)P(G = l)}$$

En notación más compacta, las verosimilitudes se escriben como $f_k(x) \equiv P(X = x|G = k)$, y las prioris de las clases como π_k:

$$P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Para obtener un modelo de clasificación, uno debe proponer las funciones f_k y las prioris π_k.

Una primera opción para la verosimilitud es una distribución gaussiana multivariada sobre los x, con centros ubicados en las clases k:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k^T)\Sigma_k^{-1}(x-\mu_k)}$$

Sin embargo, como modelo paramétrico de las K clases, resulta muy engorroso tener $K \times \frac{p(p+1)}{2}$ parámetros. Por lo tanto, un enfoque más simple es considerar que, aunque

las medias de cada clase son diferentes, todas comparten la misma matriz de covarianza $\Sigma_k = \Sigma$ para todo k.

Retomando el proceso de construcción de fronteras que mencionamos antes, si tomamos el "log-ratio"para dos clases dado un dato, tenemos:

$$\log \frac{P(G = k|X = x)}{P(G = l|X = x)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l)$$

(Ejercicio: deducir la expresión anterior)

Como la dependencia del Bayes factor con x es lineal, las fronteras entre los resultados positivos y negativos de la comparación de modelos están regidas por un hiperplano que divide ambas clasificaciones.

Ahora bien, recordemos que la comparación de modelos, o más generalmente, hipótesis, a partir de los factores de Bayes se basa en una comparación dual. Es decir, solo podemos hacer comparaciones entre dos alternativas. A la hora de introducir una tercera, debemos hacer todas las comparaciones y ver cuál predomina. Esto hace que el problema sea equivalente a proponer una función de discriminación lineal $\delta_k(x)$ de la forma:

$$\delta_k(x) = \log \pi_k - \frac{1}{2}\mu_k \Sigma^{-1} \mu_k + x^T \Sigma^{-1} \mu_k$$

Y la función de clasificación $\mathcal{G}$ resulta en $G(x) = \text{argmax}_k \delta_k(x)$.

¿Cómo procedemos en la práctica?

Una vez que definimos que las funciones de verosimilitudes son gaussianas multivariadas que comparten la matriz de covarianza, debemos estimar los parámetros de dichas distribuciones para poder hacer clasificaciones con datos futuros. Una primera alternativa es usar los estimadores de máxima verosimilitud:

- $\hat{\pi}_k = \frac{N_k}{N}$
- $\hat{\mu}_k = \frac{\sum_{i=1}^{N} x_i}{N}$
- $\Sigma = \frac{\sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{N-K}$

Habiendo obtenido estos parámetros, podemos calcular G para cada dato nuevo de entrada.

Obs

En el caso de que haya sólo dos clases y que el número de datos de cada clase coincida, la frontera resultante del factor de Bayes para la regresión lineal y para el LDA coinciden. Sin embargo, para una mayor cantidad de clases y diferente número de puntos, los resultados no coinciden y el LDA soluciona los problemas de enmascaramiento que genera la regresión lineal.

DISCUSIÓN

- ¿Cómo se podría *bayesianizar* este método?
- ¿Qué cambiaría?
- ¿Cómo afectaría el número de datos de cada clase a las fronteras de las etiquetas?

A. QDA

El siguiente paso para complejizar el modelo consiste en relajar la hipótesis donde todas las matrices de covarianza coinciden. Esto hace que a la hora de calcular la función de discriminación no se cancelen los factores cuadráticos y, por lo tanto, el resultado no sea lineal en x. En este caso, el resultado es:

$$\delta(x)_k = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k \tag{3}$$

Las fronteras, por lo tanto, están dadas por las cuádricas definidas por $\{x : \delta_k(x) = \delta_l(x)\}$.

B. Estimación de parámetros

Para hacer la estimación de parámetros, procedemos de la misma manera que LDA, solo que tenemos tantas matrices de covarianzas como etiquetas k:

$$\hat{\pi}_k = \frac{N_k}{N} \tag{4}$$

$$\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

$$\Sigma_k = \frac{1}{N-K}\sum_{g_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T \tag{6}$$

Reflexiones de Hastie

"The question arises why LDA and QDA have such a good track record. The reason is not likely to be that the data are approximately Gaussian, and in addition for LDA that the covariances are approximately equal. More likely a reason is that the data can only support simple decision boundaries such as linear or quadratic, and the estimates provided via the Gaussian models are stable. This is a bias-variance tradeoff—we can put up with the bias of a linear decision boundary because it can be estimated with much lower variance than more exotic alternatives. This argument is less believable for QDA, since it can have many parameters itself, although perhaps fewer than the non-parametric alternatives."

C. Implementación

Encontramos una base de autovectores de Σ (o Σ_k) ortonormales tal que $\Sigma = UDU^T$, donde D es diagonal.

Luego, podemos calcular el determinante de Σ utilizando los autovalores:

$$\log|\Sigma| = \sum_{l} \log d_{ll} \tag{7}$$

Para clasificar, podemos hacer la siguiente transformación a los datos:

$$X \to X^*X^* = D^{-\frac{1}{2}}U^T X \tag{8}$$

Ahora no habrá covarianza entre los datos y, a su vez, su varianza será 1.

D. Implementación

- Encontrar una base de autovectores de Σ (o Σ_k) ortonormales tal que $\Sigma = UDU^T$ donde D es diagonal.
- Luego, $\log|\Sigma| = \sum_l \log d_{ll}$

Para clasificar, podemos hacer la siguiente transformación a los datos:

$$X \to X^*X^* = D^{-\frac{1}{2}}U^TX. \tag{9}$$

Ahora no habrá covarianza entre los datos y, a su vez, su varianza será 1.

Finalmente, clasificamos los datos X^* con el centroide más cercano ponderado con la prior. En el lenguaje de las funciones discriminantes, resulta:

$$\delta(x^*)_k = -\frac{1}{2}||x^* - \mu_k||^2 + \log \pi_k \tag{10}$$

E. Reduced Rank LDA

En caso de tener K centroides en un espacio p-dimensional de los datos de entrada, existe un subespacio afin H del $\mathbb{R}^p$ que los contiene. En caso de que p sea mucho más grande que K, esto resulta en una reducción de dimensión considerable, dado que la dimensión de H será menor a $K-1$. Cuando clasificamos encontrando al centroide más cercano, podemos ignorar las distancias que son ortogonales al subespacio H, dado que contribuyen de igual manera a la clasificación. Por lo tanto, podemos sin problemas proyectar los datos X^* en este subespacio y hacer las comparaciones de distancias allí. De esta manera, existe una gran reducción de dimensionalidad en el LDA, ya que solo tenemos que considerar proyecciones de los datos en un subespacio de dimensión a lo sumo $K-1$.

En caso de querer seguir reduciendo la dimensión, una vez estando en la variedad afin H, es posible realizar un análisis de componentes principales. Este procedimiento coincide con lo que se conoce como el criterio de optimización de Fisher. La idea principal es buscar las direcciones en las que más se separan los centroides y analizar los datos proyectados sobre esas direcciones.

Formalmente, queremos encontrar una combinación lineal $Z = a^TX$ tal que la varianza entre clases sea máxima en comparación con la varianza que hay individualmente en cada clase.

En el caso de las varianzas de cada clase, cada una está representada por la matriz Σ_K, que en el LDA corresponden a una misma matriz Σ y la notaremos W por *within class*.

Para analizar lo que sucede entre clases, *between classes*, calculamos un vector μ y una matriz de covarianza B (*between classes*):

$$\mu = \sum_{k=1}^{K} \pi_k \mu_k \tag{11}$$

$$B = \sum_{k=1}^{K} \pi_k (\mu_k - \mu)(\mu_k - \mu)^T \tag{12}$$

Lo que buscó optimizar Fisher es:

$$\max_a \frac{a^T B a}{a^T W a} \tag{13}$$

Para resolver esto, comenzamos descomponiendo en autovectores a $W = V_W D_W V_W^T$. Definiendo vectores $b = W^{-\frac{1}{2}} a$, con $W^{\frac{1}{2}} = D_W^{\frac{1}{2}} V_W^T$.

La optimización resulta:

$$\max_b \frac{b^T (W^{-\frac{1}{2}})^T B W^{-\frac{1}{2}} b}{b^T b} \tag{14}$$

Si definimos $B^* = (W^{-\frac{1}{2}})^T B W^{-\frac{1}{2}}$ y lo descomponemos de forma diagonal $B^* = V_{B^*} D_{B^*} V_{B^*}^T$, resulta que la dirección b que maximiza la distancia será aquella asociada al autovector con autovalor más grande (la matriz es simétrica y real, por lo tanto, sus autovalores son positivos). Es decir, si llamamos a las columnas de $V_{B^*} = \{v_1, \ldots, v_K\}$ ordenadas según los autovalores de mayor tamaño, estos definen las direcciones de máxima separación de los centroides de las clases.

II. LOGISTIC REGRESSION

Estos modelos nacen del deseo de describir las probabilidades posterior de las K clases de forma lineal en x, manteniendo la restricción de que sumen 1 y que estén acotadas en el intervalo $[0, 1]$. La forma de especificar el modelo es:

$$\log \frac{P(G = 1|X = x)}{P(G = K|X = x)} = \beta_{10} + \beta_1^T x \tag{15}$$

$$\log \frac{P(G = 2|X = x)}{P(G = K|X = x)} = \beta_{20} + \beta_2^T x \tag{16}$$

$$\log \frac{P(G=K-1|X=x)}{P(G=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x \tag{17}$$

La elección de la comparación con la última categoría es arbitraria en el sentido de que los resultados de las estimaciones no se ven afectados.

Las posteriors bajo esta construcción resultan (Ejercicio):

$$P(G=k|X=x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{\updownarrow=1}^{K-1} (\beta_{\updownarrow 0} + \beta_l^T x)} \tag{18}$$

$$P(G=K|X=x) = \frac{1}{1 + \sum_{\updownarrow=1}^{K-1} (\beta_{\updownarrow 0} + \beta_l^T x)} \tag{19}$$

Es fácil ver que las probabilidades suman 1 y que están acotadas entre 0 y 1. Como es de costumbre, llamaremos al conjunto de todos los parámetros $\theta = \{\beta_{10}, \beta_1^T, \ldots, \beta_{(K-1)0}, \beta_{K-1}^T\}$ y escribiremos $P(G=k|X=x) = P_k(x|\theta)$.

Para determinar los valores de θ, el camino más usual es hacer maximum likelihood sobre la distribución condicional de G dado X.

A. Ejemplo

Imaginemos que tenemos N observaciones de una variable respuesta y_i que toma valores 0 o 1, es decir $K=2$. En este caso, hay dos probabilidades $P_1(x|\theta)$ y $P_0(x|\theta) = 1 - P_1(x|\theta)$. La likelihood de esas respuestas es:

$$\mathcal{L}(\theta) \propto \prod_{i=1}^{N} \left[P_1(x_i|\theta)^{y_i} + (1 - P_1(x_i|\theta))^{(1-y_i)} \right] \tag{20}$$

Para calcular el máximo tomamos logaritmo, y obtenemos:

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} y_i \log P_1(x_i|\theta) + (1-y_i) \log(1 - P_1(x_i|\theta))$$

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} y_i(\beta_{10} + \beta_1^T x_i) - \log(1 + e^{\beta_{10} + \beta_1^T x_i})$$

Tomando la notación de $x_i = (1, x_i)$ y $\beta = (\beta_{10}, \beta_1)$. Derivando e igualando a cero encontramos lo que se llaman las funciones de *score*:

$$\nabla_\beta \mathcal{L} = 0$$

$$\sum_{i=1}^{N} x_i(y_i - P(x_i|\beta)) = 0$$

Estas son $p+1$ ecuaciones, p de la dimensión de x más una del valor de la constante. La primera función de *score* resultará en $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} P(x_i|\beta)$, lo cual manifiesta que el número observado de casos 1 coincide con el número esperado de la variable aleatoria y. El resto de las ecuaciones son no lineales en β y suelen ser resueltas numéricamente con el método de Newton-Raphson. (Ver Hastie, pp. 120-122)

DIFERENCIAS ENTRE LDA Y LOGISTIC REGRESSION

Si hacemos una comparación en cómo se generan las fronteras de decisión en cada modelo tenemos:

$$\log \frac{P(G=k|X=x)}{P(G=K|X=x)} = \log \frac{\pi_k}{\pi_K} - \frac{1}{2}(\mu_k + \mu_K)^T \Sigma^{-1}(\mu_k - \mu_K) + x^T \Sigma^{-1}(\mu_k - \mu_K) \quad (21)$$

$$\log \frac{P(G=k|X=x)}{P(G=K|X=x)} = \alpha_0 + \alpha_k^T x \quad (22)$$

La linealidad en el logit es consecuencia de la suposición gaussiana y que la covarianza sea la misma para todas las clases. Luego, por construcción, la regresión logística es lineal también:

$$\log \frac{P(G=k|X=x)}{P(G=K|X=x)} = \beta_{k0} + \beta_k^T x \quad (23)$$

Desde este punto de vista, ambos modelos parecen ser idénticos. Sin embargo, la manera en que se estiman los parámetros es muy diferente. En uno utilizamos el máximo likelihood de las gaussianas e introducimos esos estimadores como los parámetros μ_k, Σ y las priors π_k. En el otro caso, sólo maximizamos la likelihood condicional de las variables respuesta G (o y). Analíticamente, lo que sucede es que al mirar la distribución conjunta de X y $G = k$ tenemos:

$$P(X, G = k) = P(X)P(G = k|X) \tag{24}$$

En la regresión logística, el término $P(X)$ no es tenido en cuenta en el ajuste de los datos. Esto puede ser considerado una bondad del método o un defecto, ¿por qué? En cambio, para el LDA:

$$P(X, G = k) = P(X|G = k)P(G = k) = \phi(X|\mu_k, \Sigma)\pi_k \tag{25}$$

Y por lo tanto:

$$P(X) = \sum_{k=1}^{K} \pi_k \phi(x|\mu_k, \Sigma) \tag{26}$$

Esto resulta en una mezcla de gaussianas que tiene incorporados en su construcción a los parámetros.

El resultado de tomar como suposición que los datos provienen de esta distribución genera que los parámetros se puedan estimar de una manera más eficiente, es decir, que tengan menos varianza. Según Efron 1975, si los datos provienen efectivamente de esta distribución, ignorar la prior marginal $P(X)$ resulta en una necesidad de un 30 % más de datos para que la likelihood resulte en los mismos resultados.

Desde otro punto de vista, puede considerarse que tener definida la prior marginal para los datos como función de los parámetros funciona como un regularizador, obligando a que las clases pueden ser destinguidas. Por ejemplo si los datos en un problema de dos clases puede ser separados perfefctamente por un hyperplano, el máximo likelihood para la regresión logistica resultan mal definidos, en cambio para el LDA no hay dificultades.

Por otro lado, al ser más informativo el LDA es menos robusto frente *outliers* por lo que cuando la suposiciones no son adecuadas el método puede fallar. En este caso podemos decir que la regresión logistica es más robusta porque tiene menos suposiciones.