Workspace 'Workspace' in 'lec8'

Page 1 (row 1, column 1)

$\hat{Y}_{sale} = 6.7502 + 0.0191 X_{TV} + 0.0289 X_{radio}$

$+ 0.0011 (X_{TV} \cdot X_{radio})$

$Y_{GPA} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

$X_1 = \begin{cases} 1 & NC \\ 0 & o.w. \end{cases}$

$X_2 = \begin{cases} 1 & Ohio \\ 0 & o.w \end{cases}$

$1 - R^2 = 27\%$

$I(Y=1) = \begin{cases} 1 & \text{if } Y=1 \\ 0 & \text{if } Y \neq 1 \end{cases}$

$Y | X=x \sim Bin(p)$

$Y = 1$ with prob p given $X=x$

$= 0$ with prob $1-p$ given $X=x$

prove $E(I(Y=1)) = ?$

$= 1 \cdot P(Y=1) + I(0=1) \cdot P(Y=0)$

$= P(Y=1) + 0 \cdot (1-p) = p$

$\rightarrow \boxed{I(X \in A)} = \begin{cases} 1 & \text{if } X \in A \\ 0 & \text{if } X \notin A \end{cases}$

$E(I(X \in A)) \overset{?}{=} 1 \cdot P(X \in A) + 0 \cdot P(X \notin A)$

$= P(X \in A)$

Module 4: Classification

Lecture 8
Jan 27th, 2023

Recap

MICHIGAN STATE UNIVERSITY

- Qualitative Predictors (One-hot encoding)
- Interactions (Hierarchy Principle)
- Beyond linearity

Y

$X \quad X^2 \quad X^3$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

C

NC	1	0	0
Oh.o	0	1	0
MI	0	0	1

X

$$\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 (X_1 \cdot X_2 X_3)$$

$$\beta_4 X_1 X_2 + \beta_6 X_2 X_3 + \beta_7 X_3 X_1$$

Classification Problems

MICHIGAN STATE UNIVERSITY

Here the response variable Y is qualitative.

Classification Problems

MICHIGAN STATE UNIVERSITY

Here the response variable *Y* is qualitative — e.g. email is one of $\mathcal{C} = (\texttt{spam}, \texttt{ham})$ (ham=good email), digit class is one of $\mathcal{C} = \{\texttt{0}, \texttt{1}, \ldots, \texttt{9}\}$. Our goals are to:

- Build a classifier *C(X)* that assigns a class label to a feature unlabeled observation *X*.
- Assess the uncertainty in each classification.
- Understand the roles of the different predictors among

$$X = (X_1, X_2, \ldots, X_p).$$

$Y = f(X) + \varepsilon$

$C(X) = \begin{cases} \text{spam} \\ \text{ham} \end{cases}$

1 2 3 4 5

1 2 3

"1" "2" "3"

Classification: some details

MICHIGAN STATE UNIVERSITY

How to measure the performance of a classifier in a training dataset Tr?

Training data:
$\{(x_1, y_1), \cdots, (x_n, y_n)\}$ with y_i qualitative

$$\min \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \beta_0 - \beta_1 x_i)$$

$$\cancel{(y_i - C(x_i))^2}$$

Classification: some details

MICHIGAN STATE UNIVERSITY

How to measure the performance of a classifier in a training dataset Tr?

Training data:
$\{(x_1, y_1), \cdots, (x_n, y_n)\}$ with y_i qualitative

Can we define it as $\mathrm{MSE}_{\mathrm{Tr}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{C}(x_i))^2$? ?

Classification: some details

How to measure the performance of a classifier in a training dataset Tr? We use the misclassification error rate:

$$\text{Error}_{\text{Tr}} = \frac{1}{N}\sum_{i\in \text{Tr}} I[y_i \neq \hat{C}(x_i)]$$

where $I[y_i \neq \hat{C}(x_i)]$ is an indicator variable that equals 1 if $y_i \neq \hat{C}(x_i)$ and 0 otherwise

$$I(y_i = \hat{c}(x_i)) = 0$$

$$I(y_i \neq \hat{c}(x_i)) = 1$$

As in the regression setting, we are most interested in the testing errors associated

$$\boxed{\text{Error}_{\text{Te}}} = \frac{1}{M}\sum_{i\in \text{Te}} I[y_i \neq \hat{C}(x_i)]$$

with a testing set $\text{Te} = \{x_i, y_i\}_1^M$:

$\hat{C}$

.

Ideal Classifier

MICHIGAN STATE UNIVERSITY

$f(x) = E(y \mid X = x)$

- Is there an ideal *C(X)*?
- Suppose the *K* elements in $\mathcal{C}$ are numbered 1, 2, …, K. Let

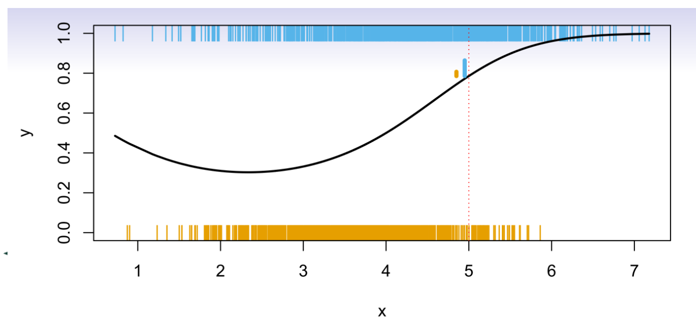
$$p_k(x) = \Pr(Y = k \mid X = x), \; k = 1, 2, \ldots, K.$$

These are the conditional class probabilities at *x*;

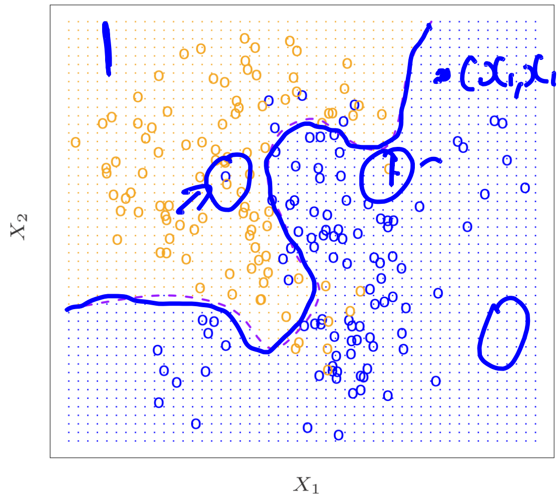
- Then the $\boxed{\textbf{Bayes classifier}}$ at *x* is oracle

$$C(x) = j \text{ if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$

, which is the optimal classifier.



Bayes decision Boundary



$(x_1, x_1) \Rightarrow P(Y=1 \mid X_1 = x_1, X_2 = x_1)$

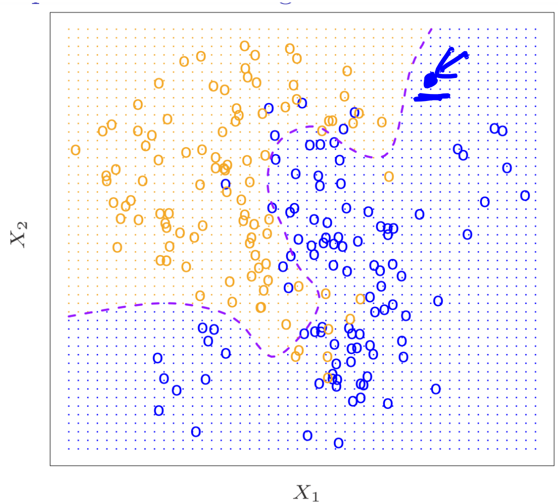
$< P(Y=0 \mid X_1 = x_1, X_2 = x_1)$

- Example where we simulated the data, so we know the probability of each
- The purple line is where we switch our predictor, called the Bayes decision boundary

Bayes Error rate

MICHIGAN STATE UNIVERSITY

$Y = 0$ or 1



$C(x_0) = 0$

$P(Y = 1 \mid X = x_0)$

$C(x_0) = 1$

$P(Y = 0 \mid X = x_0)$

- Example where we simulated the data, so we know the probability of each
- The purple line is where we switch our predictor, called the Bayes decision boundary

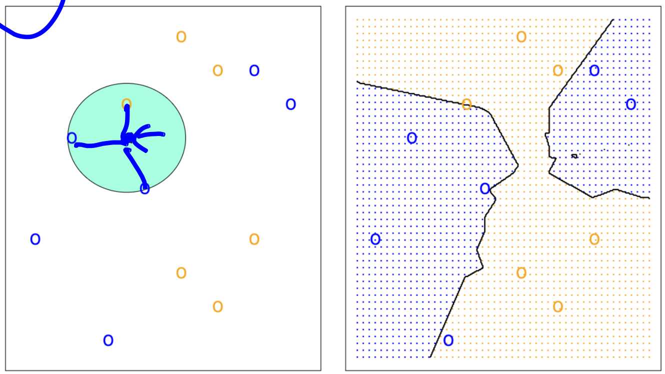
Error at $X = x_0$

$$1 - \max_j \Pr(Y = j \mid X = x_0)$$

K-nearest neighbors Classifier

Idea: Use similar training points when making predictions

k=3



- Estimate conditional proability

$$\Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in N(x_0)} I(y_i = j)$$

- Pick j with highest value

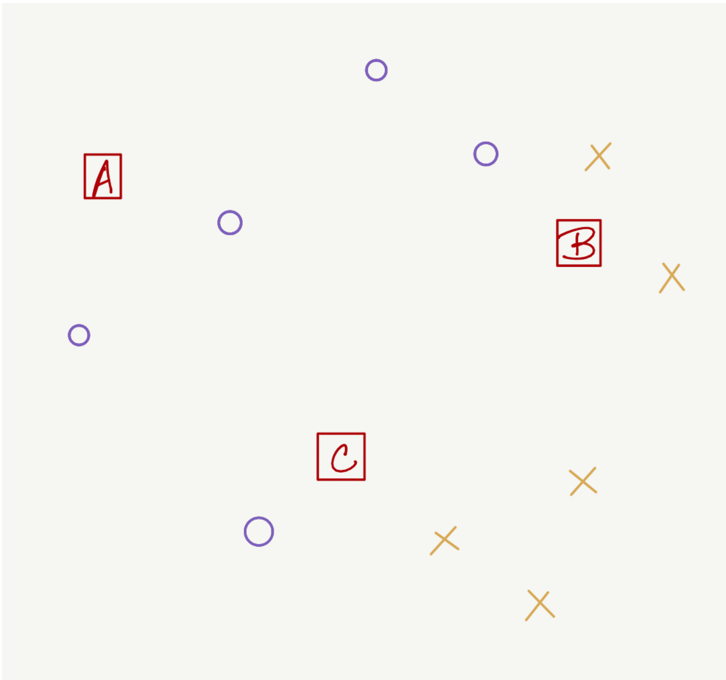
What is/are the parameters?

Is it parametric or non-parametric?

Example

MICHIGAN STATE UNIVERSITY

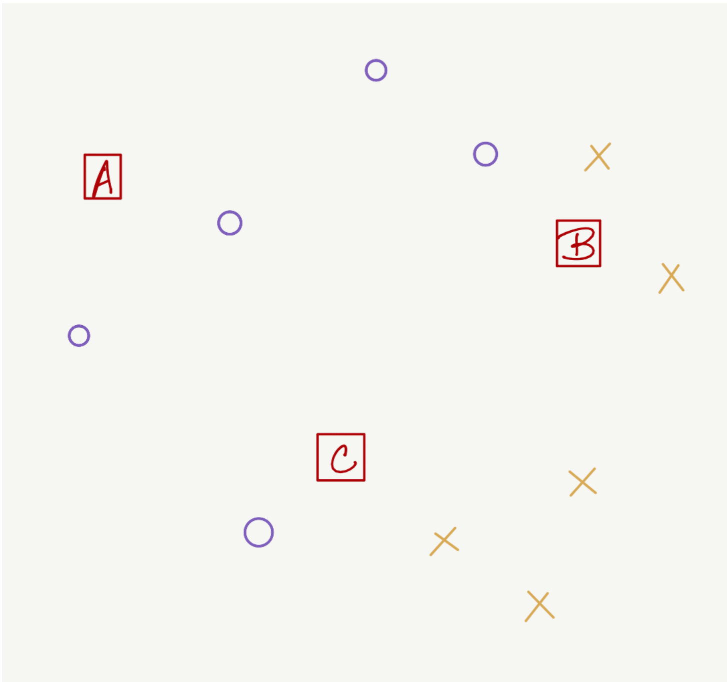
Here label is shown by O vs X. What are the *k*nn predictions for points A, B and C for k = 1
or k = 3?



	k = 1	k = 3
A	O	O
B	X	X
C	O	X ✓

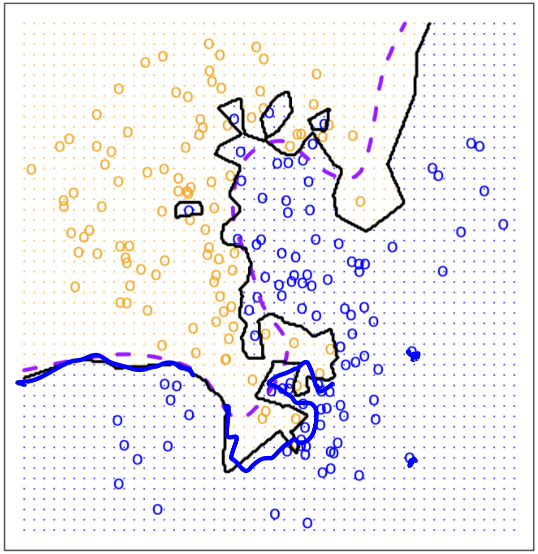
Example

Here label is shown by O vs X. What are the *k*nn predictions for points A, B and C for k = 1
or k = 3?

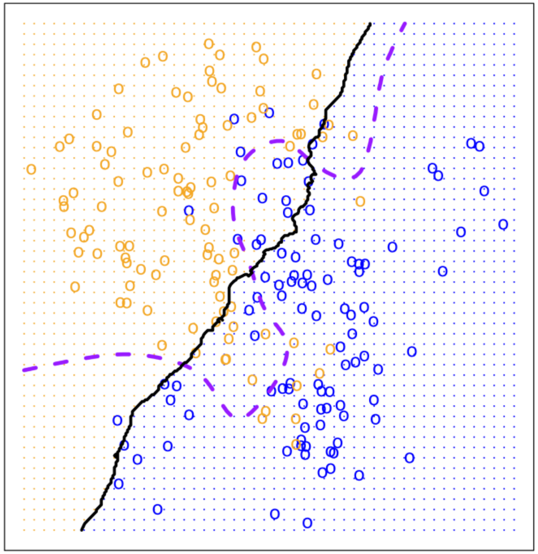


Tradeoff

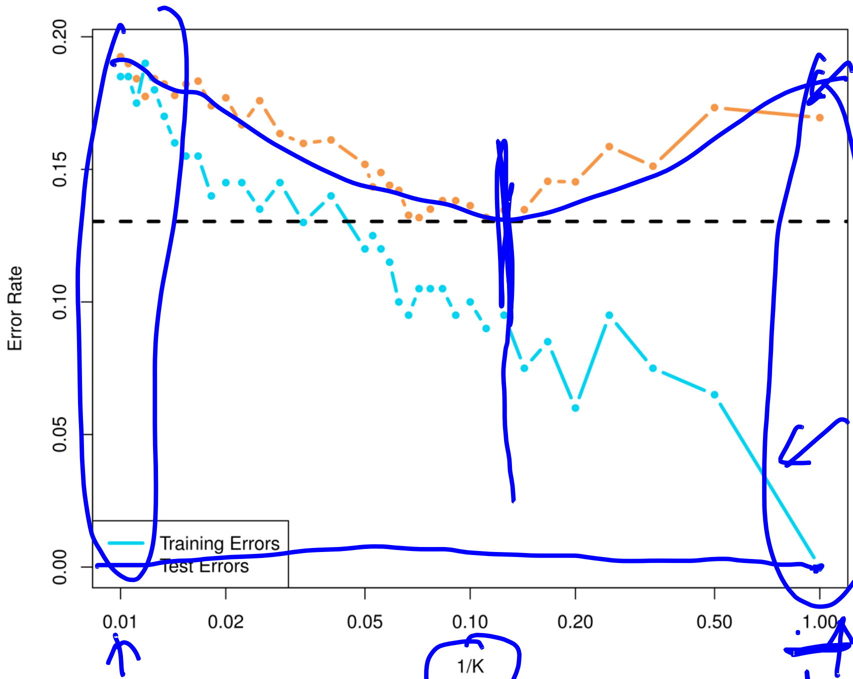
KNN: K=1



KNN: K=100



KNN: Training and Test Errors



k=100

k=1

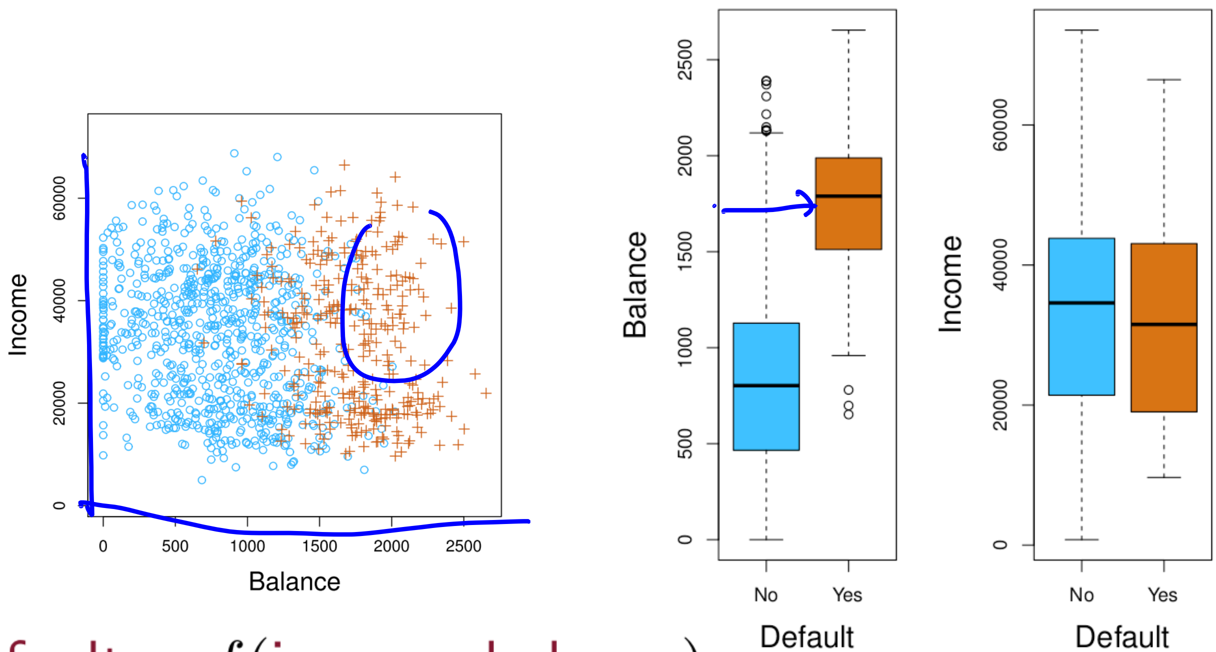
- right side, $K = 1$, training is 0 error but high test error
- Test error has same U shape as bias-variance tradeoff in regression setting: decline at first, then increase again when overfitting

training error

flexibility

Logistic Regression

Example: Predicting Default



$$\text{default} \approx f(\text{income}, \text{balance})$$

Can we use regression?

MICHIGAN STATE UNIVERSITY

Regression: $f : X \mapsto \mathbb{R}$

Classification: $C : X \mapsto \{1, 2, 3\}$

$Y_i = \underline{0}$ or $\underline{1}$

But $\{1, 2, 3\} \subseteq \mathbb{R}$

Do we even need classification?

$(\underbrace{y_i - \hat{y}_i})^2$

Yes!

Regression: Values that are close are similar

Classification: Distance of classes is meaningless

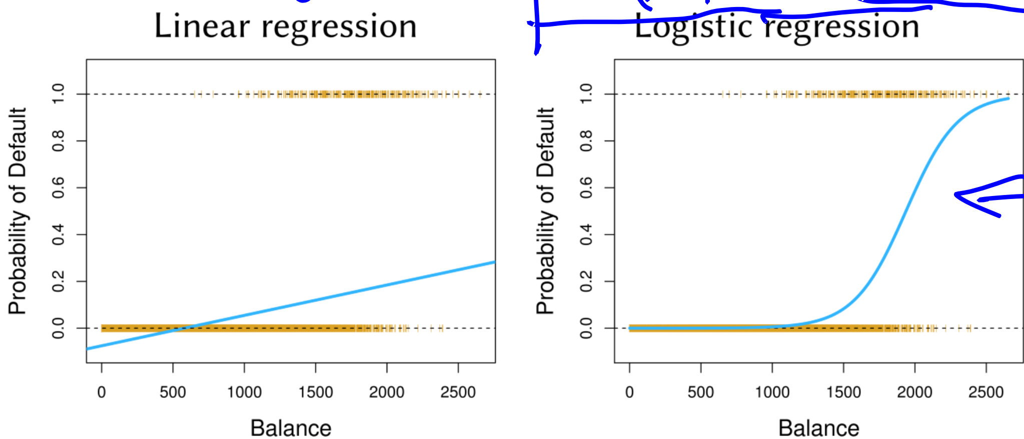
Linear Regression for 2-class

MICHIGAN STATE UNIVERSITY

$$Y = \begin{cases} 1 & \text{if default} \\ 0 & \text{otherwise} \end{cases}$$

$$E(Y|X=x) = \Pr(Y=1|X=x) = p(x)$$

$$\text{logistic}(X) = \boxed{\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$$



$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2}$$

$$\Rightarrow p(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$

$\mathbb{P}[\text{default} = \text{yes} \mid \text{balance}]$

Logistic Function

MICHIGAN STATE UNIVERSITY

Y: 1 ← or 0 ⇐

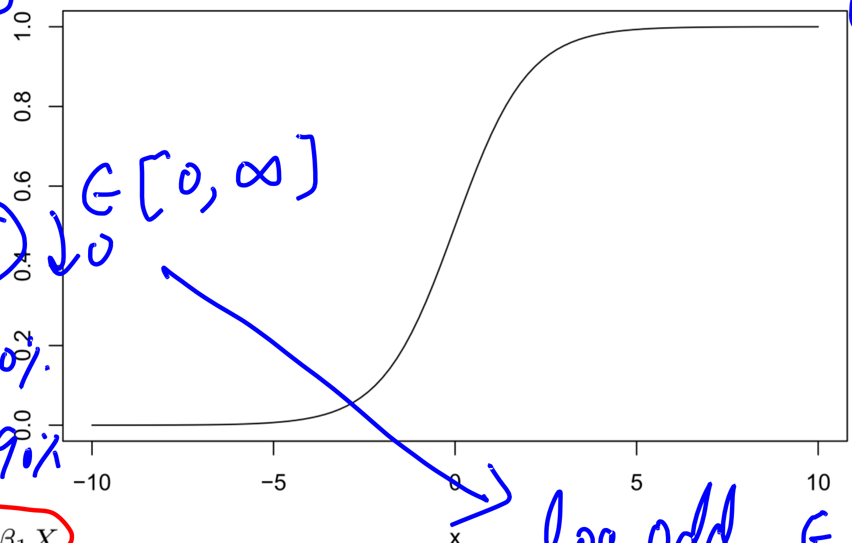
$P(Y=1 \mid X=x)$

$$y = \frac{e^x}{1+e^x}$$

$P(x) = P(Y=1 \mid X=x) \in [0, 1]$ ⇒ $(-\infty, \infty)$

prob $\in [0, 1)$ ⇓

$\text{odd} = \dfrac{P(x)}{1-P(x)} \in [0, \infty]$ ↓ 0



$P(Y=1 \mid X=x) = 10\%$

$P(Y=0 \mid X=x) = 90\%$

$\text{odd} = \frac{1}{9}$

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \in [0, 1]$$

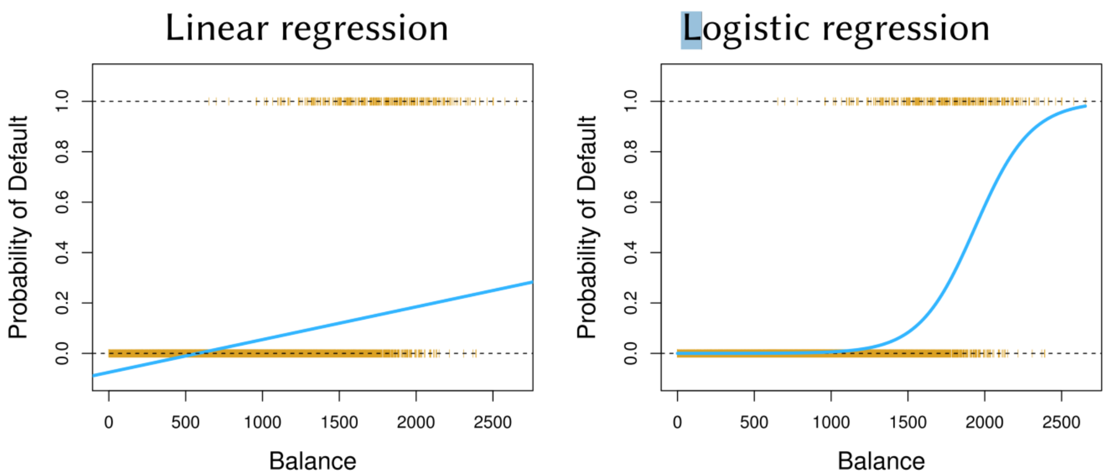
log odd $\in (-\infty, \infty)$

= logistic

Logistic Regression

MICHIGAN STATE UNIVERSITY

$$\mathbb{P}[\textsf{default} = \text{yes} \mid \textsf{balance}] = \frac{e^{\beta_0 + \beta_1 \textsf{balance}}}{1 + e^{\beta_0 + \beta_1 \textsf{balance}}}$$



Using Coefficient to predict

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

What is the estimated probability of default for someone with a balance of $1,000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

What is the estimated probability of default for someone with a balance of $2,000:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$