人工智能
——样例学习 II

饶洋辉

数据科学与计算机学院,

中山大学

raoyangh@mail.sysu.edu.cn

Expectation

- If X is a discrete random variable

$$E[X] = \sum_i x_i P\{X = x_i\}$$

- If X is a continuous random variable having probability density function f

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]$$

Expectation

- If rolling one die (6-sided) and X is the value on its face, then: $E[X]$?

Expectation

- If rolling one die (6-sided) and X is the value on its face, then: $E[X]$?

$$E[X] = \sum_{x=1}^{6} xp(x) = \frac{1}{6}\sum_{x=1}^{6} x = \frac{21}{6}$$

Median

- Sort n variables
  - $X(1) <= X(2) <= \ldots <= X(n)$
- If n is odd number
  - $X((n+1)/2)$
- If n is even number
  - $(X(n/2)+X(1+n/2))/2$

Mode

- 10 5 9 12
- 6 5 9 8 5
- 25 28 28 36 25 42

Variance

- $\mathrm{Var}(X) = E[(X\text{-}E[X])^2] = E[X^2]\text{-}(E[X])^2$

X	$E(X)$	$(X\text{-}E(X))^2$	X^2
1	2	1	1
2	2	0	4
3	2	1	9

https://blog.csdn.net/hearthougan/article/details/77859173

Covariance

- Cov(X,Y)=$E[(X\text{-}E(X))(Y\text{-}E(Y))]$

$$= E[XY - E(X)Y - XE(Y) + E(X)E(Y)]$$
$$= E[XY] - E(X)E[Y] - E[X]E(Y) + E(X)E(Y)$$
$$= E[XY] - E[X]E[Y]$$

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Pearson correlation coefficient

Linear Regression

- Least-squares solutions

$$n^{-1}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i) = 0$$

$$n^{-1}\sum_{i=1}^{n}x_i(y_i - w_0 - w_1 x_i) = 0$$

$$Q(w_0, w_1) = \min_{w_0, w_1}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)^2$$

$$\partial Q(w_0, w_1) / \partial w_0 = 0$$

$$-2\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i) = 0$$

$$\partial Q(w_0, w_1) / \partial w_1 = 0$$

$$-2\sum_{i=1}^{n}x_i(y_i - w_0 - w_1 x_i) = 0$$

Linear Regression

- Least-squares solutions

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$w_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Logistic Regression

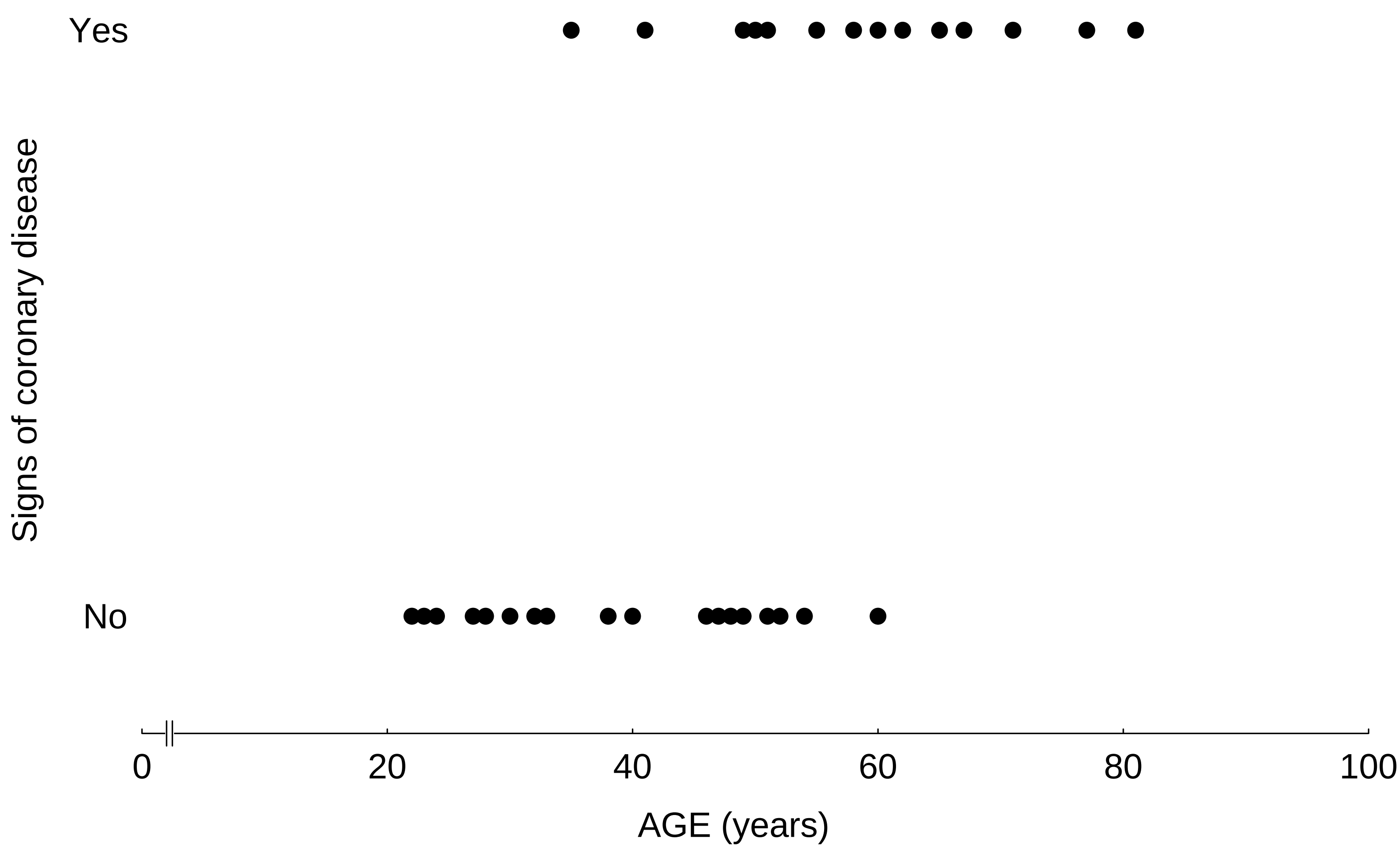
- We may use the linear regression model for binary classification

$$y = w_0 + \sum_{j=1}^{d} w_j x_j + u$$

$$= \tilde{\mathbf{W}}^{\mathrm{T}} \tilde{\mathbf{X}}$$

  - However, the predicted *y* values (预测的y值) can be greater than 1 or less than 0

Logistic Regression



Logistic Regression

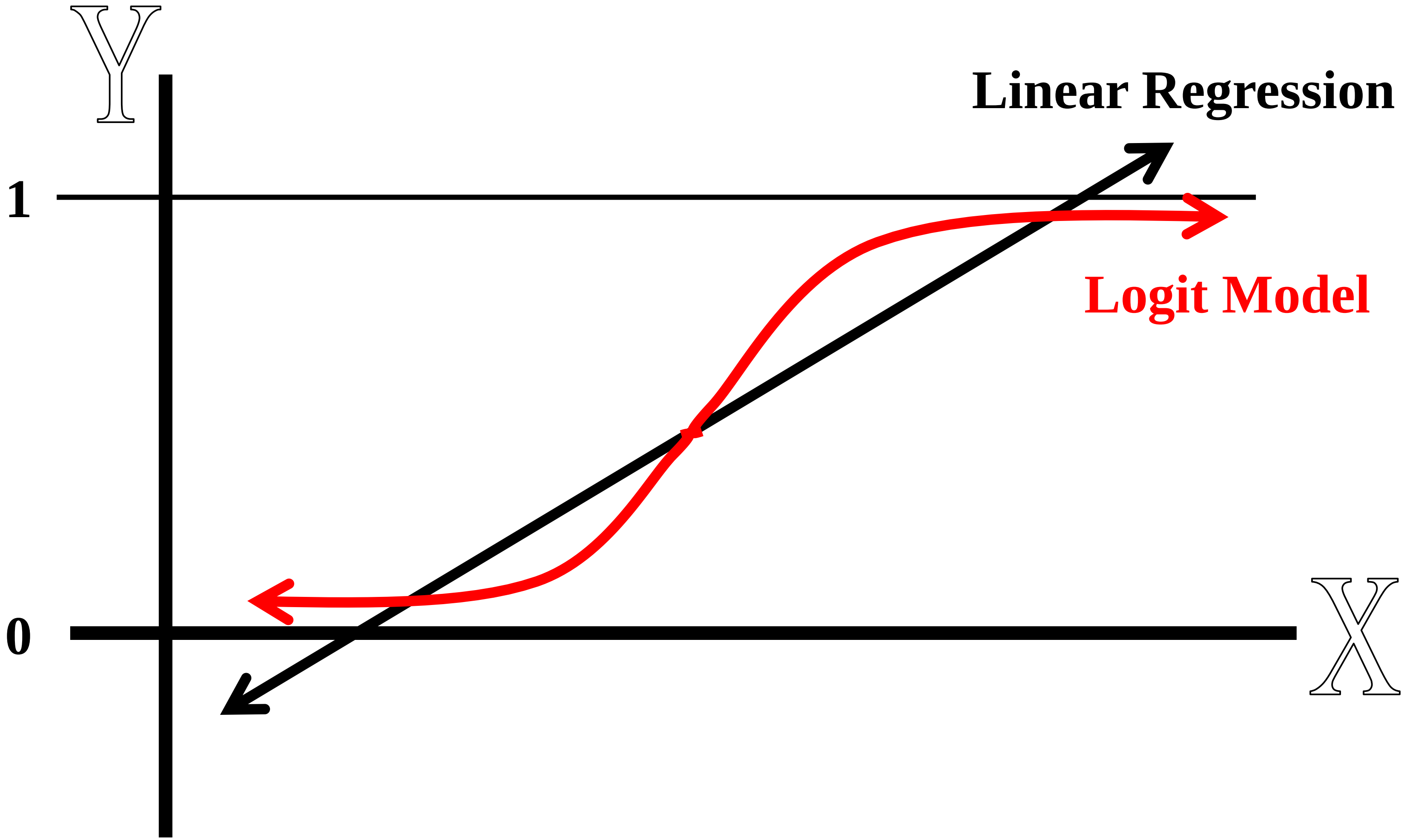
- The "logit" model solves the above problem:

$$\log\left(\frac{p}{1-p}\right) = w_0 + \sum_{j=1}^{d} w_j x_j + u$$

$$= \tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{X}}$$

- p is the probability that the event y occurs, $p(y=1 \mid \mathbf{X})$
- $p/(1-p)$ is the odds ratio (*e.g.*, odds of disease)
- $\log[p/(1-p)]$ is the log odds ratio, or "logit"

Logistic Regression

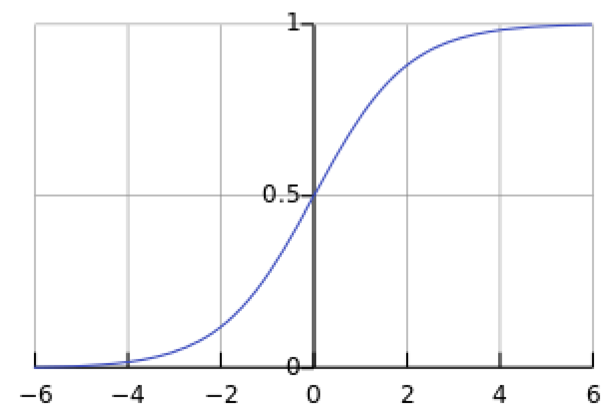


Logistic Regression

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability $p(y=1|\mathbf{X})$ is:

$$p = \frac{1}{1+e^{-w_0-\sum_{j=1}^{d} w_j x_j}} = \frac{e^{w_0+\sum_{j=1}^{d} w_j x_j}}{1+e^{w_0+\sum_{j=1}^{d} w_j x_j}}$$

$$= \frac{1}{1+e^{-\tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}}}} = \frac{e^{\tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}}}}{1+e^{\tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}}}}$$



- if you let $w_0+\sum_{j=1}^{d} w_j x_j = 0$, then $p = 0.5$
- as $w_0+\sum_{j=1}^{d} w_j x_j$ gets really big, p approaches 1
- as $w_0+\sum_{j=1}^{d} w_j x_j$ gets really small, p approaches 0

Logistic Regression

- The Logistic Regression model will be solved by an **iterative maximum likelihood** procedure.
- This is a computer dependent program that:
  - starts with arbitrary values of the regression coefficients and constructs an initial model for predicting the observed data.
  - then evaluates errors in such prediction and changes the regression coefficients so as make the likelihood of the observed data greater under the new model.
  - repeats until the model converges, meaning the differences between the newest model and the previous model are trivial.
- The idea is that you "find and report as statistics" the parameters that are most likely to have produced your data.

Logistic Regression

- The likelihood function is $\prod_{i=1}^{n}(p_i)^{y_i}(1-p_i)^{1-y_i}$
- We want to maximize the log likelihood:

$$L(\tilde{\mathbf{W}})=\sum_{i=1}^{n}\left(y_i\log p_i+(1-y_i)\log(1-p_i)\right)$$

$$=\sum_{i=1}^{n}\left(y_i\log\frac{p_i}{1-p_i}+\log(1-p_i)\right)$$

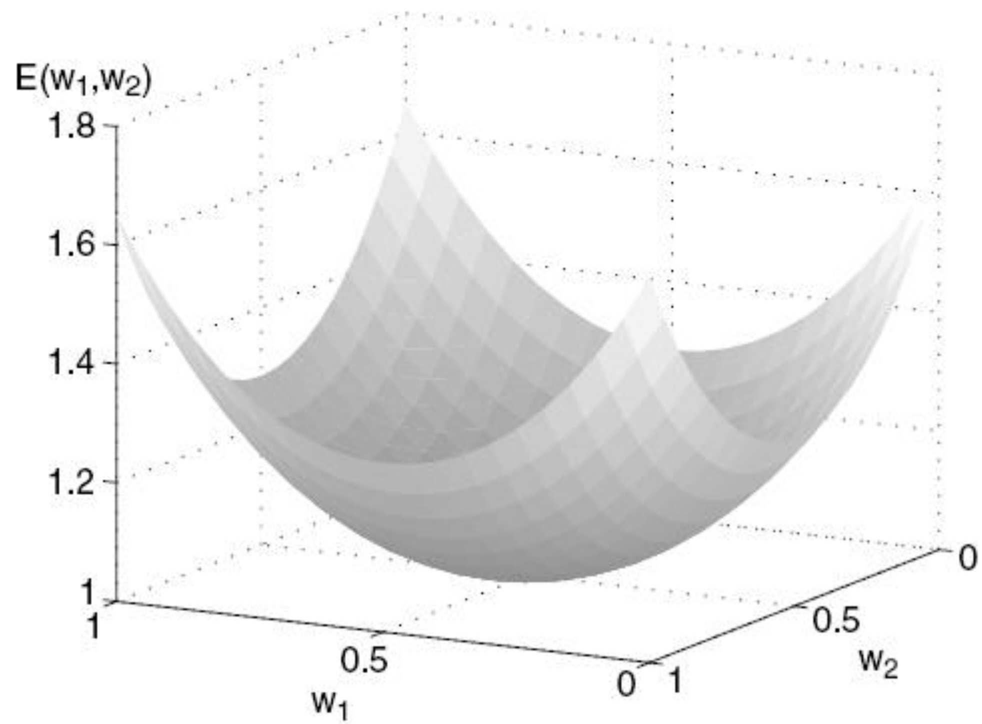
$$=\sum_{i=1}^{n}\left(y_i\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{X}}_i-\log(1+e^{\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{X}}_i})\right)$$

$$\frac{\partial L(\tilde{\mathbf{W}})}{\partial\tilde{\mathbf{W}}}=\sum_{i=1}^{n}\left[\left(y_i-\frac{e^{\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{X}}_i}}{1+e^{\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{X}}_i}}\right)\tilde{\mathbf{X}}_i\right]$$

- It is equal to minimize the cost function

$$C(\tilde{\mathbf{W}})=-L(\tilde{\mathbf{W}})=-\sum_{i=1}^{n}\left(y_i\log p_i+(1-y_i)\log(1-p_i)\right)$$ Cross-entropy

Gradient Decent



Logistic Regression

- Gradient Decent (梯度下降)
  - Calculate the gradient vector
  - Update the weighting in the opposite direction of the gradient vector at each surface point

- Repeat: $\tilde{\mathbf{w}}_{new}^{(j)} = \tilde{\mathbf{w}}^{(j)} - \eta \dfrac{\partial C(\tilde{\mathbf{W}})}{\partial \tilde{\mathbf{W}}^{(j)}}$

$$= \tilde{\mathbf{w}}^{(j)} - \eta \sum_{i=1}^{n} \left[\left(\frac{e^{\tilde{\mathbf{w}}^{\mathrm{T}} \tilde{\mathbf{x}}_i}}{1 + e^{\tilde{\mathbf{w}}^{\mathrm{T}} \tilde{\mathbf{x}}_i}} - y_i \right) \tilde{\mathbf{X}}_i^{(j)} \right]$$

- Until convergence

Gradient Decent