# Ch 2.2.3: Intro to classification

## Lecture 9 - CMSE 381


Prof. Mengsen Zhang

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


Mon, Feb 3, 2025

# Announcements

CMSE381_S2025_Schedule : Sheet1

| Lec # | Date | | Topic | Reading | HW |
|---|---|---|---|---|---|
| 1 | M | 1/13 | Intro / Python Review | 1 | |
| 2 | W | 1/15 | What is statistical learning | 2.1 | |
| 3 | F | 1/17 | Assessing Model Accuracy | 2.2.1, 2.2.2 | |
| | M | 1/20 | MLK - No Class | | |
| 4 | W | 1/22 | Linear Regression | 3.1 | |
| 5 | F | 1/24 | More Linear Regression | 3.1 | HW #1 Due Sun 1/26 |
| 6 | M | 1/27 | Multi-linear Regression | 3.2 | |
| 7 | W | 1/29 | Probably More Linear Regression | 3.3 | |
| 8 | F | 1/31 | Last of the Linear Regression | | HW #2 Due Sun 2/1 |
| 9 | M | 2/3 | Intro to classification, Bayes classifier, KNN classifier | 2.2.3 | |
| 10 | W | 2/5 | Logistic Regression | 4.1, 4.2, 4.3.1-3 | |
| 11 | F | 2/7 | Multiple Logistic Regression / Multinomial Logistic Regression | 4.3.4-5 | HW #3 Due Sun 2/9 |
| | M | 2/10 | ***Project Day & Review*** | | |
| | W | 2/12 | **Midterm #1** | | |
| 12 | F | 2/14 | Leave one out CV | 5.1.1, 5.1.2 | |
| 13 | M | 2/17 | k-fold CV | 5.1.3 | |
| 14 | W | 2/19 | More k-fold CV | 5.1.4-5 | |
| 15 | F | 2/21 | k-fold CV for classification | 5.1.5 | HW #4 Due Sun 2/23 |
| 16 | M | 2/24 | Subset selection | 6.1 | |

**Last Time:**

- Finished Linear Regression

**Announcements:**

- Homework #3 Due Sunday Feb 9
- Next Monday - Review day
  - Nothing prepped
  - Bring your questions
- Wed 2/12 - 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

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

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

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

highest Pr

Conditional Prob

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

# An example

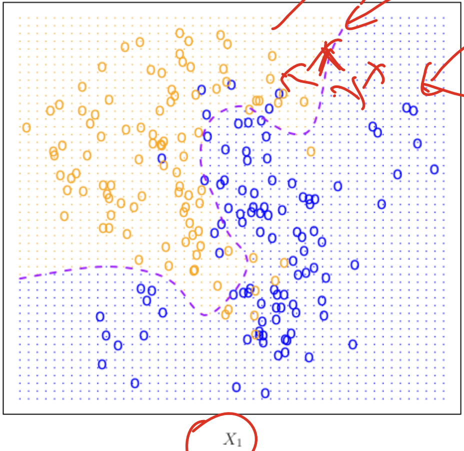
(pass, fail)

- 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.

Pr(pass | STT 380) > 0.5

> 0.9

# Bayes decision boundary



# Bayes error rate (Ideal)

- Error at $X = x_0$

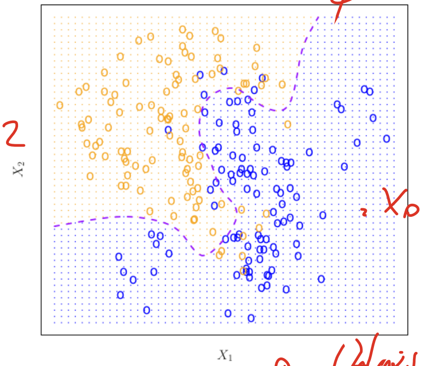
$$1 - \max_j \Pr(Y = j \mid X = x_0)$$

0.98 = 0.02

- Overall Bayes error:

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

irreducible error

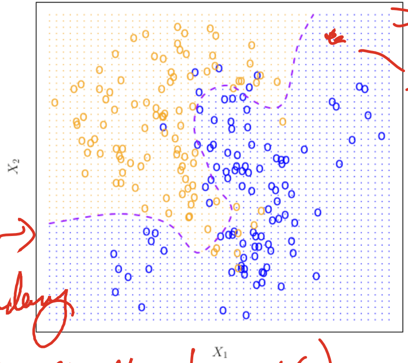


# The game

- challenge: don't know prob.

- Game: guess Bayes → prob / decision boundary

→ Pr(blue | $X_1$, $X_2$)   Pr(Yellow | $X_1$, $X_2$)




→ Pr(blue) = 0.6

= 0.4

→ blue

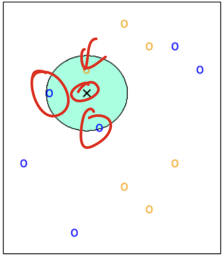
Section 3

## $K$-Nearest Neighbors Classifier

$Pr(\text{Blue} \mid X_1, X_2)$

$Pr(\text{Yellow} \mid x_1, x_2, \ldots)$

# $K$-Nearest Neighbors

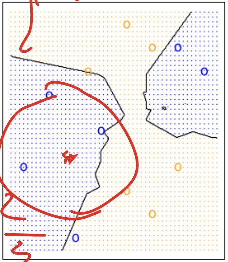


$K = 3$

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

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

- Pick $j$ with highest value

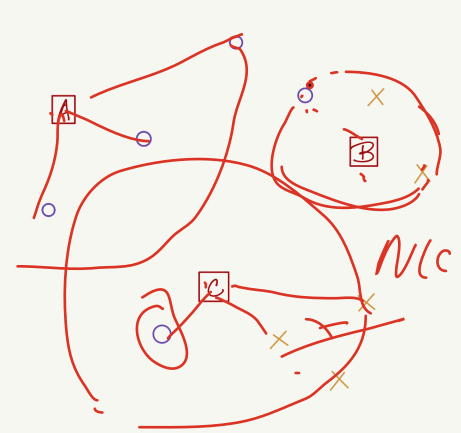


Black line: KNN decision boundary

# Example

k-NN

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

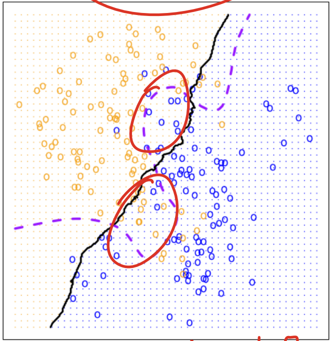



N(C)

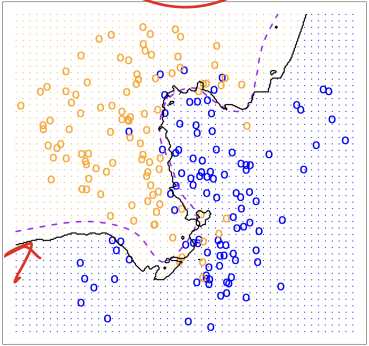
| **Point** | $k = 1$ Prediction | $k = 3$ Prediction |
|---|---|---|
| A | O | O |
| B | X | X |
| C | O | X |

# Tradeoff

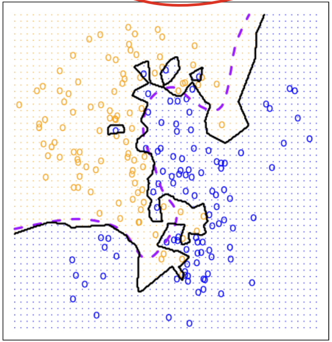
KNN: K=100



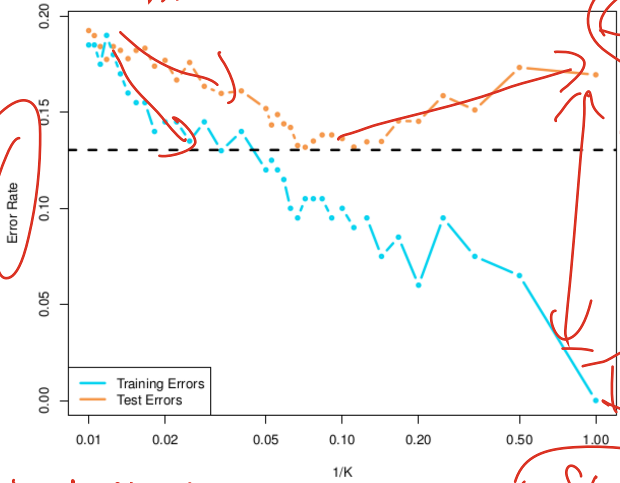
KNN: K=10



KNN: K=1



# More on tradeoff



# Jupyter notebook

# Next time

CMSE381_S2025_Schedule : Sheet1

| Lec # | Date | | Topic | Reading | HW |
|---|---|---|---|---|---|
| 1 | M | 1/13 | Intro / Python Review | 1 | |
| 2 | W | 1/15 | What is statistical learning | 2.1 | |
| 3 | F | 1/17 | Assessing Model Accuracy | 2.2.1, 2.2.2 | |
| | M | 1/20 | MLK - No Class | | |
| 4 | W | 1/22 | Linear Regression | 3.1 | |
| 5 | F | 1/24 | More Linear Regression | 3.1 | HW #1 Due Sun 1/26 |
| 6 | M | 1/27 | Multi-linear Regression | 3.2 | |
| 7 | W | 1/29 | Probably More Linear Regression | 3.3 | |
| 8 | F | 1/31 | Last of the Linear Regression | | HW #2 Due Sun 2/1 |
| 9 | M | 2/3 | Intro to classification, Bayes classifier, KNN classifier | 2.2.3 | |
| 10 | W | 2/5 | Logistic Regression | 4.1, 4.2, 4.3.1-3 | |
| 11 | F | 2/7 | Multiple Logistic Regression / Multinomial Logistic Regression | 4.3.4-5 | HW #3 Due Sun 2/9 |
| | M | 2/10 | ***Project Day & Review*** | | |
| | W | 2/12 | **Midterm #1** | | |
| 12 | F | 2/14 | Leave one out CV | 5.1.1, 5.1.2 | |
| 13 | M | 2/17 | k-fold CV | 5.1.3 | |
| 14 | W | 2/19 | More k-fold CV | 5.1.4-5 | |
| 15 | F | 2/21 | k-fold CV for classification | 5.1.5 | HW #4 Due Sun 2/23 |
| 16 | M | 2/24 | Subset selection | 6.1 | |