Sistemas de Apoio à Decisão

Ajustando um Modelo aos Dados e o Problema de Sobreajuste

Vahid Nikoofard

Universidade do Estado do Rio de Janeiro - Faculdade de Tecnologia
Laboratório Interdisciplinar Virtual de Inteligência Artificial (LIVIA)

Novembro 2022

Modelagem Paramétrica: Introdução

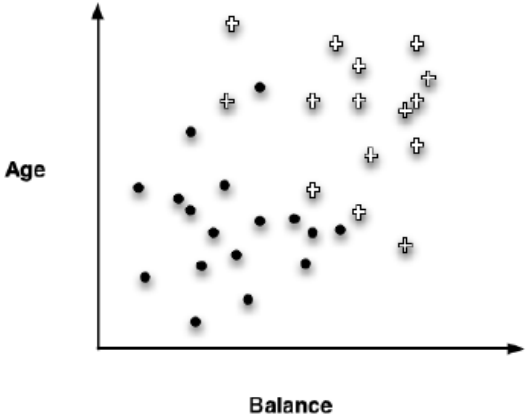
- Definir a estrutura de modelo
- Achar os melhores parâmetros usando os dados

Importante: Todos os atributos têm que ser numéricos. Os atributos categóricos tem que ser convertidos.

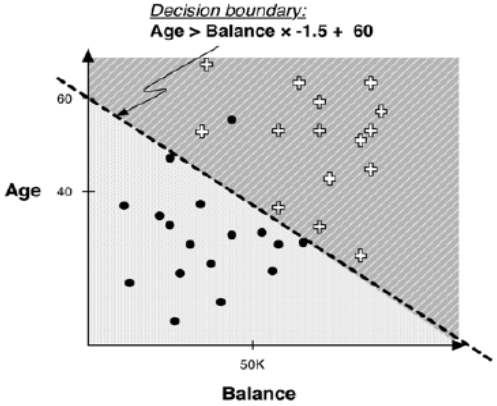
Lembrete: Modelagem por Segmentação

Classificação por Funções Matemáticas

Será que não tem outra forma de criar um *decision boundary*



Classificação por Funções Matemáticas: Classificador Linear



Funções Discriminantes Lineares

- Equação geral :

$$y = mx + b$$

- Exemplo anterior

$$\text{Age} = -1,5 \times \text{Balance} + 60$$

Funções Discriminantes Lineares

- Equação geral :

$$y = mx + b$$

- Exemplo anterior

$$\text{Age} = -1,5 \times \text{Balance} + 60$$

- Função de classificação: discriminante linear

$$\begin{cases} + & \textbf{se} \quad \text{Age} - 1,5 \times \text{Balance} + 60 > 0 \\ \circ & \textbf{se} \quad \text{Age} - 1,5 \times \text{Balance} + 60 < 0 \end{cases}$$

Funções Discriminantes Lineares

- Equação geral :

$$y = mx + b$$

- Exemplo anterior

$$\text{Age} = -1,5 \times \text{Balance} + 60$$

- Função de classificação: discriminante linear

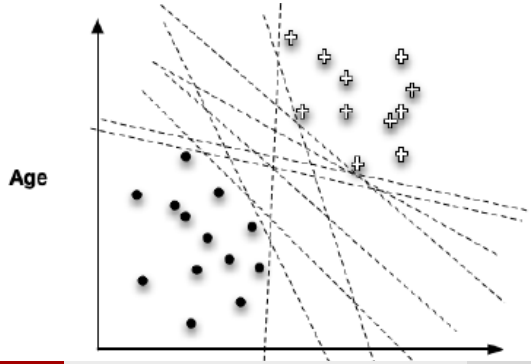
$$\begin{cases} + & \textbf{se} \quad \text{Age} - 1,5 \times \text{Balance} + 60 > 0 \\ \circ & \textbf{se} \quad \text{Age} - 1,5 \times \text{Balance} + 60 < 0 \end{cases}$$

- **Modelo linear geral**

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + ...$$

Modelo Parametrizado

Funções lineares são o carro-chefe da data science; agora, finalmente, chegamos a mineração de dados. Temos um modelo parametrizado: os pesos (*weights*) da função linear (w_i) são os parâmetros. A mineração de dados vai "ajustar" (*fit*) este modelo parametrizado para um conjunto de dados em particular — o que significa, especificamente, encontrar um bom conjunto de pesos dos atributos.



Otimizando uma Função Objetiva

- Importante: qual deve ser a nossa meta ou objetivo na escolha dos parâmetros.
- Função objetiva

Funções Discriminantes Lineares para Casos de Pontuação e Classificação

- Likelihood
- Perto de *decision boundary* $\rightarrow$ mais incerteza
- $f(x) \ll 0$ ou $f(x) \gg 0$

Regressão por Funções Matemáticas: função de custo

- Função de custo, função de perda (*loss*): A função perda determina quanta penalidade deve ser atribuída a um exemplo com base no erro no valor preditivo do modelo — em nosso contexto atual, com base em sua distância do limite de separação.
- hinge loss : (usada no SVM) Para uma saída pretendida $t = \pm 1$ e um escore de classificador y,

$$l(y) = \max(0, 1 - t \cdot y)$$

- Perda zero-um: como o próprio nome indica, atribui uma perda de zero para uma decisão correta e um para uma incorreta.
- Erro absoluto: (regressão) $|y - \hat{y}|$
- Erro quadrático:(regressão) $|y - \hat{y}|^2$

Regressão por Funções Matemáticas

A estrutura do modelo de regressão linear é exatamente a mesma que para a função discriminante linear

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + ...$$

O que é a função objetiva?: minimizar a função de custo

$$\sum_{\text{todos os exemplos}} |y - \hat{y}|^2 (\text{um exemplo})$$

Estimativa de Probabilidade de Classe e "Regressão" Logística

Um dos modelos lineares, escolhendo uma função objetiva diferente, podemos produzir um modelo projetado para dar estimativas precisas de probabilidade de classe. O procedimento mais comum para fazermos isso é chamado **regressão logística**.

- O que exatamente é uma estimativa precisa da probabilidade de pertencer à classe?
  - as estimativas de probabilidade fossem bem calibradas, o que significa que se você tiver 100 casos cuja probabilidade de membro de classe é estimada em 0,2, então, cerca de 20 deles realmente pertencerão à classe. Também gostaríamos que
  - as estimativas de probabilidade fossem discriminativas e, se possível, fornecessem estimativas de probabilidade significativamente diferentes para diferentes exemplos.

Regressão Logística

A função linear, $f(x)$, dá a distância a partir do limite de separação. No entanto, isso também mostra o problema: $f(x)$ varia de $-\infty$ a $+\infty$, e uma probabilidade deve variar de zero a um. Como transformar a distancia do *decision boundary* para probabilidade? Se a função linear calcular a chance

Probability	Corresponding odds
0.5	50:50 or 1
0.9	90:10 or 9
0.999	999:1 or 999
0.01	1:99 or 0.0101
0.001	1:999 or 0.001001

onde *odd* significa Chance, a probabilidade de acontecer dividido por a probabilidade de não acontecer.

Regressão Logística: cont.

A distância a partir do *decision boundary* (limite de decisão) fica entre $-\infty$ e $+\infty$, mas, como podemos ver no exemplo, as chances variam de 0 a $+\infty$. Calculando o logaritmo das chances (*log-odds*), uma vez que, para qualquer número no intervalo de 0 a $+\infty$, seu $\log$ será entre $-\infty$ e $+\infty$

Probability	Odds	Log-odds
0.5	50:50 or 1	0
0.9	90:10 or 9	2.19
0.999	999:1 or 999	6.9
0.01	1:99 or 0.0101	–4.6
0.001	1:999 or 0.001001	–6.9

Regressão Logística: Detalhes técnicos

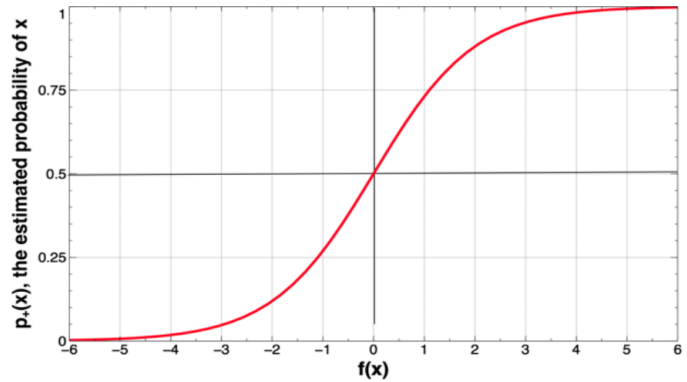
$p_+(x)$ para representar a estimativa do modelo de probabilidade de pertencer à classe + de um item de dados representado por um vetor de característica **x**. A probabilidade estimada de o evento não ocorrer é, portanto, $(1 - p_+(x))$.

$$\log\left(\frac{p_+(\mathbf{x})}{1 - p_+(\mathbf{x})}\right) = f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + ...$$

ou

$$p_+(\mathbf{x}) = \frac{1}{1 + e^{-f(\mathbf{x})}}$$

Função de Sigmoid



A figura mostra que, no limite de decisão (na distância x=0), a probabilidade é 0,5 (cara ou coroa)

Regressão Logística: função objetiva

Gostaríamos que $p_+(x_+)$ estivesse o mais próximo possível de um e que $p_+(x_\circ)$ estivesse o mais próximo possível de zero.

A seguinte função de cálculo da "probabilidade" (*likelihood*) de que um exemplo rotulado em particular pertença à classe correta, dado um conjunto de parâmetros **w** que produz estimativas de probabilidade de classe $p_+(x)$:

$$g(\mathbf{x}, \mathbf{w}) = \begin{cases} p_+(x) & \text{se } \mathbf{x} \text{ é um } + \\ 1 - p_+(x) & \text{se } \mathbf{x} \text{ é um } \circ \end{cases}$$

O modelo (conjunto de pesos) que dá a maior soma é o modelo que dá a maior "probabilidade" aos dados — o modelo de "probabilidade máxima". O modelo de probabilidade máxima "na média" dá as maiores probabilidades para os exemplos positivos e as menores probabilidades para os exemplos negativos.

Etiquetas de Classe e Probabilidades

Uma pessoa pode ser tentada a pensar que a variável alvo é uma representação da probabilidade de pertencer à classe, e os valores observados da variável alvo nos dados de treinamento simplesmente relatam probabilidades de $p(x) = 1$ para os casos que são observados na classe, e $p(x) = 0$ para exemplos que não são observados na classe. No entanto, isso não costuma ser consistente com a forma como são usados os modelos logísticos de regressão. Pegue um aplicativo para marketing direcionado, por exemplo. Para um consumidor c, nosso modelo pode estimar a probabilidade de responder à oferta ser p(c responde) = 0,02. Nos dados, vemos que a pessoa, de fato, responde. Isso não significa que a probabilidade desse consumidor responder foi, na verdade, 1,0, nem que o modelo incorreu um grande erro neste exemplo. A probabilidade dos consumidores pode, de fato, ter sido em torno de p(c responde) = 0,02, o que, na verdade, é uma alta probabilidade de resposta para muitas campanhas, e aconteceu de o consumidor responder desta vez. Uma maneira mais satisfatória de pensar sobre isso é que os dados de treinamento compreendem um conjunto de "eventos" estatísticos a partir das probabilidades subjacentes, em vez de representar as próprias probabilidades subjacentes. O procedimento de regressão logística, então, tenta estimar as probabilidades com um modelo linear de log-chances