Regressão Logística

Já vimos que a distribuição de Bernoulli representa a probabilidade associada com a variável $Y \in \{0,1\}$, sendo descrita por

$$P(Y) = p^{Y}(1-p)^{1-Y}$$

se p o parâmetro da distribuição. Também estudamos como obter p a partir de conjunto de dados $\{Y_i\}_{i=1}^{N}$.

Suponha agora que os resultados de Y_i estejam associados com uma variável contínua X, ou seja, temos um conjunto

$$\{(x_i, Y_i)\}_{i=1}^{n}$$

e queremos encontrar uma possível relação entre p e X. Tal como numa regressão linear, poderíamos supor

$$p = ax + b$$

com a e b sendo parâmetros. Entretanto, visto que p é uma probabilidade $p \in (0,1)$, devemos "envelopar" a função linear. Uma possibilidade é usar a função logística

$$\theta(s) = \frac{e^{s}}{1+e^{s}}$$

e, desse modo, temos

$$\hat{p} = \theta(ax+b) = \frac{e^{ax+b}}{1+e^{ax+b}}$$

Uma maneira mais compacta de escrever a relação anterior é usar a função logit

$$\text{logit}(t) = \cancel{\log}\left(\frac{t}{1-t}\right) = \log\left(\frac{t}{1-t}\right)$$

de modo que

$$\text{logit}(\hat{p}) = b + aX$$

No caso mais de uma variável preditora, ficamos com

$$\text{logit}(\hat{p}) = b + \sum_k a_k X_k$$

<u>Exemplo notebook</u>

Para compreender mais formalmente a regressão logística, considere a expressão para $\hat{p}$ escrita como

$$p(\bar{x}) = \frac{1}{1 + \exp(-\bar{B}^T \bar{x})}$$

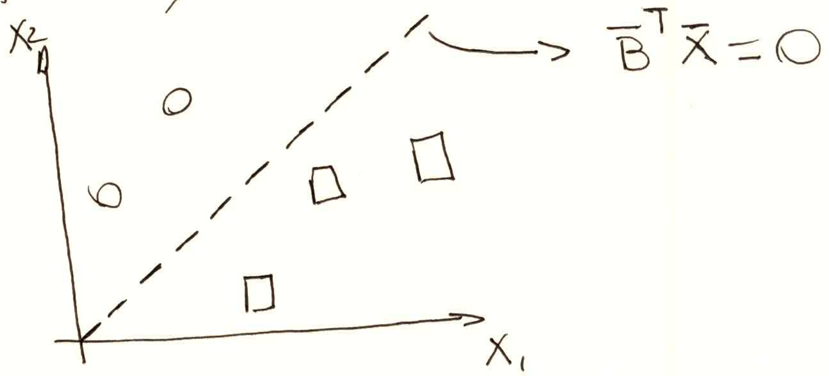
com $\bar{x}, \bar{B} \in \mathbb{R}^n$ (n é número features). Dos resultados sobre projeção, sabemos que a distância entre $\bar{x}$ e $\bar{B}$ (perpendicular) é $\frac{\bar{B}^T \bar{x}}{\|\bar{B}\|}$. Assim, a probabilidade associada com um ponto em $\mathbb{R}^n$ é uma função de quão próximo o ponto está do contorno linear definido por

$$\bar{B}^T \bar{x} = 0$$

Entretanto, notamos que para algum $\alpha \in \mathbb{R}$, temos

$$\alpha \bar{B}^T \bar{X} = 0$$

ou seja, $\alpha \bar{B}^T$ define o mesmo plano. Por outro lado, essa constante α determina a intensidade da probabilidade atribuída a $\bar{X}$. Suponha um espaço com duas features, conforme ilustra a figura abaixo



embora o plano (reta) separando os dois conjuntos seja uma só $\beta_1 X_1 + \beta_2 X_2 = 0$, a probabilidade associada a cada (X_1, X_2)

$$P(X_1, X_2) = \frac{1}{1 + \exp\left[-\alpha(\beta_1 X_1 + \beta_2 X_2)\right]}$$

depende de α. Na prática, esse problema é resolvido por um procedimento chamado regularização, que basicamente consiste em penalizar o tamanho de β adicionando um termo $\frac{\|B\|}{C}$ na quantidade a ser minimizada (least squares error).

Exemplo notebook

Modelos lineares generalizados

Regressão logística é um exemplo de uma classe de modelos lineares generalizados que envolvem transformações lineares no processo de ajuste. De modo geral, podemos pensar que nosso objetivo é estimar

$$E(\bar{Y} | \bar{X} = \bar{x})$$

No caso da regressão linear

$$E(\bar{Y} | \bar{X} = \bar{x}) \approx \bar{\beta}^T \bar{x}$$

No caso da regressão logística, como $Y \in \{0, 1\}$, temos que

$$E(\bar{Y} | \bar{X} = \bar{x}) = P(\bar{Y} | \bar{X} = \bar{x}) = r(\bar{x})$$

e a transformação faz $r(\bar{x})$ ficar linear, isto é,

$$\begin{aligned} \eta(\bar{x}) &= \bar{\beta}^T \bar{x} \\ &= \log \frac{r(\bar{x})}{1 - r(\bar{x})} \\ &= g[r(\bar{x})] \end{aligned}$$

sendo g a função link logístico e $\eta(\bar{x})$ o preditor linear. Uma vez que transformamos o dado na forma linear, é tentador usar uma regressão usual para ajustar a variável binária Y_i transformada. Entretanto, podemos obter ~~log(0)~~ log(0) ou log(1/0). Uma alternativa é usar uma série de Taylor para expandir $g(Y)$ ao redor de $r(\bar{x})$, isto é,

$$g(Y) \approx \log \frac{r(\bar{x})}{1-r(\bar{x})} + \frac{Y-r(x)}{r(x)-r^2(x)}$$

$$= n(\bar{x}) + \frac{Y-r(x)}{r(x)-r^2(x)}$$

A parte interessante é o termo $Y-r(\bar{x})$, pois é onde surge a classe $Y \in (0,1)$. Vale notar que

$$E(Y-r(x)|x) = 0$$

de modo que o segundo termo funciona com um ruído aditivo para $n(\bar{x})$. Por outro lado,

$$\mathbb{V}(g(Y)|X) = \frac{1}{r(x)(1-r(x))}$$

é uma função de x, o que significa que o valor de x altera não somente a probabilidade $P(\bar{Y}|\bar{X}=\bar{x})$ mas também sua variância.

Regularização

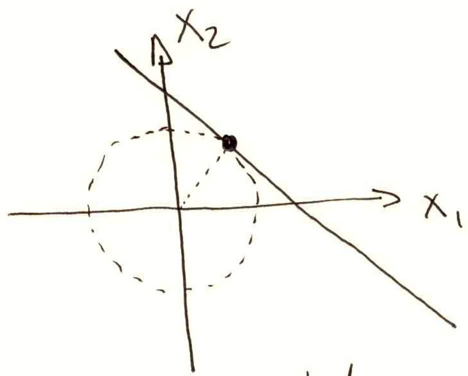
Regularização é um processo pelo qual lidamos com o trade-off bias-variance. Para começar, considere

O problema do mínimos quadrados

$$\underset{\bar{x}}{\text{minimizar}} \ \|\bar{x}\|_2^2$$

$$\text{sujeito a} \quad x_0 + 2x_1 = 1$$

como $\|\bar{x}\|_2 = \sqrt{x_0^2 + x_1^2}$ a norma L_2. Sem o vínculo, é fácil ver que $\bar{x} = 0$ é solução para o problema. Se consideramos o vínculo, podemos pensar no problema como encontrar o menor círculo que toca a linha, numa figura teríamos algo do tipo



Do ponto de vista analítico, podemos obter essa solução via multiplicadores de Lagrange

$$J(x_0, x_1, \lambda) = x_0^2 + x_1^2 + \lambda(1 - x_0 - 2x_1)$$

Vale notar que a solução para (x_0, x_1) tem uma vizinhança que pode levar a valores muito bons para o problema de minimização. Em alguns casos, isso pode ser um problema. Para contornar isso, podemos usar a norma L_1, isto é,

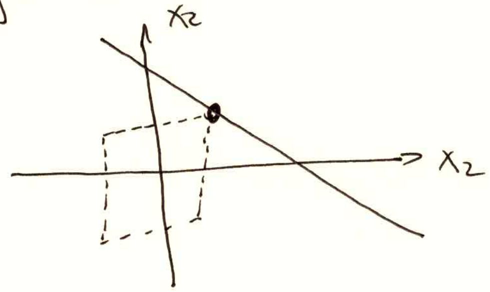
$$\|\bar{x}\|_1 = \sum_{i=1}^{d} |x_i|$$

de modo que o problema se torna

minimizar $\|\bar{x}\|_1$

sujeito a $x_1 + 2x_2 = 1$

Ocorre que esse problema é mais complicado para se obter uma solução analítica. Nesse caso, num gráfico, teríamos algo como



Note que, nesse caso, a solução tem uma vizinhança que não leva a valores tão próximos ao melhor valor. Isso se torna mais pronunciado à medida que a dimensão do espaço aumenta.

<u>Exemplo notebook</u>

Regressão Ridge

Lembrando da regressão linear usual, estamos interessados em resolver o problema

$$\min_{\beta \in \mathbb{R}^p} \|y - \bar{X}\bar{\beta}\|$$

sendo $\bar{X} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p]$ com $\bar{x}_i \in \mathbb{R}^n$. Vale notar que que se $p > n$, não existe um único $\bar{\beta}$, mas infinitos.

Outro problema que pode ocorrer é quando os $\bar{x}_i$ são colineares. Nesse caso, a inversa

$$(\bar{X}^T\bar{X})^{-1}$$

não é bem definida. Uma possibilidade para resolver esse problema é considerar o seguinte problema de minimização

$$\min_{\bar{\beta} \in \mathbb{R}^p} \|y - \bar{X}\beta\|_2^2 + \alpha \|\beta\|_2^2$$

onde α é um hiperparâmetro da regressão Ridge. Esse parâmetro controla o trade-off entre minimizar o erro quadrado $\|y - \bar{X}\beta\|_2$ e tamanho de $\|\bar{\beta}\|_2$.

Exemplo notebook

Regressão Lasso

A regressão Lasso (least absolute shrinkage and selection operator) funciona basicamente como a Ridge; porém, no lugar da norma L_2 usamos a L_1, isto é,

$$\min_{\beta \in \mathbb{R}^p} \|y - \bar{X}\bar{\beta}\|^2 + \alpha \|\beta\|_1$$

Em comparação com a Ridge, a regressão Lasso impõe uma restrição maior ao tamanho de $\bar{\beta}$, fazendo com que na prática muitos coeficientes do vetor $\bar{\beta}$ tendam para zero. Por esse motivo, a regressão Lasso é usada para selecionar e reduzir o número de features no modelo linear.

Exemplo notebook

Elastic net regularization

A regressão elástic net é uma combinação das Ridge e Lasso, de modo que tanto a norma L_1 quanto a L_2 são usadas, ou seja,

$$\min_{\beta \in \mathbb{R}^p} \|y - \bar{X}\bar{\beta}\|^2 + \alpha_1 \|\beta\|_2^2 + \alpha_2 \|\beta_1\|$$

Uma das vantagens da elastic net é que diferente do Ridge, ela tende a produzir uma solução esparsa como a Lasso; porém, mais estável e que tende agrupar características que são correlacionadas. Num gráfico teríamos algo como