Ch 2.2.3: Intro to classification

Lecture 9 - CMSE 381

Prof. Elizabeth Munch

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

Weds, Sep 18, 2023

Announcements

Lec #	Date			Reading	HW
1	Mon	8/26	Intro / First day stuff / Python Review Pt 1	1	
2	Wed	8/28	What is statistical learning?	2.1	
3	Wed	9/4	Assessing Model Accuracy	2.2.1, 2.2.2	
4	Fri	9/6	Linear Regression	3.1	HW #1 Due Sun 9/8
5	Mon	9/9	More Linear Regression	3.1	
6	Wed	9/11	Multi-linear regression	3.2	
7	Fri	9/13	Probably more linear regression	3.3	Hw #2 Due Dun 9/15
8	Mon	9/16	Last of the linear regression		
9	Wed	9/18	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
10	Fri	9/20	Logistic Regression	4.1, 4.2, 4.3.1-3	
11	Mon	9/23	Multiple Logistic Regression / Multinomial Logistic Regression /Project day	4.3.4-5	Hw #3 Due Sun 9/22
	Wed	9/25	***Review***		
	Fri	9/27	**Midterm #1**		

Last Time:

- Finished Linear Regression

Announcements:

- Homework #3 Due Sunday Sep 22
- Next Wednesday - Review day
  - Nothing prepped
  - Bring your questions
- Friday 9/27 - Exam #1
  - Bring 8.5x11 sheet of paper
  - Handwritten both sides
  - Anything you want on it, but must be your work
  - You will 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:

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

Best ever classifier

We can't have nice things

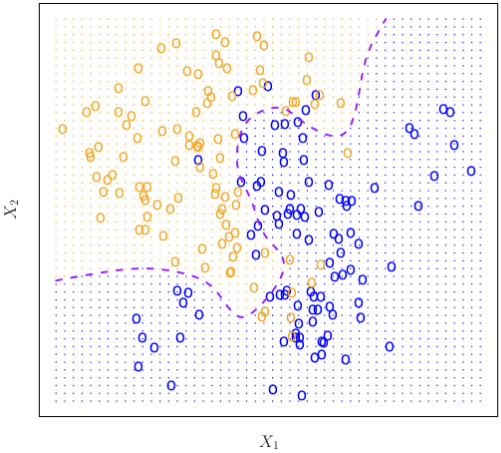
Bayes Classifier:
Give every observation the highest probability class given its predictor variables

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

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



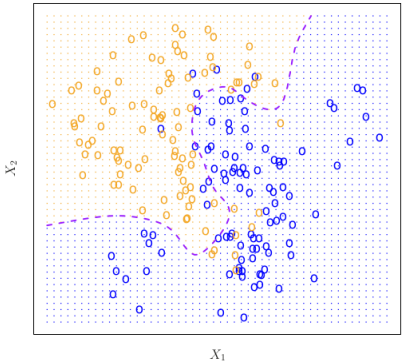
Bayes error rate

- Error at $X = x_0$

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

- Overall Bayes error:

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

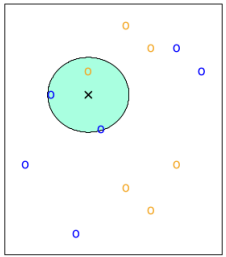


The game

Section 3

K-Nearest Neighbors Classifier

K-Nearest Neighbors

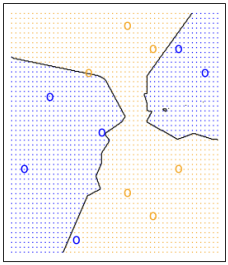


$K = 3$

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

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

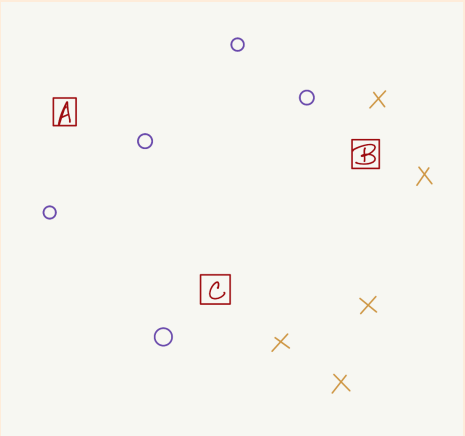
- Pick j with highest value



Black line: KNN decision boundary

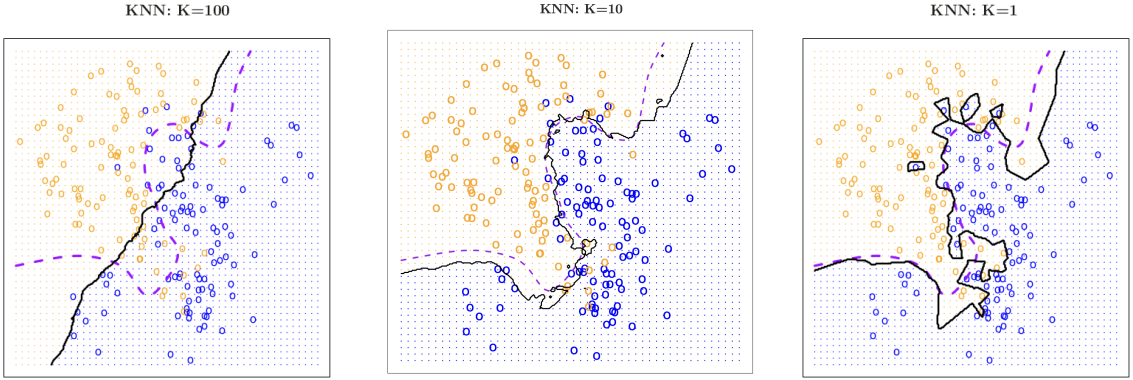
Example

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

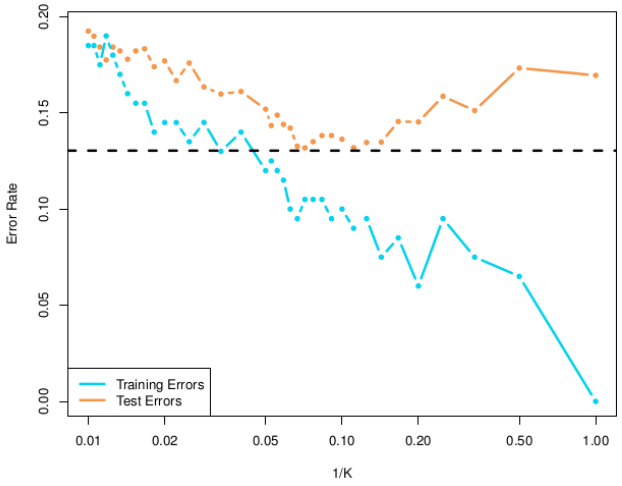


Point	$k = 1$ Prediction	$k = 3$ Prediction
A		
B		
C		

Tradeoff



More on tradeoff



Jupyter notebook

Next time

Lec #	Date			Reading	HW
1	Mon	8/26	Intro / First day stuff / Python Review Pt 1	1	
2	Wed	8/28	What is statistical learning?	2.1	
3	Wed	9/4	Assessing Model Accuracy	2.2.1, 2.2.2	
4	Fri	9/6	Linear Regression	3.1	HW #1 Due Sun 9/8
5	Mon	9/9	More Linear Regression	3.1	
6	Wed	9/11	Multi-linear regression	3.2	
7	Fri	9/13	Probably more linear regression	3.3	Hw #2 Due Dun 9/15
8	Mon	9/16	Last of the linear regression		
9	Wed	9/18	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
10	Fri	9/20	Logistic Regression	4.1, 4.2, 4.3.1-3	Hw #3 Due Sun 9/22
11	Mon	9/23	Multiple Logistic Regression / Multinomial Logistic Regression /Project day	4.3.4-5	
	Wed	9/25	***Review***		
	Fri	9/27	**Midterm #1**		