

Ch 2.2.3: Intro to classification and Logistic regression

Lecture 6 - CMSE 381

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

January 29, 2024

Announcements:

- Homework #3 Due Wednesday
 - No class on Feb 14th
 - The exam review will be on that Monday
 - Will give sample homework solutions
- The requirement for the honors option has been determined: a final project that involves using two of the major methods we learned in this class to tackle a real world problem related to your field/major of study.
 - ▶ send me a message on slack indicating the interest in selecting the honors option
 - ▶ propose a topic of the project
 - ▶ submit the final project before the exam and give a presentation of 10min.

Covered in this lecture

- Ch 2.2.3
- Error rate (classification)
- Bayes Classifier
- K -NN classification
- Logistic regression

Section 1

Classification Overview

What is classification

Classification: When the response variable is qualitative

- qualitative: Unordered set
- Examples
 - ▶ Eye color $\in \{\text{brown}, \text{blue}, \text{green}\}$
 - ▶ email $\in \{\text{spam}, \text{notspam}\}$

- Given feature vector X and qualitative response Y in the set S , the goal is to find a function (classifier) $C(X)$ taking X as input and predicting its value for Y .
- We are more interested in estimating the probabilities that X belongs to each category

Some examples

Discuss what set Y comes from in each case

- Predict whether a COVID19 vaccine will work on a patient given patient's age
- An online banking service wants to determine whether a transaction being performed is fraudulent on the basis of the user's IP address, past transactions, etc.

Section 2

Ch 2.2.3: Classification

Error rate

- Training data:
 $\{(x_1, y_1), \dots, (x_n, y_n)\}$ with y_i qualitative
- Estimate $\hat{y} = \hat{f}(x)$
- Indicator variable
 $I(y_i \neq \hat{y}_i)$ is 1 if not the same (misclassified) and 0 else

Training error rate:

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i)$$

Test error rate:

$$\text{Ave}(I(y_0 \neq \hat{y}_0))$$

Good classifier: one that minimizes test error rate

Best ever classifier

We can't have nice things

Bayes Classifier:

Give every observation the highest probability class given its predictor variables

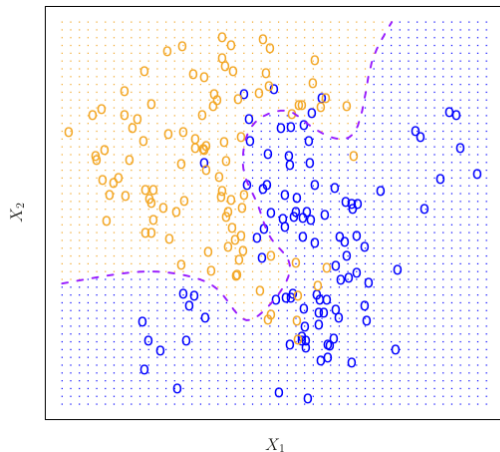
$$\Pr(Y = j \mid X = x_0)$$

- It is possible to show (though the proof is outside of the scope of this book) that the test error rate is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values.

An example

- Survey students for amount of programming experience, and current GPA
- Try to predict if they will pass CMSE 381.
- If we have a survey of all students that could ever exist, we can determine the probability of failure given combo of those features.
- No single classifier will get it right 100% since students with the same features might pass or fail

Bayes decision boundary



- 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

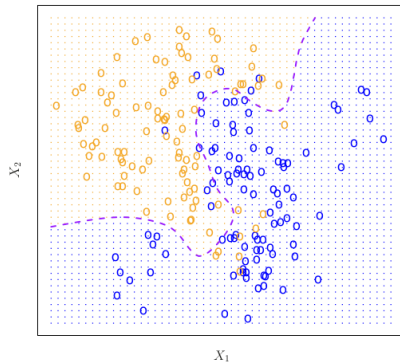
- Error at $X = x_0$ *The j with the highest value is what's predicted by Bayes for that x_0 , so this is probability that the Bayes classifier is wrong*

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

- Overall Bayes error: *This is expectation over all x_0 . This is the analogous to irreducible error*

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

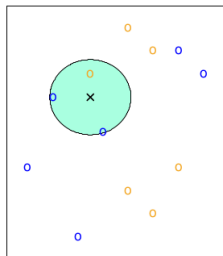
This is the irreducible error



Section 3

K-Nearest Neighbors Classifier

K-Nearest Neighbors

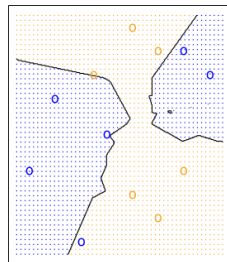


$K = 3$

- Fix K positive integer
- $N(x)$ = the set of K closest neighbors to x
- Estimate conditional probability

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

- Pick j with highest value

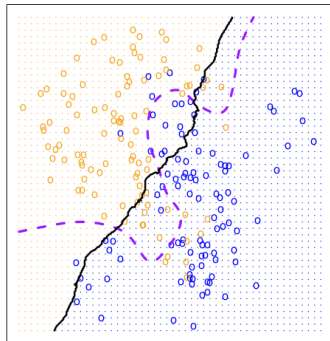


Black line: KNN
decision boundary

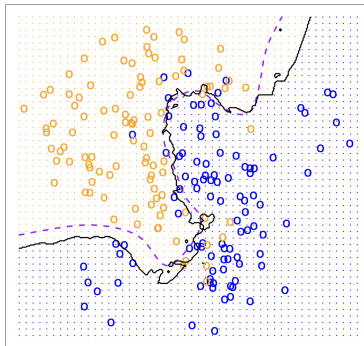
*Note this is not
Bayes like in 2
slides! stupid
textbook*

Tradeoff

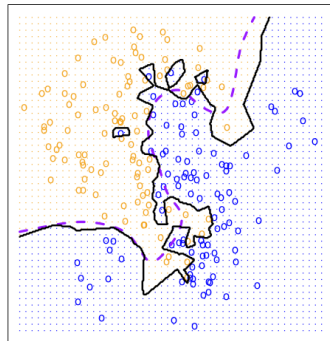
KNN: $K=100$



KNN: $K=10$

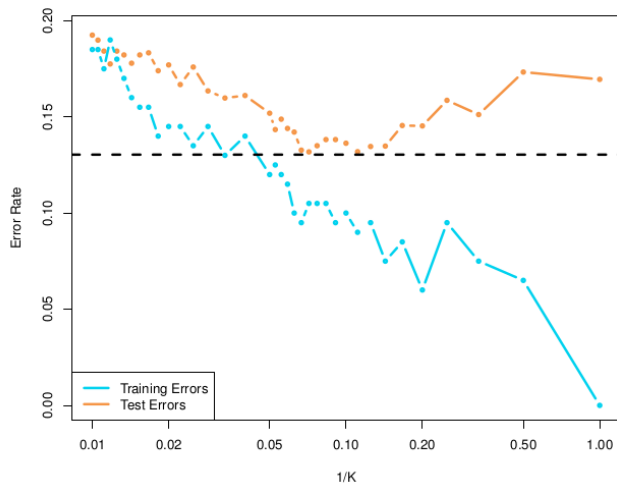


KNN: $K=1$



- Purple is Bayes decision boundary. Same on both
- $K = 1$, boundary is overly flexible, finds patterns not there
- $K = 100$ not enough flexibility

More on tradeoff

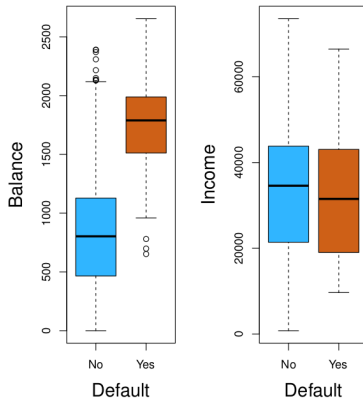
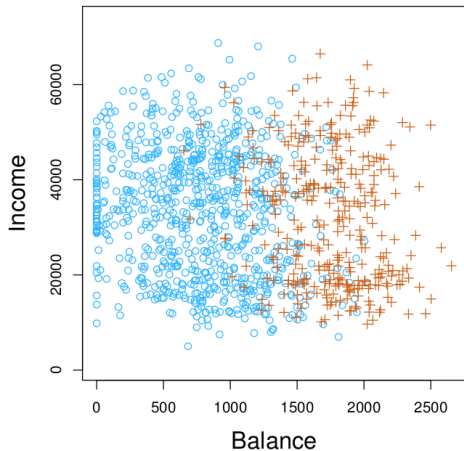


- right side, $K = 1$, training is 0 error but high test error
- Test error has same U shape as bias-variance tradeoff in regression setting: decline at first, then increase again when overfitting

Section 4

Logistic Regression

Simulated Default data set

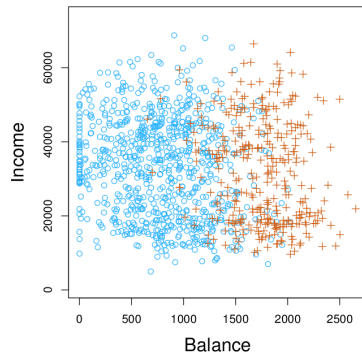


- 10,000 data points
- defaulted on credit card payments: failed to make at least the minimum payment for 180 days
- Individuals who defaulted shown in orange (reality is default rate is actually 3% so only a fraction of those who didn't default are plotted)

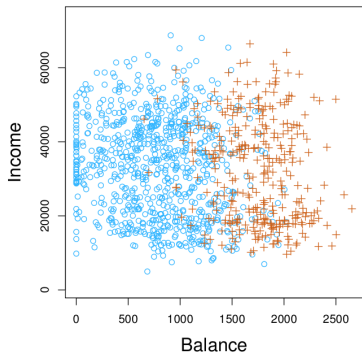
What is classification

- Classification: When the response variable is qualitative
- Goal: Model the probability that Y belongs to a particular category

$$p(\text{balance}) = \Pr(\text{default} = \text{yes} \mid \text{balance})$$



Goal for Balance data set



Goal: Model the probability that Y belongs to a particular category

Ex.

$\Pr(\text{default} = \text{yes} \mid \text{balance})$

- Use notation $p(\text{balance})$
- Choice of threshold depends on comfort with risk
- Note that highly contrived fake data means that largely decidable by balance, so we'll just look at that

Let's just use regression!

JK that's a bad idea

Bad idea:

- Set Y to be a dummy variable taking values in $\{0, 1, 2, \dots\}$
- Run regression, and choose k based on what integer value \hat{y} is closest to

Ex.

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

vs.

$$Y = \begin{cases} 1 & \text{if mild} \\ 2 & \text{if moderate} \\ 3 & \text{if severe} \end{cases}$$

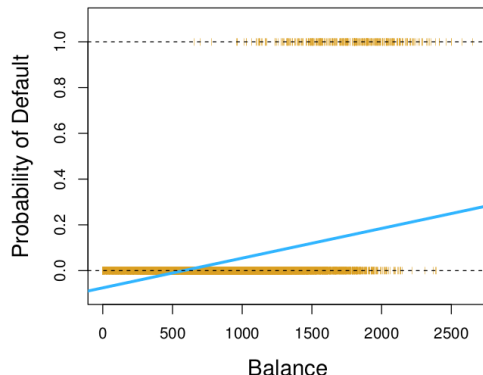
Ordering implies something about closeness. Draw a number line to emphasize this distinction.

Bad idea is still not a great idea for two levels

$$p(\text{balance}) = \Pr(\text{default} = \text{yes} \mid \text{balance})$$

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

- Fit linear regression
- Predict default if $\hat{y} > 0.5$; not default otherwise
- Could spit out values outside of $[0,1]$
- Can't interpret this as probability

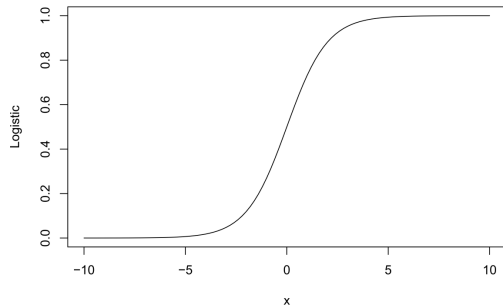


$$p(\text{balance}) = \beta_0 + \beta_1 \text{balance}$$

- Regression methods can't accomodate qualitative response with more than two classes
- Doesn't provide meaningful estimates of $\Pr(Y \mid X)$ even with

Logistic function

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



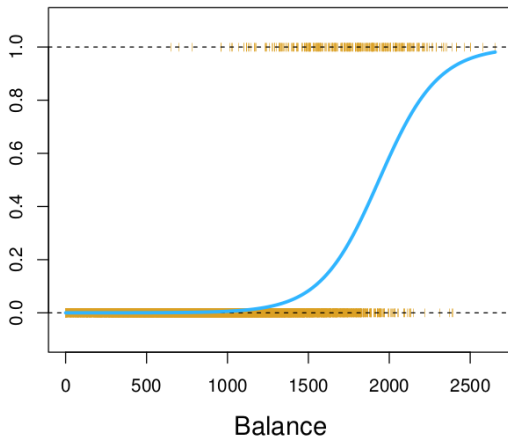
$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Try it out:

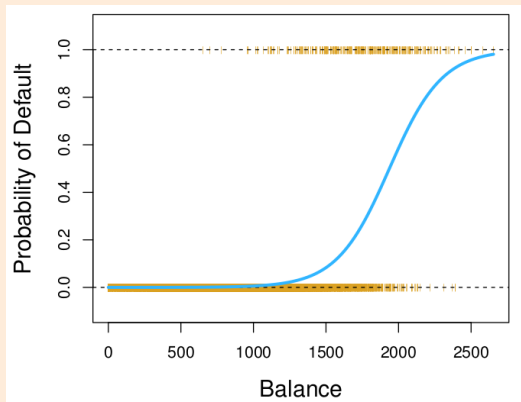
desmos.com/calculator/cw1pyzzqci

Logistic Regression

$$\Pr(\text{default} = \text{yes} \mid \text{balance}) = \frac{e^{\beta_0 + \beta_1 \text{balance}}}{1 + e^{\beta_0 + \beta_1 \text{balance}}}$$




What will the drawn logistic regression classifier predict for each of the following values of Balance




Balance	Prediction
0	No
500	No
1000	No
1500	Yes
2000	Yes
2500	Yes

$$\frac{p(x)}{1 - p(x)} = \frac{\Pr(Y = 1 \mid X = x)}{1 - \Pr(Y = 1 \mid X = x)} = \frac{\Pr(Y = 1 \mid X = x)}{\Pr(Y = 0 \mid X = x)}$$

Examples:

Probability
or risk $= \frac{p}{p+q}$ 

Odds $= p : q$ 

- Logistic function is chosen so that odds are linear
- Can take any value from 0 (low odds) to ∞ (high odds)

- If the probability of default is 90% what are the odds?

- ▶ $p(x) = 0.9$
- ▶ $\frac{0.9}{1-0.9} = 9$

- If the odds are 1/3, what is the probability of default?

- ▶ $\frac{p}{1-p} = 1/3$
- ▶ $3p = 1 - p$
- ▶ $4p = 1$
- ▶ $p = 1/4$

How to get logistic function

Assume the (natural) log odds (logits) follow a linear model

$$\log \left(\frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x$$

Solve for $p(x)$:

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

- Note that we can't use the linear interpretation from before since one unit increase in β_1 doesn't correspond to one unit of odds if that made sense
- But can say that positive β_1 corresponds to increasing $p(X)$ and vice versa

Playing with the logistic function: desmos.com/calculator/cw1pyzzqci

Using coefficients to make predictions

	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:

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

Interpreting the coefficients

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

$$\log \left(\frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x$$

	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

- β_1 means increasing x by one unit increases the log odds by β_1 unit
- In this case, increasing the balance by 1 means increasing the log odds of default by 0.0055

Confusion Matrix: Predicting default from balance

		<i>True default status</i>		
		No	Yes	Total
<i>Predicted default status</i>	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- Overall error $(c + d)/N = 2.75\%$
- This table is from the LDA section, but we can use this for understanding confusion matrix

		True		Total
		Yes	No	
Predicted	Yes	<i>a</i>	<i>b</i>	$a + b$
	No	<i>c</i>	<i>d</i>	$c + d$
Total		$a + c$	$b + d$	N