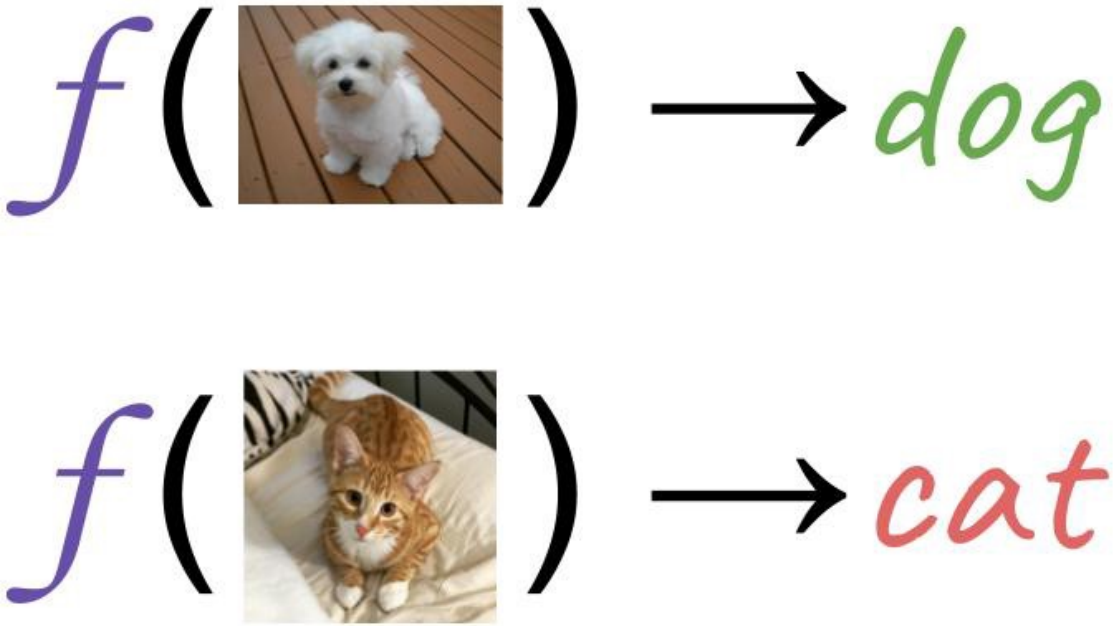
Logistic Regression and Classifications

Shaobo Li

University of Kansas

From Continuous to Categorical Outcome



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

Classification Methods

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

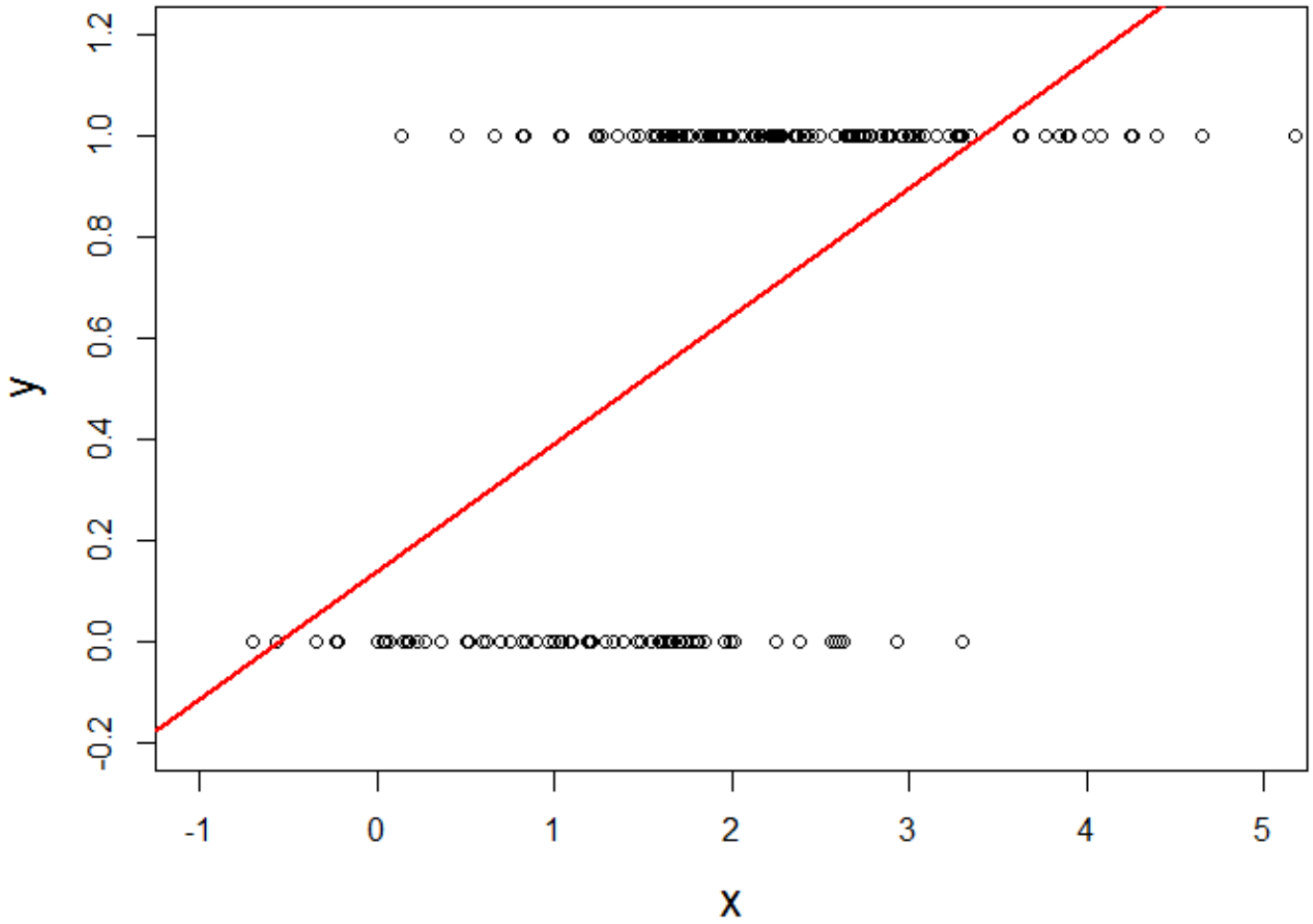
Suppose the estimated linear regression is

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

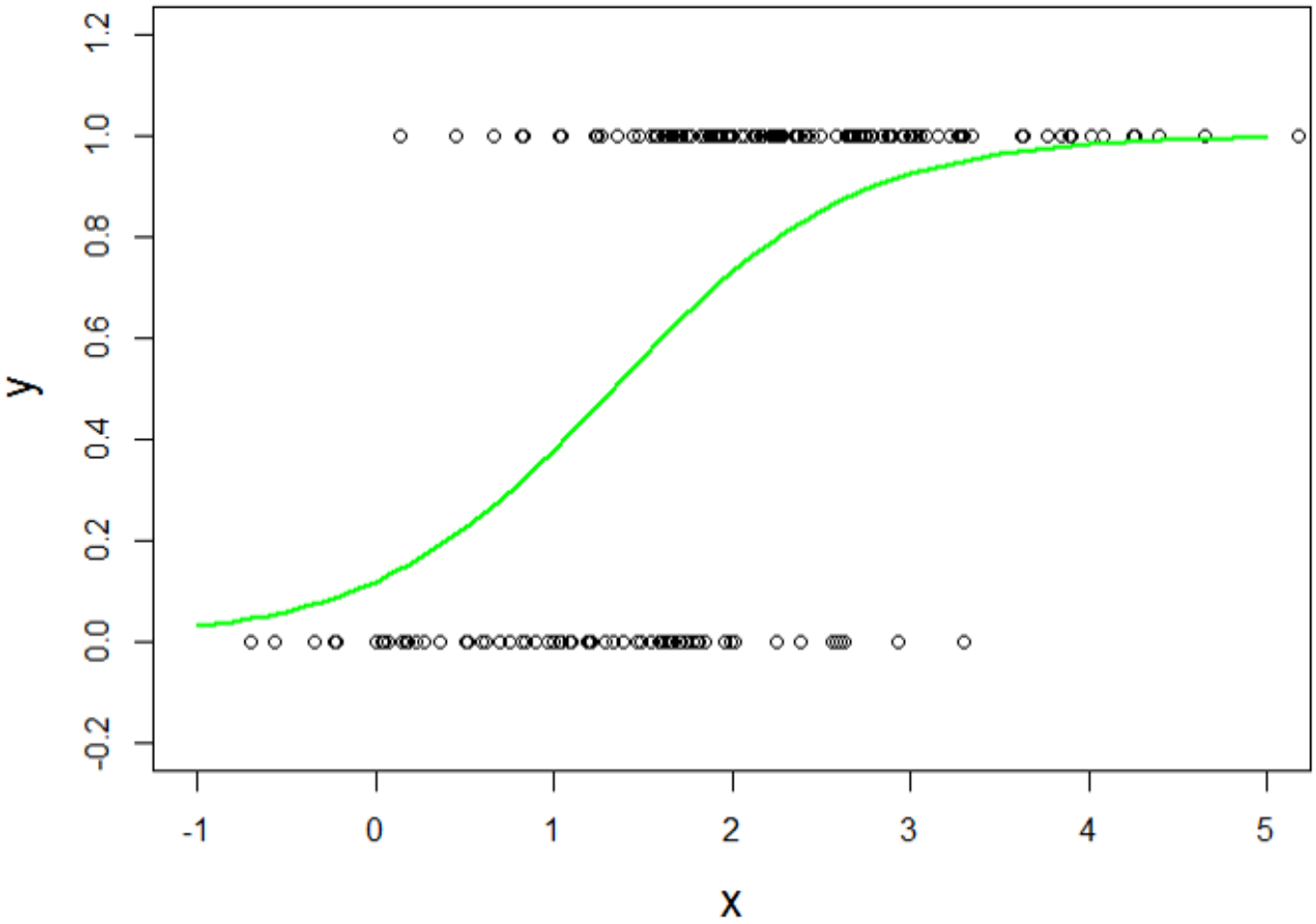
What is the predicted value if a person's balance level is 1 and income level is 3?

How to interpret this value?

An Illustration



An Illustration



Classification

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

Logistic Regression

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

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

- 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 $Y = 1, 2, ..., K$, 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

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

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 ith observation $y_i|\mathbf{x}_i$:

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

Likelihood function of all observations, $i = 1, ..., n$:

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

log-likelihood for all observations:

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

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) $= \frac{\text{FP}+\text{FN}}{\text{Total}}$

True positive rate (TPR) $= \frac{\text{TP}}{\text{TP}+\text{FN}}$: Sensitivity or Recall

True negative rate (TNR) $= \frac{\text{TN}}{\text{FP}+\text{TN}}$: Specificity

False positive rate (FPR) $= \frac{\text{FP}}{\text{FP}+\text{TN}}$: 1 – Specificity

True negative rate (FNR) $= \frac{\text{FN}}{\text{TP}+\text{FN}}$: 1 – Sensitivity

ROC Curve

Receiver Operating Characteristic

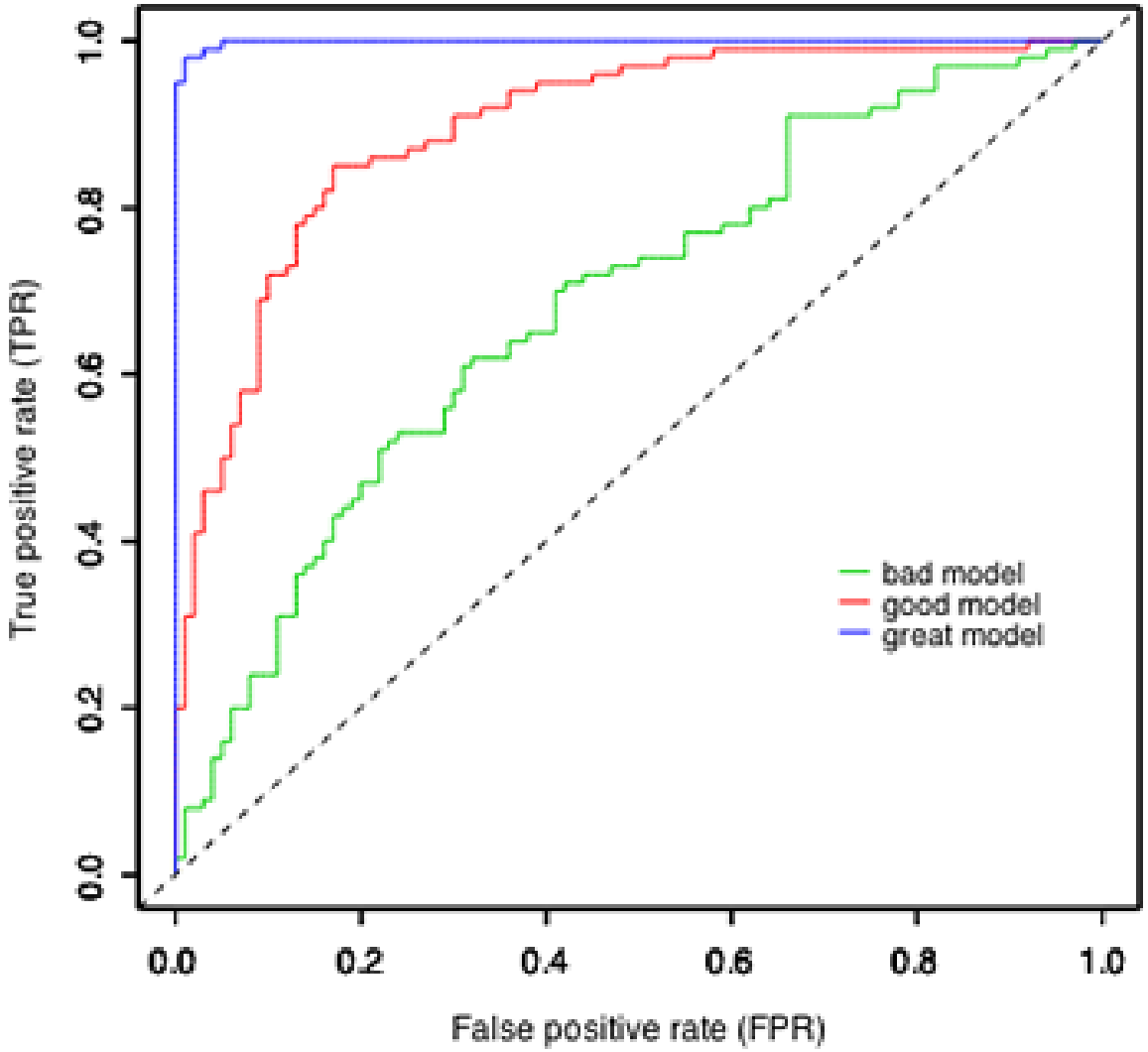
Plot of FPR (X) against TPR (Y) at various p-cut values

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


True positive rate (TPR)
False positive rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
bad model
good model
great model

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)

Grid search with cross-validation