Introduction to Classification

DS 301

Iowa State University

Classification

Classification

- Goal: carry out classification of a response Y using predictors $X_1, \ldots, X_p$.
- Main difference with the regression models we've covered so far is that now Y is qualitative (categorical).
- Still in the supervised learning setting. Classification is NOT clustering!!

Example of classification problems:

- Email is spam or not spam?
- Is this transaction a fraud or not fraud?
- Does an individual have a disease or not a disease? Is this DNA mutation harmful or not?
- Is this image of a dog or not a dog (image classification)?
- ....

Example

Suppose we are trying to classify the medical condition of ER patients:

$$Y = \begin{cases} 1 & \text{if stroke} \\ 0 & \text{if heart attack} \end{cases}$$

Why don't we use a least squares model here to predict Y? $(\hat{Y})$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p.$$

Here, $\hat{Y}$ is an estimate of $P(\text{stroke}|X)$.

$P(Y=1|X)$

$P(Y=0|X) =$
$1 - P(Y=1|X).$

$\Rightarrow$ estimates $\hat{Y}$ might not be in the interval $[0,1]$.

$\Rightarrow$ this setting cannot accommodate more than 2 classes.

Instead we need to use classification techniques that are specifically designed to handle this kind of problem:

1. Logistic regression
2. Linear/quadratic discriminant analysis (LDA/QDA)
3. k-nearest neighbor (non-parametric)
4. SVM

Gold standard for classification is **Bayes Rule.**

Bayes Rule

$$\begin{pmatrix} 1 & X_1 & X_2 & X_3 & \cdots \\ 0 & & & & \\ 1 & & & & \\ 0 & & & & \end{pmatrix}$$

Let's start with a Y that only takes 2 values:

$$Y = \begin{cases} 1 & \text{if default on credit} \\ 0 & \text{if do not default on credit} \end{cases}$$

Target: $P(Y = 1|X_1, \ldots, X_p)$.

If I knew this probability, I could use Bayes Rule directly:

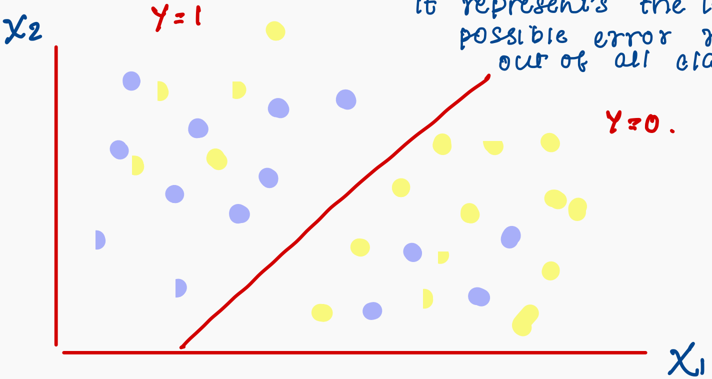
$$\begin{cases} P(Y = 1|X) > 0.5 & \text{then } \hat{Y} = 1 \\ P(Y = 1|X) \leq 0.5 & \text{then } \hat{Y} = 0 \end{cases}$$

Bayes Rule

There is still error associated with Bayes Rule because data is hardly ever perfectly separated. → This error is called Bayes error.

(well. seperated)

It represents the lowest possible error rate out of all classifiers.



Logistic Regression

Instead of modeling Y directly, let's try to model the **probability** that Y belongs to a category.
Suppose Y only takes 2 values:

$$Y = \begin{cases} 1 & \text{if default on credit} \\ 0 & \text{if do not default on credit} \end{cases}$$

Target: $P(Y = 1|X_1, \ldots, X_p) = p(x)$.

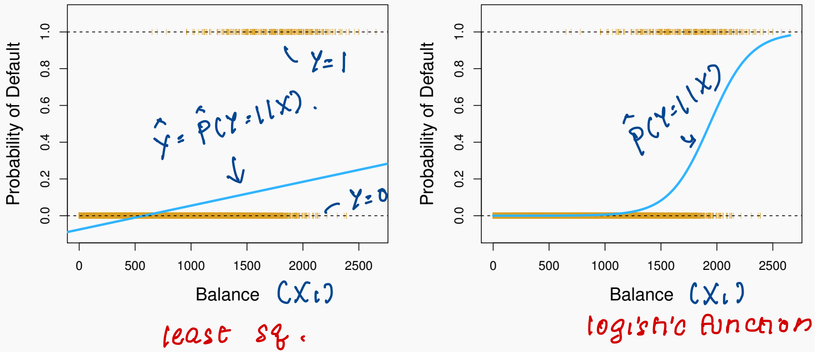
- $p(x)$ is usually unknown to us, so we need to estimate it from the data.
- We must model $p(x)$ using a function that can guarantee outputs fall between 0 and 1.

Logistic Regression

In logistic regression, we use the logistic function:

$$P(Y=1|X) = p(x) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)}$$

$= exp(B_0 + B_1 X_1) / 1 + exp(B_0 + B_1 X_1)$ if only X_1.



$P(Y=1 \mid X_1, X_2, \ldots, X_p)$.

$\downarrow$

$$p(x) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)} \leftarrow$$

$\rightarrow$ estimate $\hat{p}(x)$ from our data.

$$\log\Big(\underbrace{\frac{p(x)}{1-p(x)}}_{\text{odds.}}\Big) = \underbrace{\log(\text{odds})}_{\text{logit}} = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_p X_p$$

- logit is linear with respect to $X_1 \ldots X_p$.

$$\log(\text{odds}) = \underset{\downarrow}{B_0} + \underset{\downarrow}{B_1} X_1 + \ldots \underset{\downarrow}{B_p} X_p.$$

parameters (unknown).

we need to estimate them (training set)

Estimating β's

$\beta_0, \beta_1, \ldots, \beta_p$ are parameters. These are unknown quantities and we need to estimate them from our data $\Rightarrow$ Maximum likelihood estimation.

- Intuition: We seek estimate for $\beta_0, \beta_1, \ldots, \beta_p$ such that the predicted probability $\hat{p}(x) = \hat{P}(Y = 1|X)$ is large for observations where $Y = 1$ and small for observations where $Y = 0$.
- Following our example, that means
  - If an individual defaults on credit ($Y = 1$), then we want $\hat{p}(x)$ to be close to 1.
  - If an individual does not default on credit ($Y = 0$), then we want $\hat{p}(x)$ to be close to 0.

Maximum likelihood estimation technical details (sort of)

Intuition can be formalized using a likelihood function: probability of seeing your data given set of parameters.

$$\underbrace{l(\beta_0, \beta_1, \ldots, \beta_p)}_{\text{likelihood function}} = P(Y_1 \mid X, \beta_0, \beta_1, \ldots \beta_p) \times P(Y_2 \mid X, \beta_0, \beta_1, \ldots, \beta_p) \times P(Y_3 \mid X, \beta_0, \ldots \beta_p) \times \cdots P(Y_n \mid X, \beta_0, \ldots, \beta_p).$$

goal: maximize this

$$= \prod_{i=1}^{n} P(Y_i \mid X, \beta_0, \ldots \beta_p) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

$$p(x_i) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p)}{1 + \exp(\beta_0 + \cdots \beta_p x_p)}$$

$\Rightarrow \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are chosen that maximize this likelihood

(mle : max likelihood estimates).

Maximum likelihood estimation

No closed form analytical solution to
logistic regression model.

↳ we can solve for it numerically.
↳ optimization to solve.

✗ (Newton's method
⋮
gradient descent).

R solve this problem (mles)

See R script: `logit_intro.R`