

Logistic Regression and Classifications

Shaobo Li

University of Kansas

From Continuous to Categorical Outcome

$$f(\text{image of a white dog}) \rightarrow \text{dog}$$

$$f(\text{image of an orange cat}) \rightarrow \text{cat}$$

- Response Y : discrete value
 - e.g., $Y = \{\text{dog}, \text{cat}\}$
 - or $Y = \{0, 1\}$, 1 - dog; 0 - not dog

- K-Nearest Neighbor
- Logistic regression
- Classification tree
- Random forest
- Boosted tree
- Support vector machine
- Neural networks
- Deep learning
- ...
- Is clustering a classification model?

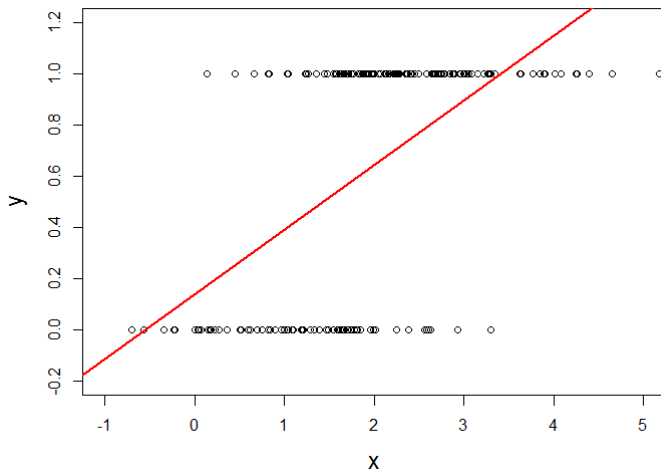
Why Not Linear Regression

- Example: default prediction
 - Default ($Y = 1$) vs. Nondefault ($Y = 0$)
 - X_1 : credit card balance level, X_2 : income level
- Suppose the estimated linear regression is

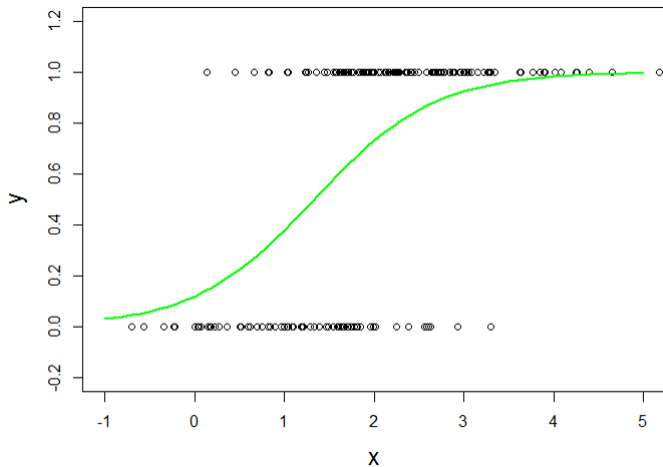
$$\hat{Y} = -1.5 + 2X_1 - X_2$$

- What is the predicted value if a person's balance level is 1 and income level is 3?
- How to interpret this value?

An Illustration



An Illustration



- Denote $\mathcal{C}(X)$ as a classifier
- Most DM algorithms produce probabilistic outcome
 - e.g. probability that X belongs to each class
- Classification is based on certain decision rules
- Example: The model prediction tells that the probability of default is 0.2, then

Threshold	<0.1	>0.1
Class	Nondefault	Default

- For binary response:

$$\mathbb{P}(y_i = 1 | \mathbf{x}_i) = \frac{1}{1 + \exp(-\beta^T \mathbf{x}_i)}$$

- Sigmoid function: $s(u) = \frac{1}{1 + e^{-u}}$
- Interpretation: **probability** of event conditional on X
- More than two classes: *multinomial logistic model*
- Can you re-write the model such that the right-hand side is the linear predictor?

Odds and Interpretation of β

- Let $P = \mathbb{P}(Y = 1|\mathbf{X})$, then

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

- The probability ratio $\frac{P}{1-P}$ is called odds, a function of X
- Logistic model is also called log-odds model

- Interpretation of β_1

- with 1 unit increase on X_1 , the log odds changes by β_1
- Can we say the odds changes by e^{β_1} ?

- By simple algebra, given all X 's are fixed except for X_j

$$\beta_j = \log\left(\frac{\text{Odds}(X_j + 1)}{\text{Odds}(X_j)}\right)$$

this is *log of odds ratio*.

Odds and Interpretation of β

Exercise: Suppose we are interested in predicting corporate bankruptcy using the logistic model. The parameter estimate for Earning is -1.29. This number means:
Holding all other predictors fixed, for every one-unit increase in earning,

- a) the probability of bankruptcy decreases 1.29%.
- b) the odds of bankruptcy decreases 1.29%.
- c) the odds of bankruptcy changes by $(e^{\beta_1} - 1) * 100\%$.
- d) the odds of bankruptcy decreases 1.29.
- e) the log of odds of bankruptcy decreases by 1.29.
- f) the odds of bankruptcy is 1.29 times lower.
- g) the odds of bankruptcy is $e^{1.29}$ times lower.

Multinomial Logit Model

- Response $Y = 1, 2, \dots, K$, K classes
- Given predictors \mathbf{x}_i

$$\log \left(\frac{\mathbb{P}(y_i = 2)}{\mathbb{P}(Y_i = 1)} \right) = \beta_2^T \mathbf{x}_i$$

$$\log \left(\frac{\mathbb{P}(y_i = 3)}{\mathbb{P}(Y_i = 1)} \right) = \beta_3^T \mathbf{x}_i$$

\vdots

$$\log \left(\frac{\mathbb{P}(y_i = K)}{\mathbb{P}(Y_i = 1)} \right) = \beta_K^T \mathbf{x}_i$$

- The first class “1” is the reference
- There are $(K - 1) \times (p + 1)$ coefficients need to be estimated.

Loss Function (in machine learning)

- Recall OLS

$$L_{OLS}(\beta) = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

- For logistic regression, we have a different loss

$$L_{logit}(\beta) = \sum_{i=1}^n -2 [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

where $p_i = \frac{1}{1 + \exp(-\beta_0 - \sum_{j=1}^p \beta_j x_{ij})}$

Maximum Likelihood Estimation (in statistics)

- $y_i = \begin{cases} 1 & \text{with Prob. } p_i \\ 0 & \text{with Prob. } 1 - p_i \end{cases}, p_i = \frac{1}{1 + \exp(-\beta_0 - \sum_{j=1}^p \beta_j x_{ij})}$

- Likelihood function of i th observation $y_i | \mathbf{x}_i$:

$$Likelihood_i = p_i^{y_i} \times (1 - p_i)^{1-y_i}$$

- Likelihood function of all observations, $i = 1, \dots, n$:

$$Likelihood = \prod_{i=1}^n p_i^{y_i} \times (1 - p_i)^{1-y_i}$$

- log-likelihood for all observations:

$$\log Lik(\beta) = \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

How to solve for β ?

- Unlike OLS, there is no analytical solution for logit model
- Numeric solution (below is a univariate example)
 - Gradient descent

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} - \alpha * L'(\hat{\beta}^{(n)})$$

where α is called learning rate.

- Newton's method (a very good [tutorial](#))

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} - \frac{L'(\hat{\beta}^{(n)})}{L''(\hat{\beta}^{(n)})}$$

- We call derivative for univariate: $f'(x)$, $f''(x)$
- We call gradient for multivariate: $\nabla f(x)$, $\nabla^2 f(x)$

Prediction — From Probability to Class

- Direct outcome of model: probability
- Next step: classification
- Need decision rule (cut-off probability – p-cut)
- Not unique

Confusion Matrix

- Classification table based on a specific cut-off probability
- Used for model assessment

	Pred=1	Pred=0
True=1	True Positive (TP)	False Negative (FN)
True=0	False Positive (FP)	True Negative (TN)

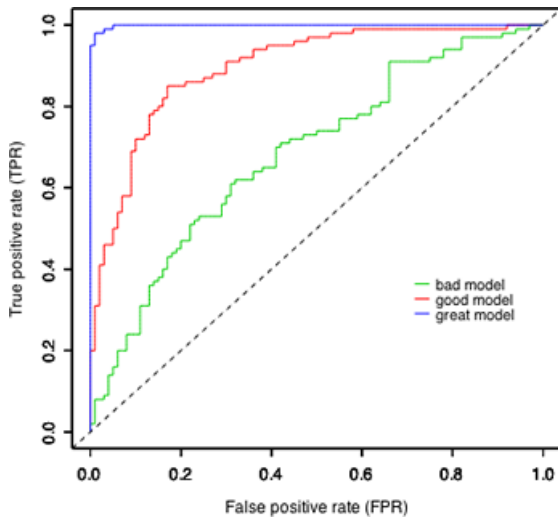
- FP: type I error; FN: type II error
- Different p-cut results in different confusion matrix
- Try to understand this table instead of memorizing!

Some Useful Measures

- Misclassification rate (MR) = $\frac{FP+FN}{Total}$
- True positive rate (TPR) = $\frac{TP}{TP+FN}$: Sensitivity or Recall
- True negative rate (TNR) = $\frac{TN}{FP+TN}$: Specificity
- False positive rate (FPR) = $\frac{FP}{FP+TN}$: $1 - \text{Specificity}$
- False negative rate (FNR) = $\frac{FN}{TP+FN}$: $1 - \text{Sensitivity}$

- Receiver Operating Characteristic
- Plot of FPR (X) against TPR (Y) at various p-cut values
- Overall model assessment (not for a particular decision rule)
- Unique for a given model
- Area under the curve (AUC): a measure of goodness-of-fit

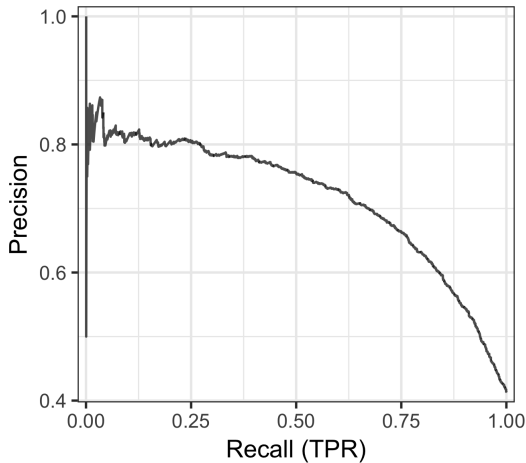
ROC Curve



Precision and Recall

- More accurate measure for imbalanced data
- Widely used in document retrieval (e.g., spam classification)
- Precision = $\frac{TP}{TP+FP}$
 - fraction of retrieved instances that are relevant
- Recall = $\frac{TP}{TP+FN}$, which is the same as TPR
 - fraction of relevant instances that are retrieved
- This pair of measures emphasizes the positive cases
- There is a trade-off between precision and recall
- F1-score: $F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$. (harmonic mean)
 - a popular measure to balance the tradeoff.

Precision-Recall Curve



Exercise

The following R code demonstrate the idea of measuring the overall discriminant power of a model (predicted probability). Complete the code and draw ROC and PR curve. In addition, find the the optimal cut-off with respect to accuracy rate and F1-score.

```
1 set.seed(750)
2 n <- 200
3 p <- runif(n)
4 y <- rbinom(n,1,p)
5 head(cbind(y,p))
6
7 pcut <- seq(0.1,0.9,0.05)
8 v1<- NULL
9 for(i in 1:length(pcut)){
10   yhat <- (p>pcut[i])*1
11   confMat <- table(y,yhat,dnn = c("True", "Pred"))
12   TPR <- sum(y==1 & yhat==1)/sum(y==1)
13   TNR <-
14   FPR <-
15   FNR <-
16   Precis <-
17   v1 <- rbind(v1, c(TPR=TPR, TNR=TNR, FPR=FPR, FNR=FNR, Precis=Precis))
18   cat("pcut:", pcut[i], "TPR:", TPR, "TNR:", TNR, "FPR:", FPR, "FNR:", FNR, "Precision", Precis, '\n')
19   print(confMat)
20 }
21
```

Download the code [here](#).

Asymmetric Cost

- Example: compare following two confusion matrices based on two p-cut values

	Pred=1	Pred=0
True=1	10	40
True=0	10	440

	Pred=1	Pred=0
True=1	40	10
True=0	130	320

- Which one is better? In terms of what?
- What if this is about loan application
 - $Y = 1$: default customer
 - Default will cost much more than reject a loan application

Choice of Decision Threshold (p-cut)

- Do NOT simply use 0.5!
- In general, we use **grid search** method to optimize a measure of classification accuracy/loss
 - Cost function (symmetric or asymmetric) based on FP and FN
 - F1-score based on precision and recall

- Based on Bayes theorem:

$$\mathbb{P}(Y = k|X = x) = \frac{\mathbb{P}(X = x|Y = k) \times \mathbb{P}(Y = k)}{\mathbb{P}(X = x)}$$

- Discriminant analysis

$$\mathbb{P}(Y = k|X = x) = \frac{f_k(x) \times \pi_k}{\sum_{l=1}^K f_l(x) \times \pi_l}$$

- $f_k(x)$ is the assumed density function of X in class k
- π_k can be simply calculated as the fraction of $Y = k$

Linear Discriminant Function

- Given x , find the k such that $f_k(x) \times \pi_k$ is largest
- Therefore, only $f_k(x) \times \pi_k$ is of interest
- We assume $f_k(x)$ to be Gaussian density

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right)$$

- By taking log and discard terms without k , we have

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \log(\pi_k)$$

- This is called *linear discriminant score function*

Comparison Between Logistic Model and LDA

- Logistic regression is a very popular classifier especially for binary classification problem
- LDA is often used when n is small and classes are well separated, and Gaussian assumption is reasonable. Also when $K > 2$.
- Both are linear methods