1/3/2021

REGRESSION:

① (parametric) $\widehat{E[Y|X]} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$

② (non-parametric) K-nn $\widehat{E[Y|X]} = Ave(y_i \mid x_i \in N_k(x))$

CLASSIFICATION

K-NEAREST NEIGHBOUR (K-nn)

K positive integer

$(\underline{x}_i, y_i)$, $i = 1, \ldots, n$ training data

$\underline{x}_0$

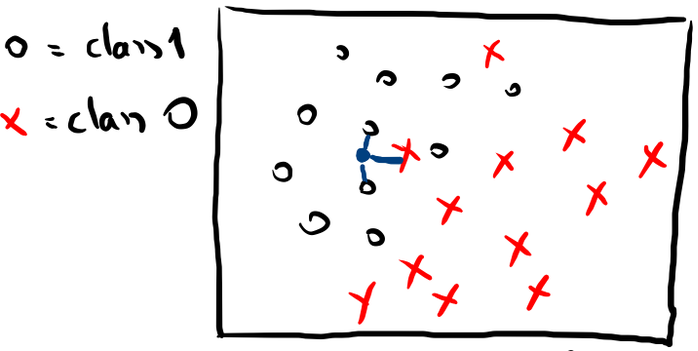
Estimate $P(Y = j \mid X = x_0)$, $j = 1, \ldots, C$ ← number of classes

← according to some metric

① Identify the K training points closest to x_0

② Denote the indices of these points with N_0

$$\hat{P}(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) \quad , j = 1, \ldots, C$$



K = 3 $\quad \hat{P}(Y = 1 \mid X = x_0) = \frac{2}{3}$

$P(Y = 0 \mid X = x_0) = \frac{1}{3}$

CHOICE OF K

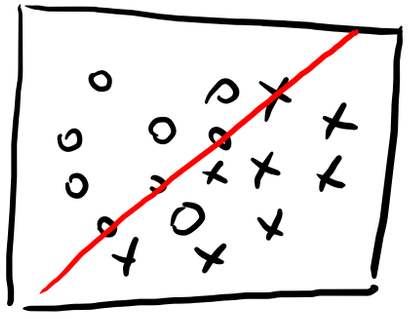
① How complex is this model?

"Effective" number of parameters $\simeq \frac{n}{K}$ ($\frac{n}{K}$ non-overlapping regions)

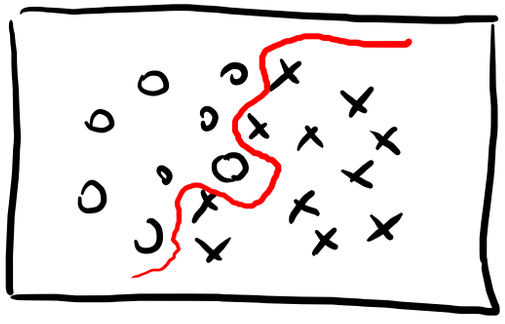
K large → simple model

K small → complex model

② Complexity of decision surface $P(Y=1 \mid X=x) = 0.5$

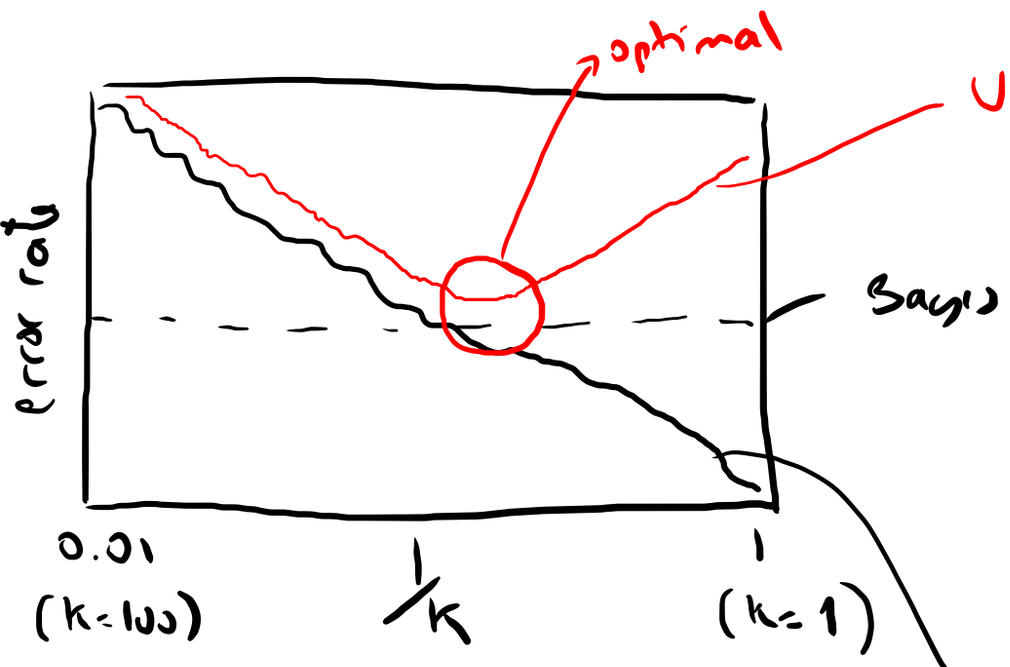


K=100



K=1

Choose K on some test data (unseen, not used for estimating $\hat{P}(Y=j|X)$)



U-curve

Bayes error rate (BER)

↑

only on simulation

(if you know True $p(1|x)$)

error rate on training data

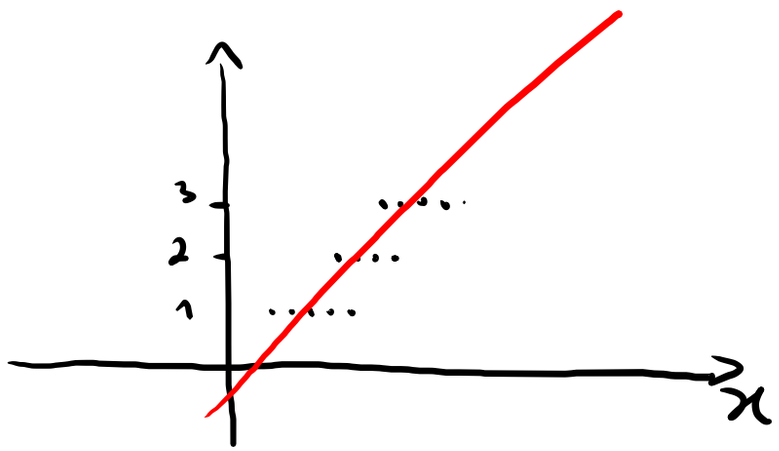
LOGISTIC REGRESSION

Why not linear regression?

Ex 1 Emergency room

$$y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{drug overdose} \\ 3 & \text{epileptic seizure} \end{cases}$$

Predict from symptoms



It could even work but

① No ordering between the categories, but we impose one

② We are imposing equal differences between categories

In most situations, we have no knowledge about ① and ②

Ex 2 Only two classes

Predict only between stroke and drug overdose

$$y = \begin{cases} 0 & \text{if stroke} \\ 1 & \text{if drug overdose} \end{cases} \quad \text{(dummy variable)}$$

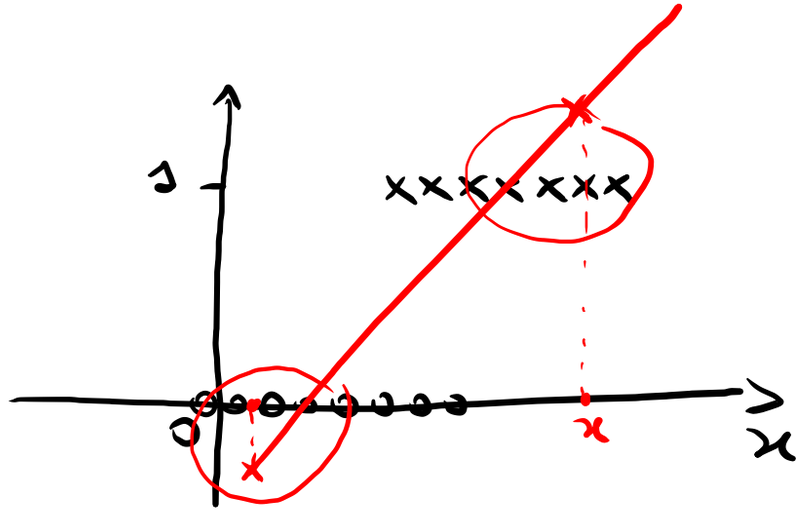
Fit linear regression, get $\hat{y}$ and predict to 1 if $\hat{y} > 0.5$

"$E[\hat{Y}|X]$"

$$E[Y|X] = 0 \cdot p(0|x) + 1 \cdot p(1|x) = p(1|x)$$

As a linear regression model:

$$\underbrace{p(1|x)}_{[0,1]} = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}_{(-\infty, \infty)}$$

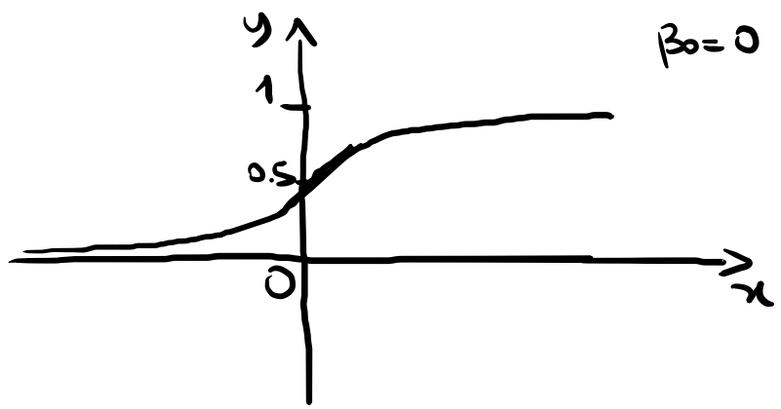


LOGISTIC FUNCTION (sigmoid function)

$$p(1|\underline{x}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} = \frac{e^{\underline{\beta}^t \underline{x}}}{1+e^{\underline{\beta}^t \underline{x}}} \qquad (x_0 = 1)$$

$$p(1|\underline{x}) \in (0,1)$$

$p = 1$ $\quad p(1|x) = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$



$\beta_0 = 0 \quad \beta_1 = 1 \quad \dfrac{e^x}{1+e^x} = y$

ODDS $\quad \dfrac{p(1|\underline{x})}{1-p(1|\underline{x})} = \dfrac{e^{\underline{\beta}^t \underline{x}} / (1+e^{\underline{\beta}^t \underline{x}})}{1 - \dfrac{e^{\underline{\beta}^t \underline{x}}}{1+e^{\underline{\beta}^t \underline{x}}}} =$

$$= \frac{\dfrac{e^{\underline{\beta}^t \underline{x}}}{1+e^{\underline{\beta}^t \underline{x}}}}{\dfrac{1}{1+e^{\underline{\beta}^t \underline{x}}}} = e^{\underline{\beta}^t \underline{x}}$$

LOG-ODDS (or LOGIT)

$$\log\left(\frac{p(1|\underline{x})}{1-p(1|\underline{x})}\right) = \underline{\beta}^t \underline{x} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

LOGISTIC REGRESSION MODEL

INTERPRETATION OF PARAMETERS

β_j As x_j increases by one unit (while everything else remain constant), the log-odds increase/decrease by β_j units

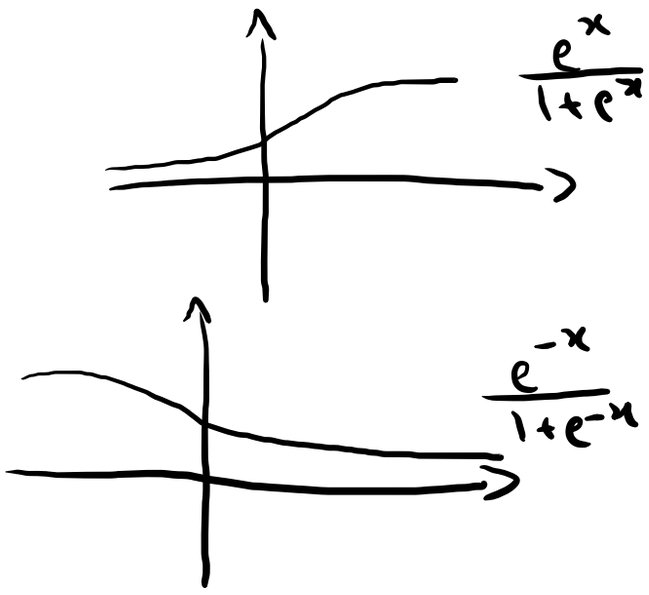
LOG ODDS RATIO

$$\log\left(\frac{p(1|x_0+1)}{1-p(1|x_0+1)}\right) - \log\left(\frac{p(1|x_0)}{1-p(1|x_0)}\right) =$$

$$= \beta_0 + \beta_1(x_0+1) - \beta_0 - \beta_1 x_0 = \beta_1$$

$\beta_j > 0 \rightarrow p(1|x)$ increases if x_j increases

$\beta_j < 0 \rightarrow p(1|x)$ decreases as x_j increases



PARAMETER ESTIMATION

Typically by maximum likelihood

Training data $(\underline{x}_i, y_i)$, $i = 1, \dots, n$

y binary response with a Bernoulli distribution:

$$y \mid X = x = \begin{cases} 1 & p(1|\underline{x}) = p(\underline{x}) \\ 0 & p(0|x) = 1 - p(\underline{x}) \end{cases}$$

LIKELIHOOD

$$L(\underline{\beta}) = \prod_{i=1}^{n} f(\underline{x}_i, y_i; \underline{\beta})$$

$$= \prod_{i=1}^{n} p(\underline{x}_i)^{y_i} (1 - p(\underline{x}_i))^{1-y_i}$$

$y_i = 1 \rightarrow p(\underline{x}_i) = p(1|\underline{x}_i)$

$y_i = 0 \rightarrow 1 - p(\underline{x}_i) = p(0|\underline{x}_i)$

LOG-LIKELIHOOD

$$L(\underline{\beta}) = \prod_{i=1}^{n} p(\underline{x}_i)^{y_i} (1-p(\underline{x}_i))^{1-y_i}$$

$$\ell(\underline{\beta}) = \log(L(\underline{\beta})) = \sum_{i=1}^{n} \left[y_i \log(p(\underline{x}_i)) + (1-y_i) \log(1-p(\underline{x}_i)) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \log p(\underline{x}_i) + \log(1-p(\underline{x}_i)) - y_i \log(1-p(\underline{x}_i)) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \log \frac{p(\underline{x}_i)}{1-p(\underline{x}_i)} + \log(1-p(\underline{x}_i)) \right]$$

$\frac{p(\underline{x}_i)}{1-p(\underline{x}_i)}$: $p(1|\underline{x}_i)$

$$\log\left(1 - \frac{e^{\underline{\beta}^t \underline{x}_i}}{1+e^{\underline{\beta}^t \underline{x}_i}}\right) = \log\left(\frac{1}{1+e^{\underline{\beta}^t \underline{x}_i}}\right) = -\log(1+e^{\underline{\beta}^t \underline{x}_i})$$

$$= \sum_{i=1}^{n} \left[y_i \underline{\beta}^t \underline{x}_i - \log(1+e^{\underline{\beta}^t \underline{x}_i}) \right]$$

$\underline{\beta}^t \underline{x}_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \dots + \beta_j x_{ij} + \dots + \beta_p x_{ip}$

FUNCTION OF $\beta_0, \beta_1, \dots, \beta_p$

MAXIMUM LOG-LIKELIHOOD

$$\frac{\partial \ell(\underline{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i x_{ij} - \frac{e^{\underline{\beta}^t \underline{x}_i}}{1+e^{\underline{\beta}^t \underline{x}_i}} x_{ij} \right] = \sum_{i=1}^{n} \left[y_i x_{ij} - p(\underline{x}_i) x_{ij} \right] \qquad j = 0, 1, \dots, p$$

$$= 0$$

$p+1$ equations in $p+1$ parameters

Numerical methods such as Newton-Raphson are used to solve these equations, returning $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

Using these estimates, we calculate probabilities for any $\underline{x}$ value:

$$\hat{p}(1|\underline{x}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p}}$$

- > 0.5 classify $\underline{x}$ to class 1
- < 0.5 classify $\underline{x}$ to class 0

Or plot a ROC curve on a test data, etc etc

A SIMPLE EXAMPLE where $\hat{p}(1|x)$ can be obtained by hand:

y binary response

X binary

So data can be summarised in a table:

x \ y	0	1
0	a	b
1	c	d

with $a, b, c, d \neq 0$

X \ Y	0	1
0	a	b
1	c	d

$$\log\left(\frac{P(Y=1 \mid X=1)}{1-P(Y=1 \mid X=1)}\right) = \beta_0 + \beta_1 \cdot 1$$

$$\log\left(\frac{P(Y=1 \mid X=0)}{1-P(Y=1 \mid X=0)}\right) = \beta_0$$

So paramiters can be easily estimated by :

$$\hat{\beta}_0 = \log\left(\frac{\hat{P}(Y=1 \mid X=0)}{\hat{P}(Y=0 \mid X=0)}\right) = \log\left(\frac{\frac{b}{a+b}}{\frac{a}{a+b}}\right) = \log\left(\frac{b}{a}\right)$$

$$\hat{\beta}_1 = \log\left(\frac{\hat{P}(Y=1 \mid X=1)}{1-\hat{P}(Y=1 \mid X=1)}\right) - \hat{\beta}_0 = \log\left(\frac{d}{c}\right) - \log\left(\frac{b}{a}\right) =$$

$$= \log\left(\frac{a \cdot d}{b \cdot c}\right)$$

Final comments

① More than 2 classes?

Could fit a model for $p(1|x)$ (versus all the other classes) and another for $p(2|x)$.

Then $p(3|x) = 1 - p(1|x) - p(2|x)$

But logistic regression is not as suited to the case of more than 2 classes as other methods (LDA).

② Predictors can take any form: continuous, categorical, ordinal, etc...

③ Although the model is linear (in the parameters), it can accommodate rather complex structures, such as "polynomial" terms ($X_j^2, X_j^3, \ldots$) and interaction effects ($X_j \cdot X_k$, $X_j^2 \cdot X_k, \ldots$)