Ch 2.2.3: Intro to classification

Lecture 9 - CMSE 381

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

Mon, Feb 2, 2026

Announcements

11	F	2/6	Multiple Logistic Regression / Multinomial Logistic Regression	4.3.4-5	HW #2 Due Sun 2/8	
	M	2/9	***Project Day & Review***			
	W	2/11	**Midterm #1**			
12	F	2/13	Leave one out CV	5.1.1, 5.1.2		
13	M	2/16	k-fold CV	5.1.3		
14	W	2/18	More k-fold CV	5.1.4-5		Q5
15	F	2/20	k-fold CV for classification	5.1.5		
16	M	2/23	Subset selection	6.1		
17	W	2/25	Shrinkage: Ridge	6.2.1		
18	F	2/27	Shrinkage: Lasso	6.2.2	HW #3 Due Sun 3/1	
	M	3/2	Spring Break			
	W	3/4	Spring Break			
	F	3/6	Spring Break			
19	M	3/9	PCA	6.3		
20	W	3/11	PCR	6.3		Q6

Last Time:

- Finished Linear Regression

Announcements:

- Homework #2 Due Sunday Feb 8
- Next Monday - Review day
  - depends on what you ask!
- Wed 2/11 - Exam #1
  - Bring 8.5x11 sheet of paper
  - **Handwritten** both sides
  - Anything you want on it, but must be your work
  - Must have your name and group number
  - You must turn it in

Covered in this lecture

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

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:

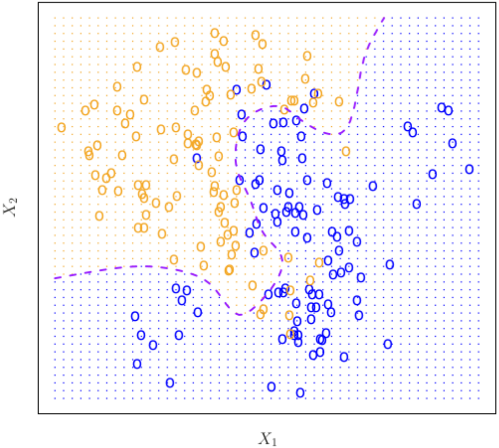
Give every observation the highest probability class given its predictor variables

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

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



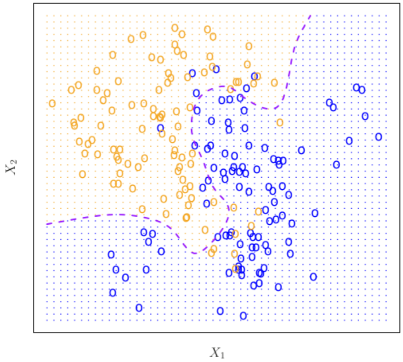
Bayes error rate

- Error at $X = x_0$

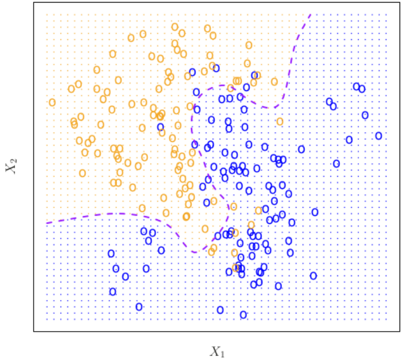
$$1 - \max_j \Pr(Y = j \mid X = x_0)$$

- Overall Bayes error:

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



The game

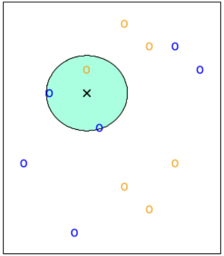


Test your understanding: PollEv

Section 3

K-Nearest Neighbors Classifier

K-Nearest Neighbors

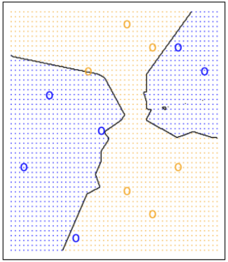


$K = 3$

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

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

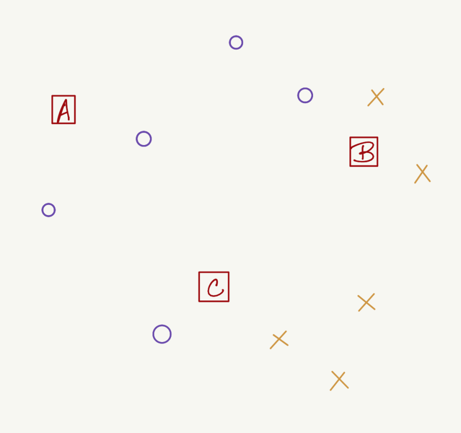
- Pick j with highest value



Black line: KNN decision boundary

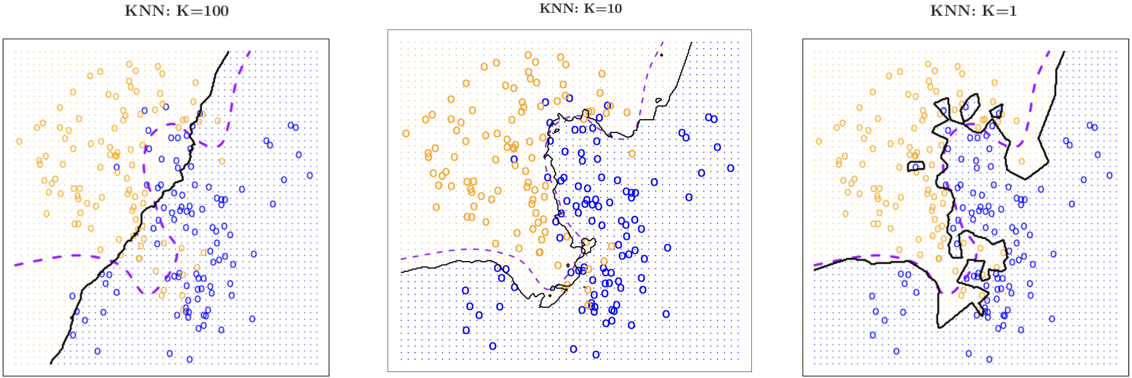
Example

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

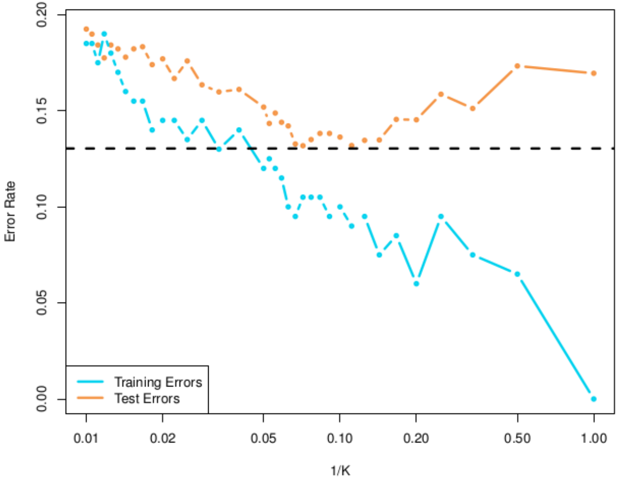


Point	$k = 1$ Prediction	$k = 3$ Prediction
A		
B		
C		

Tradeoff



More on tradeoff



Jupyter notebook

Next time

CMSE381_S2026_Schedule : Sheet1

Lec #	Date		Topic	Reading	HW	Pop Quizzes	Notes
1	M	1/12	Intro / Python Review	1			
2	W	1/14	What is statistical learning	2.1		Q1	
3	F	1/16	Assessing Model Accuracy	2.2.1, 2.2.2			
	M	1/19	MLK - No Class				
4	W	1/21	Linear Regression	3.1		Q2	
5	F	1/23	More Linear Regression	3.1	HW #1 Due Sun 1/25		
6	M	1/26	Multi-linear Regression	3.2			
7	W	1/28	Probably More Linear Regression	3.3		Q3	
8	F	1/30	Last of the Linear Regression				
9	M	2/2	Intro to classification, Bayes classifier, KNN classifier	2.2.3			
10	W	2/4	Logistic Regression	4.1, 4.2, 4.3.1-3		Q4	