LECTURE 05:

CLASSIFICATION PT. 1

September 25, 2017
SDS 293: Machine Learning

Q&A: homework format

Q: *What file format should we use for our homework?*

A: PDF is fine for conceptual exercises; Jupyter notebook is preferable for applied exercises but .Rmd and .py are also acceptable

Q&A: confidence vs. prediction intervals

Q: *Problem 3.8 asks for the predicted* **`mpg`** *of a car with 98* **`horsepower`***, and then asks for the associated prediction and confidence intervals. I thought that for single number predictions you could only use prediction intervals and confidence intervals were for means and coefficients?*

A: Let's step back and consider what each one is telling us:

Q&A: confidence vs. prediction intervals

Confidence interval

We have 95% confidence that the

mean of all samples with these predictors

will fall within this interval

Prediction interval

We have 95% confidence that the

next sample with these predictors

will fall within this interval

Outline

- Motivation
- Bayes classifier
- K-nearest neighbors
- Logistic regression
  - Logistic model
  - Estimating coefficients with maximum likelihood
  - Multivariate logistic regression
  - Multiclass logistic regression
  - Limitations
- Linear discriminant analysis (LDA)
  - Bayes' theorem
  - LDA on one predictor
  - LDA on multiple predictors
- Comparing Classification Methods

Motivation

- **So far:** predicted *quantitative* responses

$$height = \beta_1 \left(\right) + \beta_2 \left(\right) + \beta_3 \left(\right)$$

Image credit: Ming Malaykham

Motivation

- **Question:** what if we have a *qualitative* response?

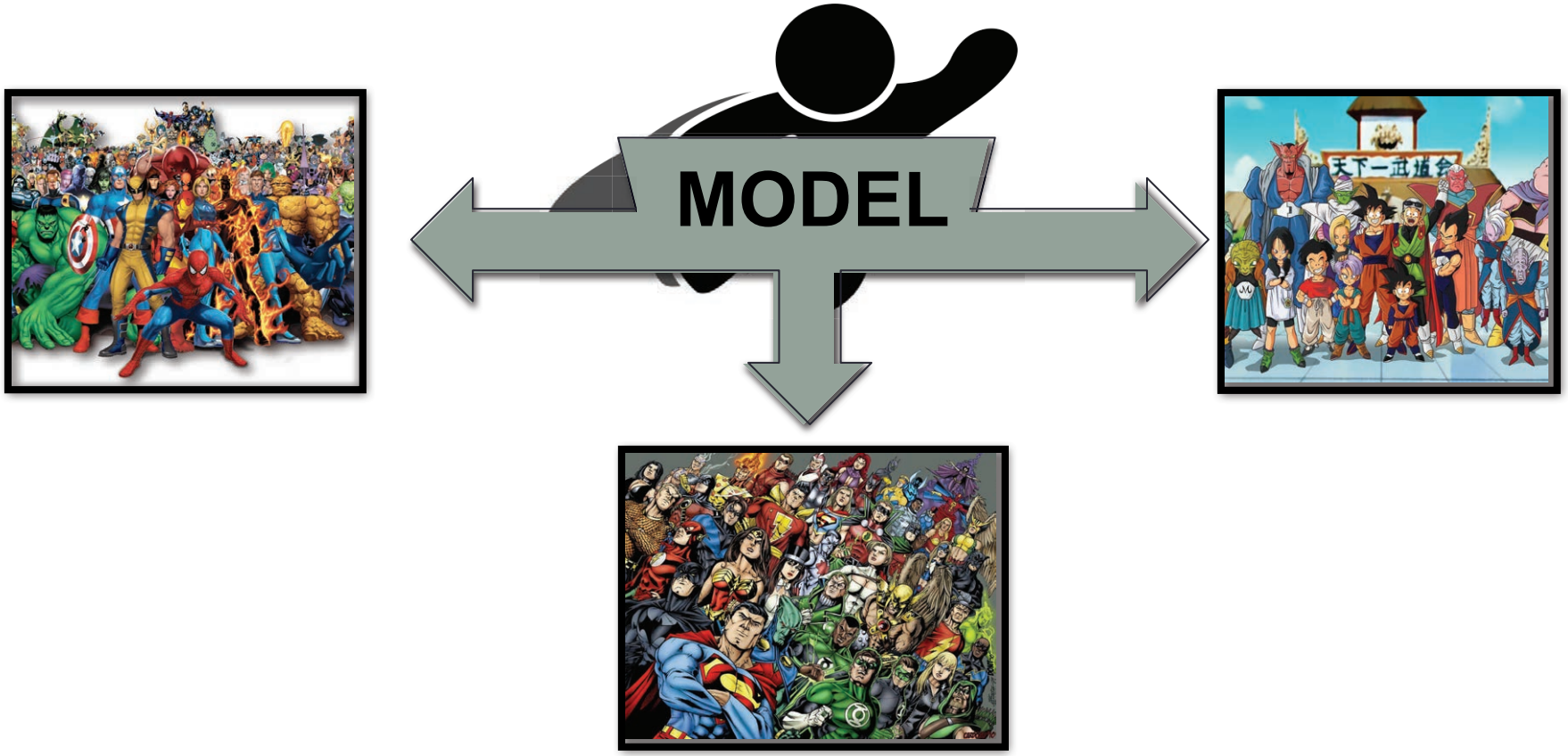
Motivation

- Like in regression, we have a set of training *observations*:

- Want to build a model to predict the (qualitative) response for **new** observations

Classification

- Predicting a qualitative response for an observation can be thought of as **classifying** that observation



Quantifying accuracy

- With quantitative responses, we measured a model's accuracy using the *mean squared error*:

$$MSE = avg\left(\left(\hat{f}(x_0) - y_0\right)^2\right)$$

- **Problem:** how do we take the difference of two classes

$$\left(\text{MARVEL} - \text{DC}\right)?$$

- **Solution:** we'll try to minimize the proportion of misclassifications (the *test error rate*) instead:

$$TE = avg\left(I(y_0 \neq \hat{y}_0)\right)$$

*$I(\ldots)$ is an indicator function that outputs 1 if the input is true, 0 otherwise

Why won't linear regression work?

- LR only works on **quantitative** responses (why?)
- Could try approach we took with qualitative **predictors**:

$$Y = \begin{cases} 1 & if \\ 2 & if \\ 3 & if \end{cases}$$

- What's the problem?

Why won't linear regression work?

- Is it any better if we only have a binary response?

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

or

- In this case, how might we interpret $\beta_1 X$?

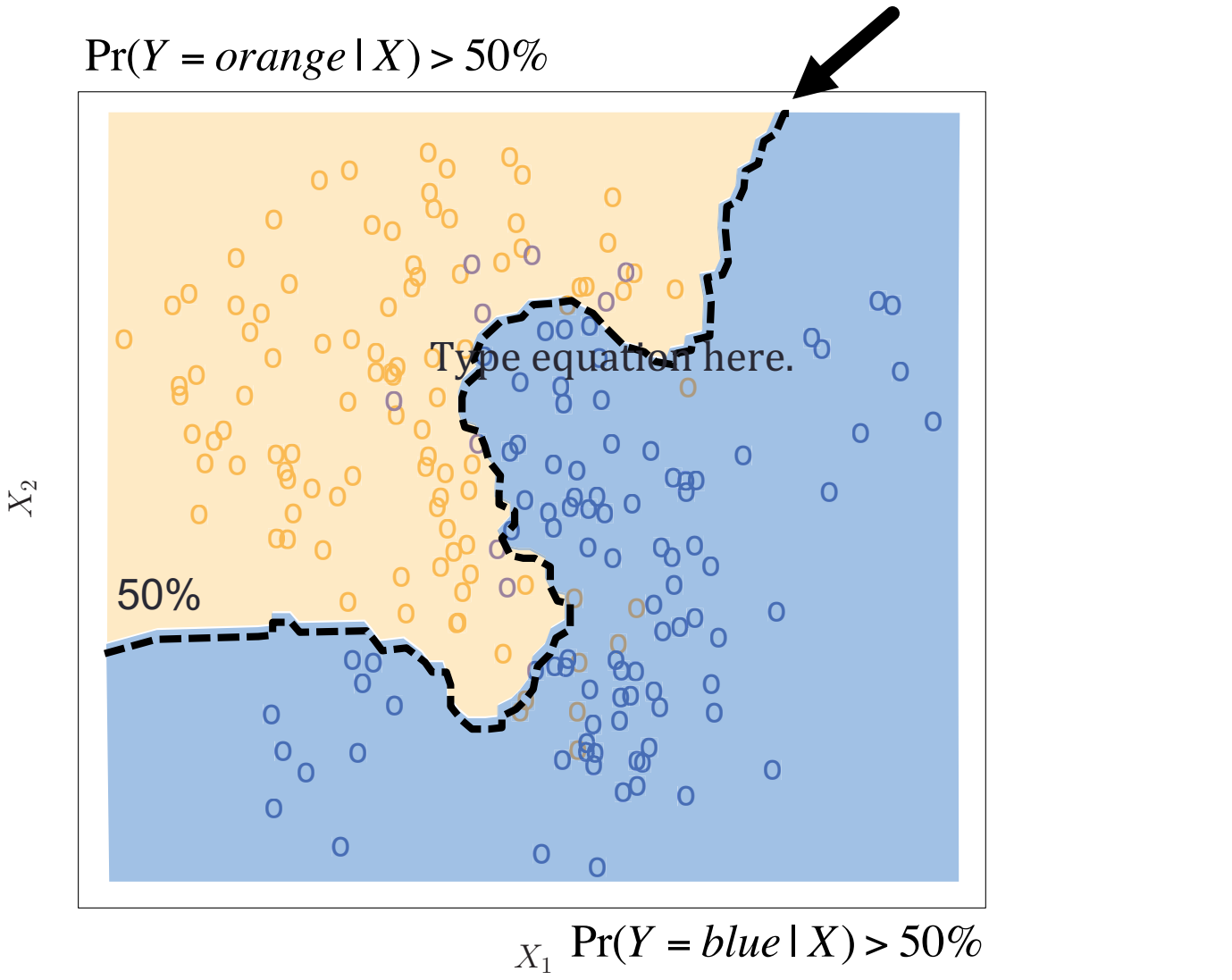
Bayes' classifier

- A simple, elegant classifier: *assign each observation to the most likely class, given its predictor values*

- Mathematically: assign a test observation with predictor vector x_0 to the class j that maximizes:

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

Toy example


Bayes' Decision Boundary
$\Pr(Y = orange \mid X) > 50\%$
Type equation here.
50%
X_2
X_1
$\Pr(Y = blue \mid X) > 50\%$

Bayes' classifier

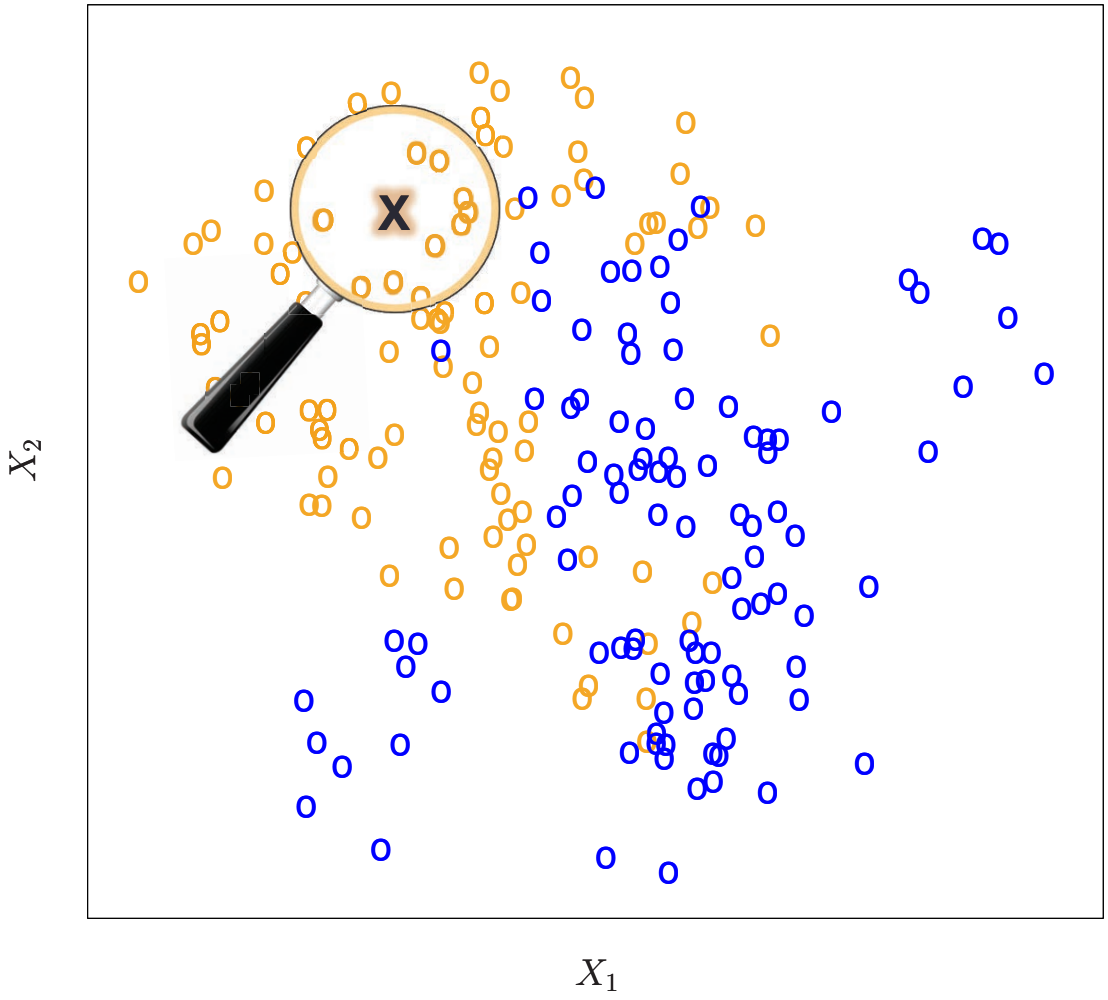
- Test error rate of the Bayes classifier:

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

- Great news! This error rate is provably* optimal!
- Just one problem…
- Okay, let's estimate!

* proof left to the reader

Back to our toy example



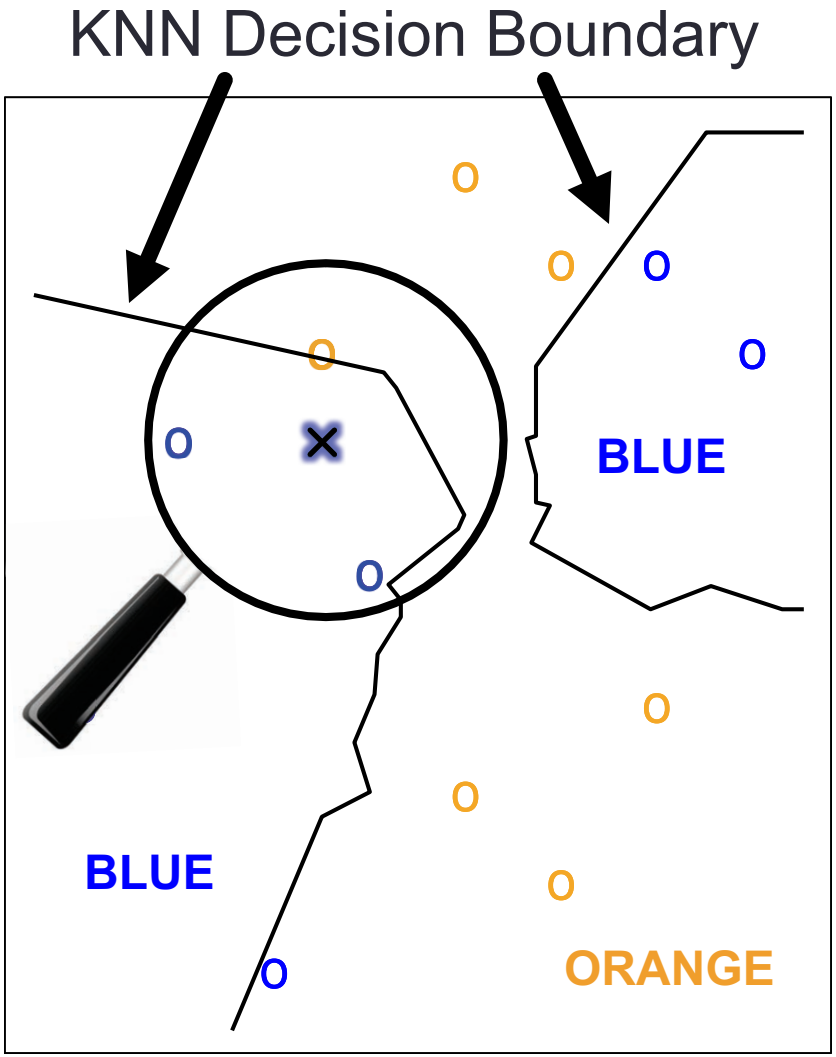
K-nearest neighbors

- **Input**: a positive integer K and a test value x_0
- **Step 1**: Identify the K training examples closest to x_0 (call them N_0)
- **Step 2**: Estimate:

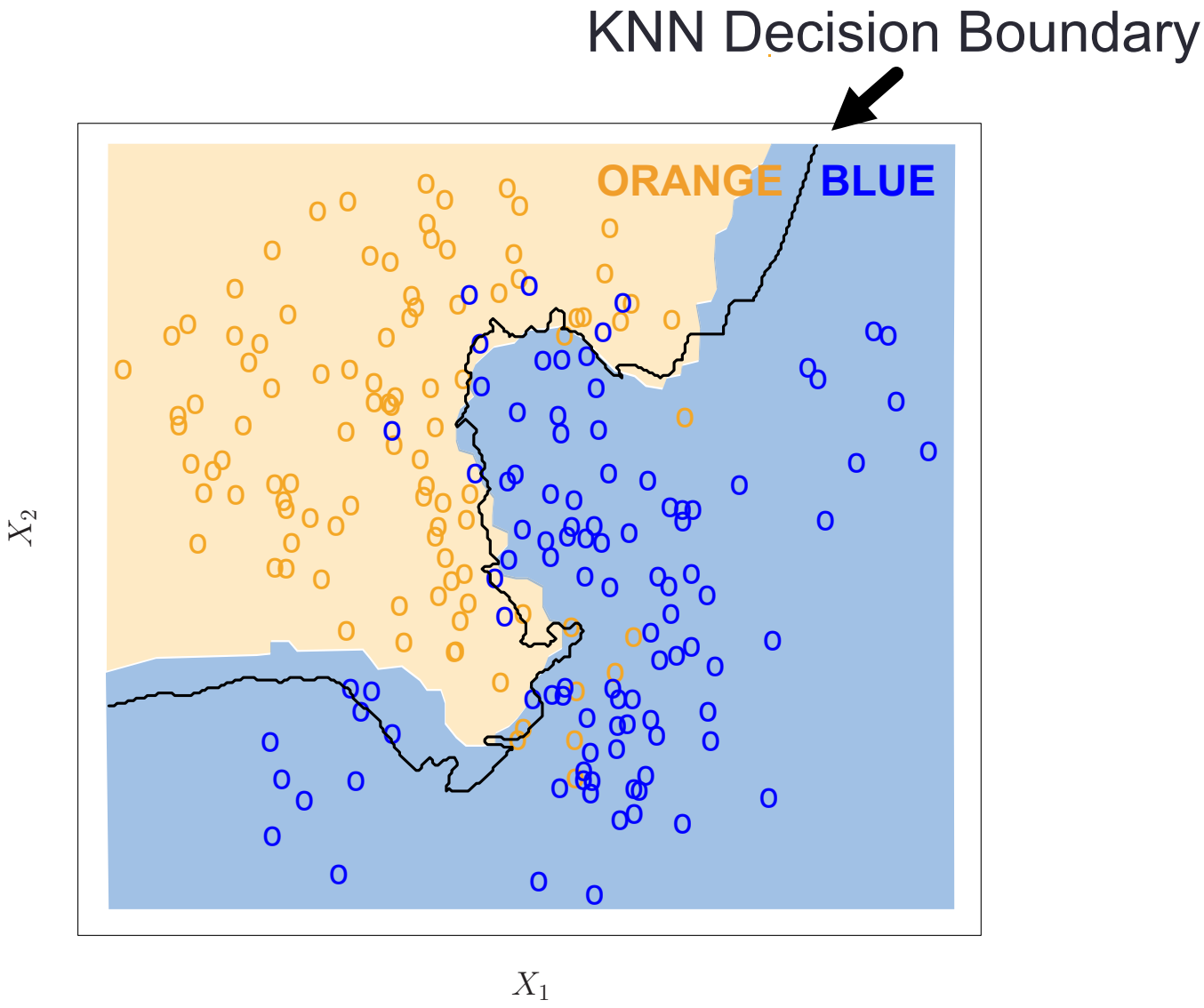
$$\Pr(Y = j \mid X = x_0) \approx \frac{1}{k} \sum_{i \in N_0} I(y_i = j)$$

- **Step 3**: Assign x_0 to whichever class has the highest (estimated) probability

K-nearest neighbors example: K = 3



K-nearest neighbors example: K = 10

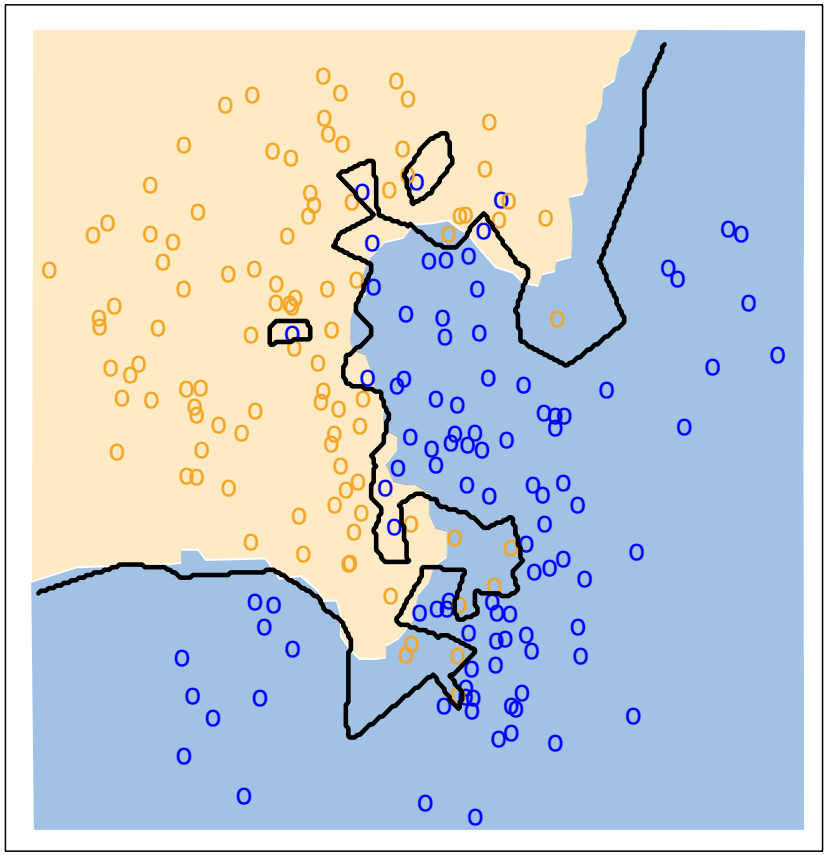


KNN vs. Bayes

- Despite being extremely simple, KNN can produce classifiers that are close to optimal (is this surprising?)
- **Problem:** what's the right K?
- **Question:** does it matter?

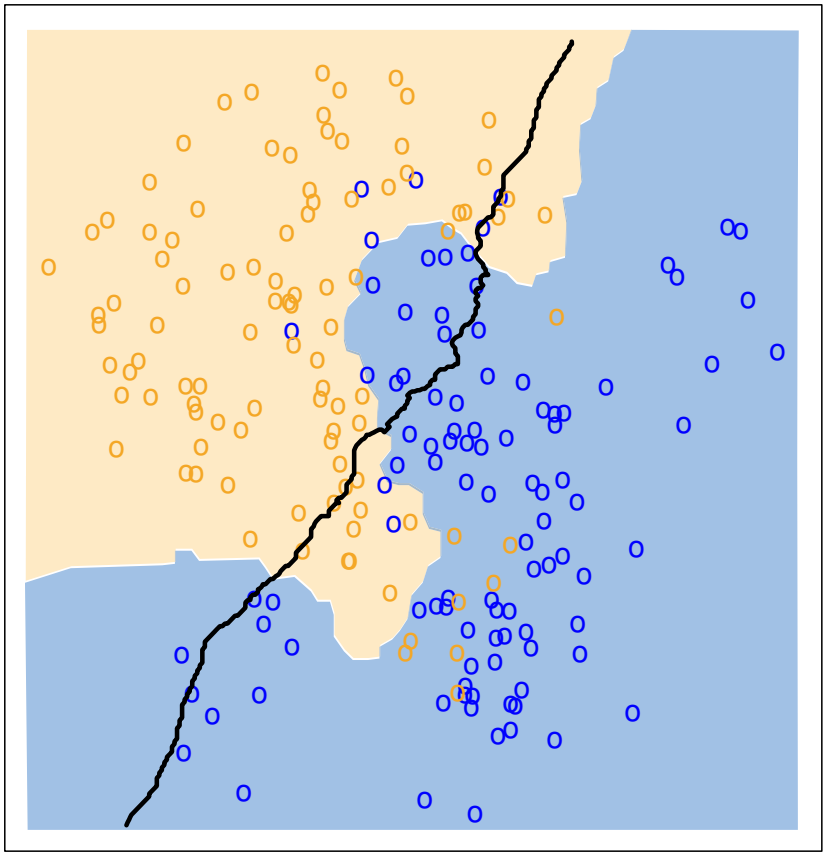
Choosing the right K

KNN: K=1



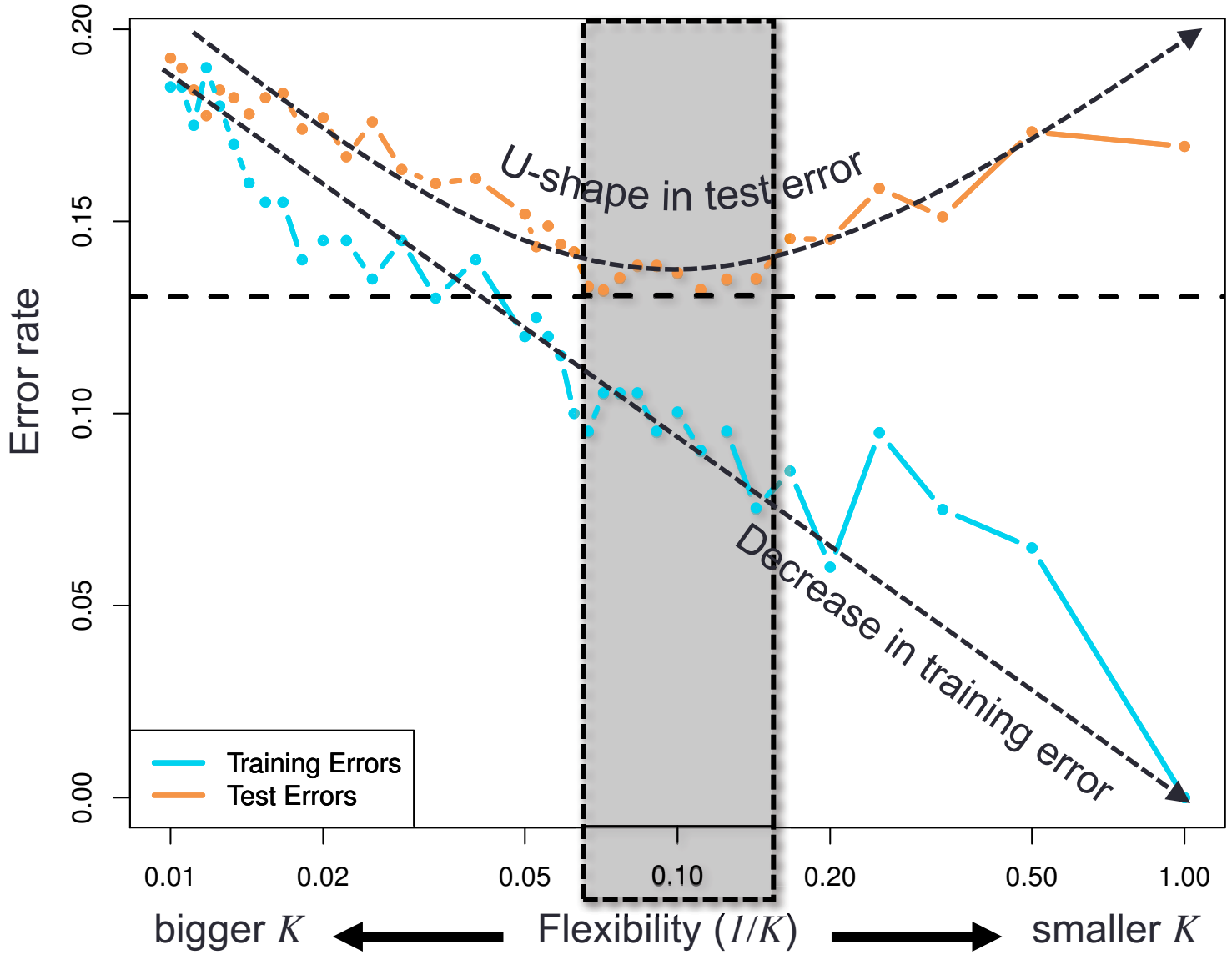
low bias, **high** variance
TE = 0.1695

KNN: K=100



high bias, **low** variance
TE = 0.1925

KNN training vs. test error



Lab: K-nearest neighbors

- To do today's lab in R: `class` package
- To do today's lab in python: `pandas`, `numpy`, `sklearn`
- Instructions and code:

  http://www.science.smith.edu/~jcrouser/SDS293/labs/lab3.html
- Full version can be found beginning on p. 163 of ISLR
- **Note**: we're going a little out of order, so you may want to stick with the demo code

Discussion: KNN on quantitative responses

- **Question 1**: is there any reason we couldn't use KNN to predict quantitative responses?
- **Question 2:** what (if anything) would need to change?

$$\Pr(j \mid x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j) \longrightarrow \hat{f}(x_0) = \frac{1}{k} \sum_{i \in N_0} y_0$$

- **Question 3**: how does it compare to LR?

LR vs. KNN

Linear Regression

- Parametric (meaning?)
  - We assume an underlying functional form for $f(X)$
- Pros:
  - Coefficients have simple interpretations
  - Easy to do significance testing
- Cons:
  - Wrong about the functional form → poor performance

K-Nearest Neighbors

- Non-parametric (meaning?)
  - No explicit assumptions about the form for $f(X)$
- Pros:
  - Doesn't require knowledge about the underlying form
  - More flexible approach
- Cons:
  - Can accidentally "mask" the underlying function

Discussion: which method?

- **Question 1**: would you expect LR to outperform KNN when the underlying relationship **is linear**? Why?
  - **Yes:** KNN won't get a reduction in bias as it increases in variance
- **Question 2:** what happens as the # of dimensions increases, but the # of observations stays the same?
  - **"Curse of dimensionality"**: the more dimensions there are, the farther away each observation's "nearest neighbors" can be

Coming up

- Wednesday: **Logistic regression**
  - Logistic model
  - Estimating coefficients with maximum likelihood
  - Multivariate logistic regression
  - Multiclass logistic regression
  - Limitations

- A1 due **Weds. 9/27 by 11:59pm** (submit using Moodle)