Indice

Capitolo 1

Metodi lineari di classificazione

1.1 Introduzione

In questo capitolo rivisitiamo il problema di classificazione e ci concentriamo in metodi lineari di classificazione. Dato che il nostro predittore $G(x)$ prende valori di un insieme discreto $\mathcal{G}$, possiamo sempre dividere lo spazio di input in una collezione di regioni secondo la classificazione. Abbiamo visto nel Capitolo 2 che le frontiere di questi regioni possono essere ruvide o lisce, dipendendo della funzione di predizione. Per una classe importante di procedure, queste *frontiere di decisione* sono lineari; questo è ciò che intendiamo con metodi lineari di classificazione.

Ci sono diversi modi in cui le frontiere di decisione lineari possono essere trovate. Nel Capitolo 2 abbiamo adattato modelli di regressione lineare alle variabili indicatori di classi, e dopo classifichiamo secondo il **largest fit**. Supponiamo che ci sono K classi, etichettate come $1, 2, \ldots, K$ per comodità, ed il modello lineare adattato per la k-esima variabile indicatore di risposta sia $\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$. La frontiera di decisione tra la classe k e l è il insieme di punti per cui $\hat{f}_k(x) = \hat{f}_l(x)$, cioè, il insieme $\{x : \left(\hat{\beta}_{k0} - \hat{\beta}_{l0}\right) + \left(\hat{\beta}_k - \hat{\beta}_l\right)^T x = 0\}$, che è un insieme affine o un iperpiano[1]. Siccome lo stesso è vero per ogni paio di classi, lo spazio di input è diviso in regioni di classificazione costante, con frontiere di decisione che sono iperplanare a tratti. Questo approccio con regressione fa parte di una classe di metodi che modellano *funzioni discriminanti* $\delta_k(x)$ per ogni classe, e dopo classificano x alla classe con il valore più grande nella sua funzione discriminante. Metodi che modellano le probabilità posteriori $\Pr(G = k|X = x)$ appartengono anche a questa classe. Chiaramente, se $\delta_k(x)$ o $\Pr(G = k|X = x)$ sono lineare in x, allora le frontiere di decisione saranno anche lineare.

In realtà, tutto ciò di cui abbiamo bisogno è che alcuna trasformazione monotona di δ_k o $\Pr(G = k|X = x)$ sia lineare affinché le frontiere di decisione siano lineari. Per esempio, se ci sono due classi, un modello popolare per le probabilità posteriori è

[1] In senso stretto, un iperpiano attraversa l'origine, mentre un insieme affine non necessariamente. A volte ignoriamo la distinzione e intendiamo iperpiani in generale

$$\begin{aligned}\Pr(G=1|X=1) &= \frac{e^{\beta_0+\beta^T x}}{1+e^{\beta_0+\beta^T x}},\\ \Pr(G=2|X=1) &= \frac{1}{1+e^{\beta_0+\beta^T x}}.\end{aligned} \tag{1.1}$$

La trasformazione monotona qui è la trasformazione *logit*: $\ln[p/(1-p)]$[2], e infatti vediamo che

$$\ln\frac{\Pr(G=1|X=1)}{\Pr(G=2|X=1)} = \beta_0+\beta^T x. \tag{1.2}$$

La frontiera di decisione è il insieme di punti dove le log-probabilità sono zero, e questo è un iperpiano definito da $\{x|\beta_0+\beta^T x=0\}$. Discutiamo qui due metodi molto popolari ma diversi che risultano in log-probabilità lineari o *logits*: il analisi discriminante lineare e la regressione logistica lineare. Sebbene sono diversi nella sua derivazione, la differenza essenziale tra loro è nel modo in cui la funzione lineare si adatta ai dati di addestramento.

Un approccio più diretto è modellare esplicitamente le frontiere tra le classi come lineari. Per un problema di due classi in uno spazio di input p-dimensionale, questo equivale a modellare la frontiera di decisione come un iperpiano — in altre parole, un vettore normale e un punto di taglio. Vedremo due metodi che esplicitamente cercano "iperpiani di separazione". Il primo è il ben noto modello del percettrone di Rosenblatt (1958), con un algoritmo che trova un iperpiano di separazione nei dati di allenamento, se esiste. Il metodo secondo, dovuto a Vapnik (1996), trova un iperpiano di separazione ottimo se esiste uno, altrimenti trova un iperpiano che minimizza alcuna misura di sovrapposizione nei dati di allenamento. Trattiamo il caso separabile qui, e differiamo il trattamento del caso non separabile al Capitolo 12.

Mentre questo capitolo intero è dedicato a frontiere di decisione lineari, c'è uno spazio considerabile di generalizzazione. Per esempio, possiamo espandere il nostro insieme di variabile $X_1,\ldots,X_p$ includendo le loro quadrati e prodotti incrociati $X_1^2,X_2^2,\ldots,X_1X_2,\ldots$, aggiungendo in tal modo $p(p+1)/2$ variabile addizionali. Le funzioni lineari nello spazio aumentato mappano a funzione quadratiche nello spazio originale — quindi mappano frontiere di decisione lineari a frontiere di decisione quadratiche. La figura 1.1 illustra l'idea. I dati sono i stessi: il disegno della sinistra usa frontiere di decisione lineari nello spazio due-dimensionale mostrato, mentre che il disegno della destra usa frontiere di decisione lineare nello spazio aumentato cinque-dimensionale descritto sopra. Questo approccio può essere usato con qualsiasi trasformazione di base $h(X)$ dove $h:\mathbb{R}^p\to\mathbb{R}^q$ con $q>p$, e sarà esplorato in capitoli posteriori.

1.2 Regressione lineare di una matrice indicatore

Qui ognuna delle categorie di risposta si codifica tramite una variabile indicatore. Così, se $\mathcal{G}$ ha K classi, ci saranno K tali indicatori Y_k, $k=1,\ldots,K$, con $Y_k=1$ se $G=k$ e 0 in

[2]Se $p_1=\Pr(G=1|X=1)$ e $p_2=\Pr(G=2|X=1)$, poi abbiamo che $\ln[p_1/(1-p_1)]=\beta_0+\beta^T x$ e $\ln[p_2/(1-p_2)]=-\left(\beta_0+\beta^T x\right)$.

Figura 1.1: Il disegno della sinistra mostra alcuni dati di tre classi, con frontiere di decisione lineari trovate con il analisi discriminante lineare. Il disegno della destra mostra frontiere di decisione quadratiche. Quelle sono state ottenute trovando frontiere lineari nello spazio cinque-dimensionale $X_1, X_2, X_1X_2, X_1^2, X_2^2$. Disuguaglianze lineari in questo spazio sono disuguaglianze quadratiche nello spazio originale.

caso contrario. Quelle vengono raccolte in un vettore $Y = (Y_1, \ldots, Y_k)$, e le N istanze di allenamento di quelle formano una matrice indicatore di risposta $\mathbf{Y}$ di $N \times K$. Adattiamo simultaneamente un modello di regressione lineare ad ognuna delle colonne di $\mathbf{Y}$, e il *fit* è dato da

$$\hat{\mathbf{Y}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{1.3}$$

Il Capitolo 3 ha più dettagli sulla regressione lineare. Nota che abbiamo un vettore di coefficienti per ogni colonna di risposta $\mathbf{y}_k$, e quindi abbiamo una matrice di coefficienti $\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ di $(p+1) \times K$. Ecco, $\mathbf{X}$ è la matrice modello con $p+1$ colonne corrispondente ai p input, e una prima colonna di 1 per l'intercettazione.

Una nuova osservazione con input x è classificata come segue:

- calcola il risultato adattato $\hat{f}(x)^T = (1, x^T)\hat{\mathbf{B}}$, che è un K-vettore;
- individua il componente più grande e classifica perciò:

$$\hat{G}(x) = \text{argmax}_{k \in \mathcal{G}}\hat{f}_k(x). \tag{1.4}$$

Qual è la logica di questo approccio? Una giustificazione piuttosto formale è vedere la regressione come una stima della aspettazione condizionata[3]. Per la variabile aleatoria Y_k, $E(Y_k|X = x) = \Pr(G = k|X = x)$, quindi la aspettazione condiziona per ognuna delle Y_k sembra di essere un obiettivo ragionevole. Il vero problema è: quanto è buono come approssimazione della aspettazione condizionata il modello rigido della regressione lineare? In alternativa, sono le $\hat{f}_k(x)$ stime ragionevoli delle probabilità posteriori $\Pr(G = k|X = x)$, e più importante, questo importa?

È abbastanza semplice verificare che $\sum_{k \in \mathcal{G}} \hat{f}_k(x) = 1$ per ogni x, finché c'è una intercettazione nel modello (la colonna di 1 in $\mathbf{X}$). Tuttavia, le $\hat{f}_k(x)$ possono essere negative o maggiore di 1, e in genere alcuni lo sono. Questo è una conseguenza della natura rigida della regressione lineare, soprattutto se facciamo predizioni fuori *dall'ambito* dei dati di allenamento. Questi violazioni da sole non garantiscono che questo approccio non funzionerà, e infatti in molti problemi dà risulti simili a metodi lineari di classificazione più standard. Se facciamo regressione lineare su espansioni di base $h(X)$ delle inputs, questo approccio può portare a stime coerenti delle probabilità. Man mano che le dimensioni N del insieme di allentamento aumentano, includiamo più elementi di base, così che la regressione lineare su queste funzioni di base si avvicina alla aspettazione condizionata. Discutiamo tali approcci nel Capitolo 5.

Un punto di vista più semplicistico è quello di costruire *obiettivi* t_k per ogni classe, dove t_k è la k-esima colonna della matrice identità di $K \times K$. Il nostro problema di predizione è quello di

[3] Questo si vede nel Capitolo 3

provare e riprodurre l'obiettivo adeguato per una osservazione. Con la stessa codifica di prima, il vettore di risposta y_i (i-esima riga di $\mathbf{Y}$) per la osservazione i ha il valore $y_i = t_k$ se $g_i = k$. Poi potremmo adattare il modello lineare *ai* minimi quadrati[4]

$$\min_{\mathbf{B}} \sum_{i=1}^{N} \|y_i - [(1, x_i^T)\mathbf{B}]^T\|^2. \tag{1.5}$$

[4]Nota che $(1, x_i^T)$ è la i-esima riga di $\mathbf{X}$.