Ch 2.2: Assessing Model Accuracy

Lecture 3 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Weds, Sep 7, 2022

Announcements

Lec #		Date	Topic	Reading	Homeworks
1	W	Aug 31	Intro / First day stuff / Python Review Pt 1	1	
2	F	Sep 2	What is statistical learning?	2.1	
	M	Sep 5	No class - Labor day		
3	W	Sep 7	Assessing Model Accuracy	2.2.1, 2.2.2	HW #1 Due
4	F	Sep 9	Linear Regression	3.1	
5	M	Sep 12	More Linear Regression	3.1/3.2	
6	W	Sep 14	Even more linear regression	3.2.2	HW #2 Due
7	F	Sep 16	Probably more linear regression	3.3	
8	M	Sep 19	Intro to classification, Logisitic Regression	2.2.3, 4.1, 4.2, 4.3	
9	W	Sep 21	More logistic regression		HW #3 Due
10	F	Sep 23	Review		
11	M	Sep 26	***Midterm #1***		
12	W	Sep 28	[No class, Dr Munch out of town]		
13	F	Sep 30	[No class, Dr Munch out of town]		

Last Time:

-

Announcements:

- Homework #3 Due Wednesday on Crowdmark
- Friday - Review day
  - Nothing prepped
  - Bring your questions
- Monday - Exam #1
  - Bring 8.5x11 sheet of paper
  - Handwritten both sides
  - Ahything you want on it, but must be your work
  - You will turn it in

Covered in this lecture

- Ch 2.2.3, 4.1, 4.2, part of 4.3
- Error rate (classification)
- Bayes Classifier
- K-NN classification
- Start Logistic Regression

Note: No jupyter notebook today

Section 1

Classification Overview

What is classification

Classification: When the response variable is qualitative

- 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

- 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: $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ with y_i qualitative
- Estimate $\hat{y} = \hat{f}(x)$
- Indicator variable

Training error rate:

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

Test error rate:

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

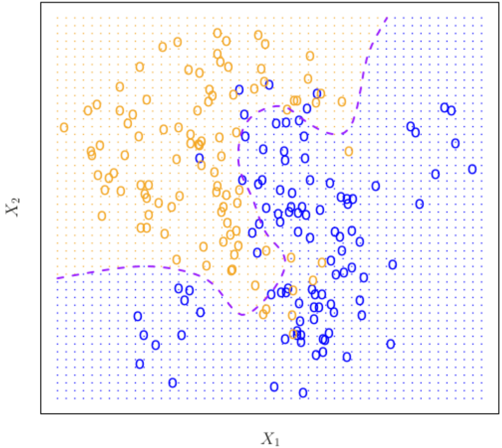
Bayes Classifier:
Give every observation the highest probability class given its predictor variables

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

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.

Bayes decision boundary



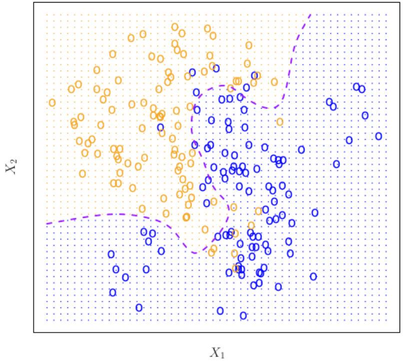
Bayes error rate

- Error at $X = x_0$

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

- Overall Bayes error:

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

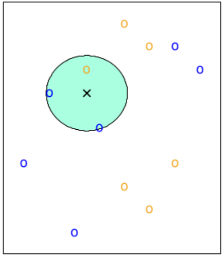


The game

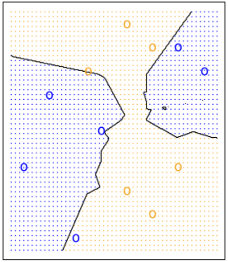
Section 3

K-Nearest Neighbors Classifier

K-Nearest Neighbors



$K = 3$



decision boundary

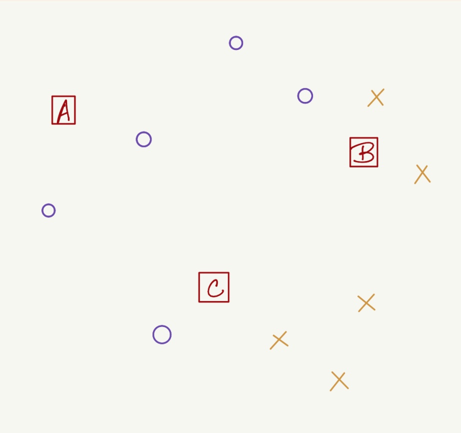
- Fix K positive integer
- $N(x)$ = the set of K closest neighbors to x
- Estimate conditional proability

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

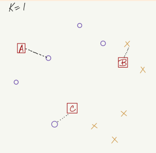
- Pick j with highest value

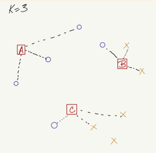
Example

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

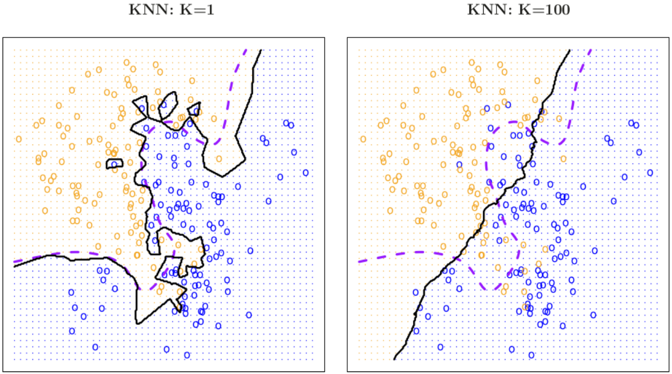


Point **Point**	$k = 1$ Prediction	$k = 3$ Prediction
A		
B		
C		

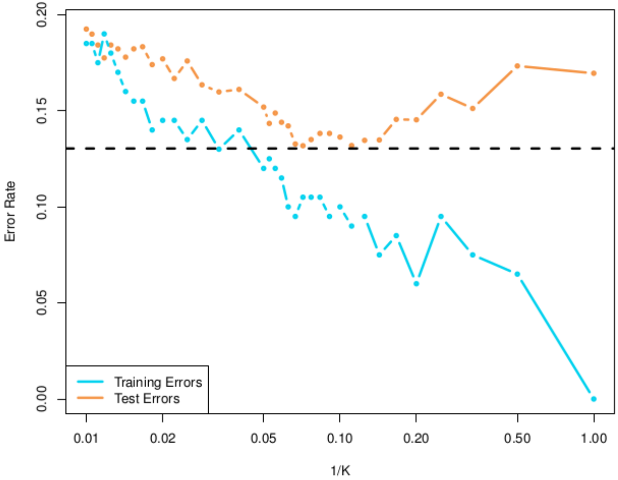




Tradeoff



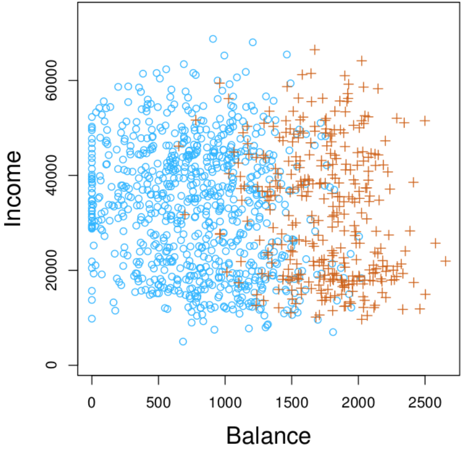
More on tradeoff

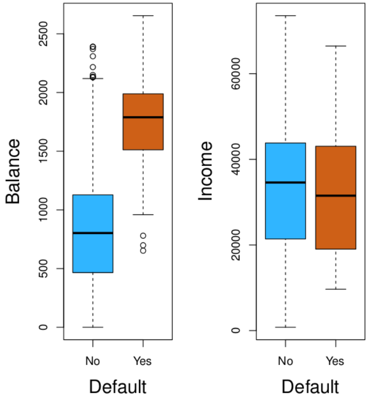


Section 4

Intro to Logistic Regression

Simulated Default data set





Let's just use regression!

JK that's a bad idea

Bad idea:

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

Ex.

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

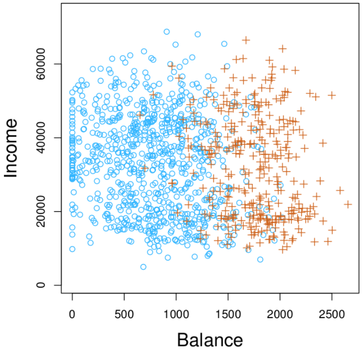
vs.

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

Bad idea is still not a great idea for two levels

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

- Fit linear regression
- Predict overdose if $\hat{y} > 0.5$; stroke otherwise



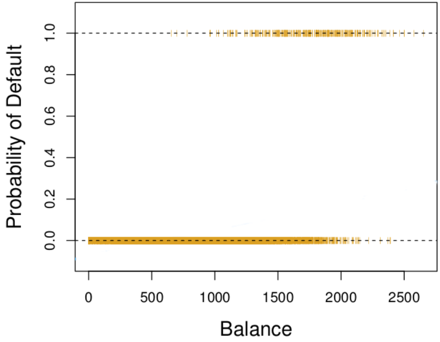
Goal: Model the probability that Y belongs to a particular category

Ex.

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

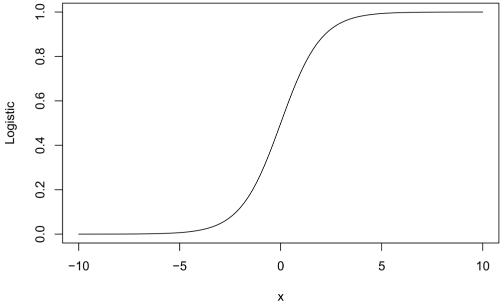
Approximating the probability

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



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

- Let $p(X) = \Pr(Y = 1 \mid X)$.
- Turn this into something with range $\mathbb{R}$

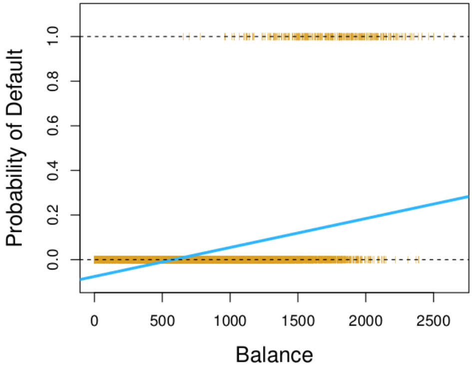
$$\ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

- Called *log odds*, or *logit*
- Solving for $p(X)$ gets

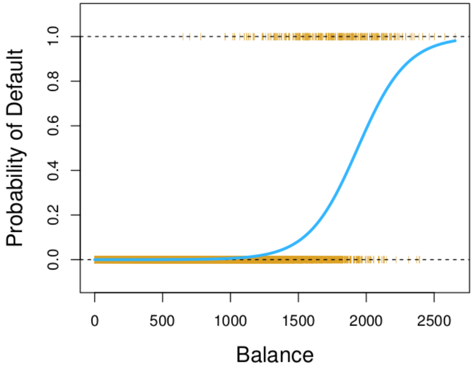
$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Logistic Regression

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

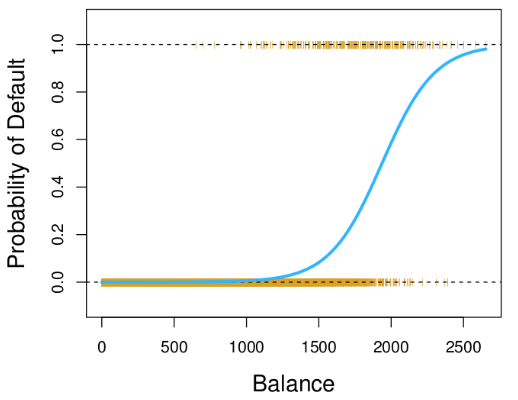


Linear Regression



Logistic Regression

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



Point	**Prediction**
0	
500	
1000	
1500	
2000	
2500	

Next time

Lec #		Date	Topic	Reading	Homeworks
1	W	Aug 31	Intro / First day stuff / Python Review Pt 1	1	
2	F	Sep 2	What is statistical learning?	2.1	
	M	Sep 5	No class - Labor day		
3	W	Sep 7	Assessing Model Accuracy	2.2.1, 2.2.2	HW #1 Due
4	F	Sep 9	Linear Regression	3.1	
5	M	Sep 12	More Linear Regression	3.1/3.2	
6	W	Sep 14	Even more linear regression	3.2.2	HW #2 Due
7	F	Sep 16	Probably more linear regression	3.3	
8	M	Sep 19	Intro to classification, Logisitic Regression	2.2.3, 4.1, 4.2, 4.3	
9	W	Sep 21	More logistic regression		HW #3 Due
10	F	Sep 23	Review		
11	M	Sep 26	***Midterm #1***		
12	W	Sep 28	[No class, Dr Munch out of town]		
13	F	Sep 30	[No class, Dr Munch out of town]		
14	M	Oct 3	Leave one out CV	5.1.1, 5.1.2	
15	W	Oct 5	k-fold CV	5.1.3	
16	F	Oct 7	More k-fold CV	5.1.4	
17	M	Oct 10	CV for classification	5.1.5	HW #4 Due
18	W	Oct 12	Resampling methods: Bootstrap	5.2	
19	F	Oct 14	Subset selection	6.1	
20	M	Oct 17	Shrinkage: Ridge	6.2.1	HW #5 Due
21	W	Oct 19	Shrinkage: Lasso	6.2.2	
22	F	Oct 21	Dimension Reduction	6.3	