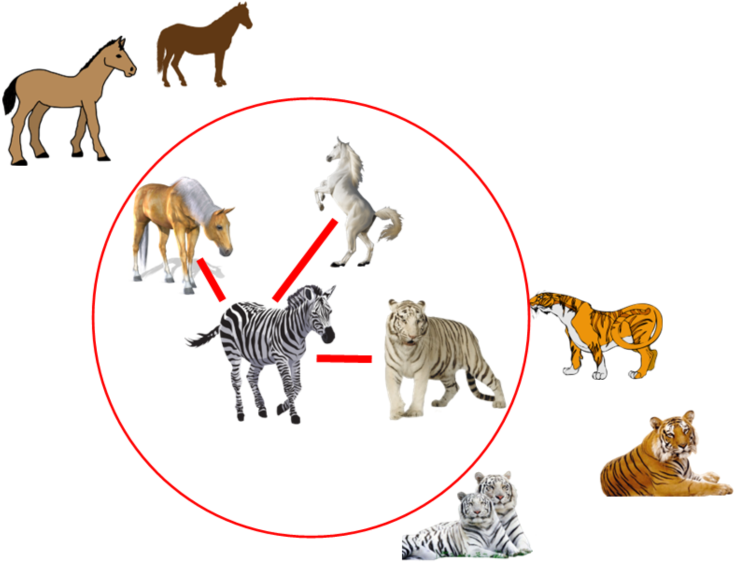
k-nearest neighbors (KNN)

December 02, 2022

Questions

Here are some questions that we should be able to answer by the end of class:

1. What is KNN, and what types of problems is it used for?
2. How does KNN differ from traditional predictive or classification models like logistic regression?
3. How is KNN impacted by samples, neighbors, and features?



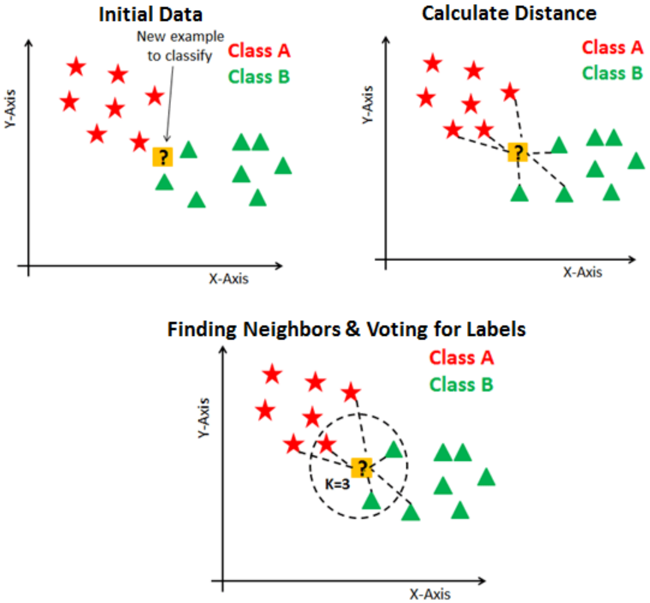
What is KNN?

- Supervised
- Nonparametric
- Model-less (memory-based)
- Used for classification or regression

KNN Algorithm

1. Compute the distance between a new observation and the observations in the dataset
2. Find the k nearest neighbors
3. Calculate the mean outcome (regression) or vote on labels for group membership (classification)

$$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(X)} y_i \qquad \text{or} \qquad \hat{\mathcal{G}}(x) = \underset{\mathcal{G}_k}{\operatorname{argmax}} \; Pr(\mathcal{G}_k | X = x)$$



source: https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn

Distance

A *distance* in any context satisfies a few properties

1. $d(x, x) = 0$ Identity
2. $d(x, y) > 0$ for all $x \neq y$ Non-negativity
3. $d(x, y) = d(y, x)$ Symmetry
4. $d(x, y) \leq d(x, z) + d(z, y)$ Triangle inequality

Examples of distance

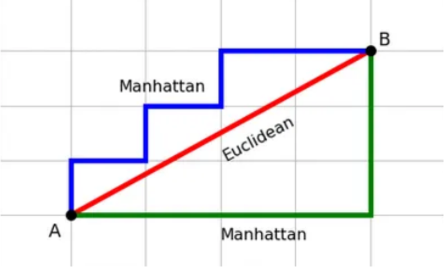
1. Euclidean distance

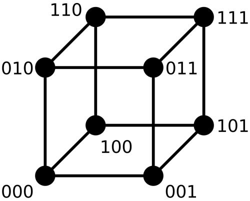
$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}$$

2. Manhattan distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

3. Hamming distance





Types of Data

Special strength when data is irregular and there are many prototypes

Also important to not be worried about inference

Examples of use cases and data:

- Spell check (text strings)
- Google search
- Image and video classification
- Context search and document retrieval

Of course, there is also utility for general regression and classification

Applied example

West et. al (2001) DNA microarray analysis on tumors from 49 breast cancer patients

Want to create genetic profile characterizing tumor type for improved diagnosis and treatment of disease

7,129 genes collected on each tumor

We have interest in predicting patient estrogen receptor (ER) status (+/-) and to predict lympth node involvement (yes/no)

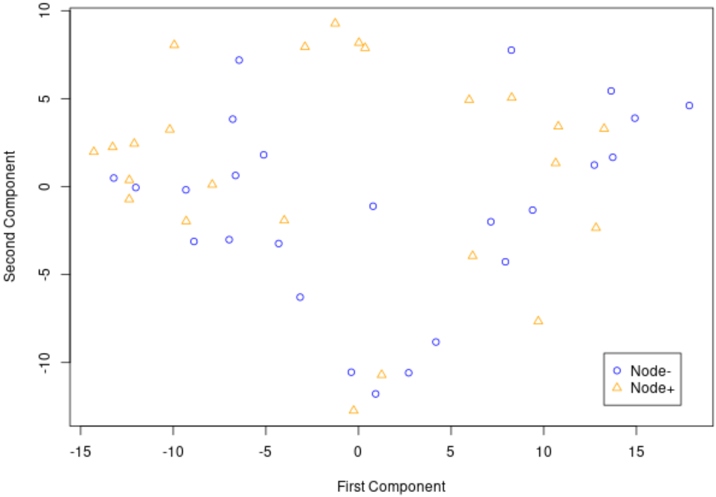
Breast Cancer

Rather than using the entire microarray, we will try to capture relevant features using the first two principal components

Our outcome of interest will be classification – specifically, we will let $\mathcal{G}$ denote membership to the group having lymph node involvement in diagnosis

We begin with a plot of our observations against the first two principal components

Principal Components



Linear Model

One could run a typical linear model of the form

$$\mathrm{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where our solution minimizes the residual sum of squares:

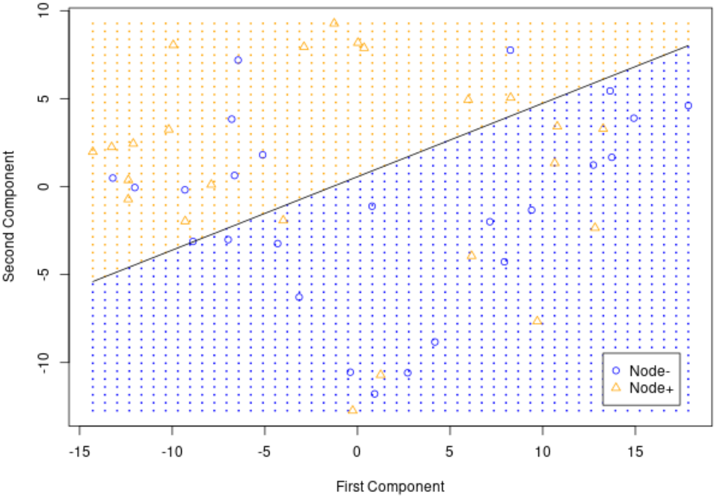
$$\hat{\beta} = (X^T X)^{-1} X^T \mathrm{y}.$$

As our outcome is binary, a decision rule is necessary,

$$\hat{\mathcal{G}} = \begin{cases} \text{Node+} & \hat{Y} > 0.5 \\ \text{Node-} & \hat{Y} \leq 0.5 \end{cases}.$$

This creates what is known as a decision boundary

3



10
5
0
-5
-10
Second Component
-15
-10
-5
0
5
10
15
First Component
○ Node-
△ Node+

KNN Algorithm

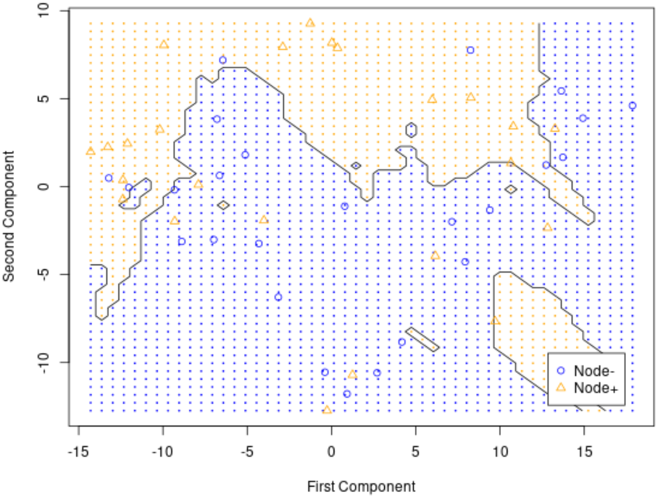
Alternatively, we can use the KNN algorithm

Here, at each grid point we have determined the nearest five neighbors and computed the proportion in each group

We will use the same classification boundary as before, though we note that the boundary is more irregular

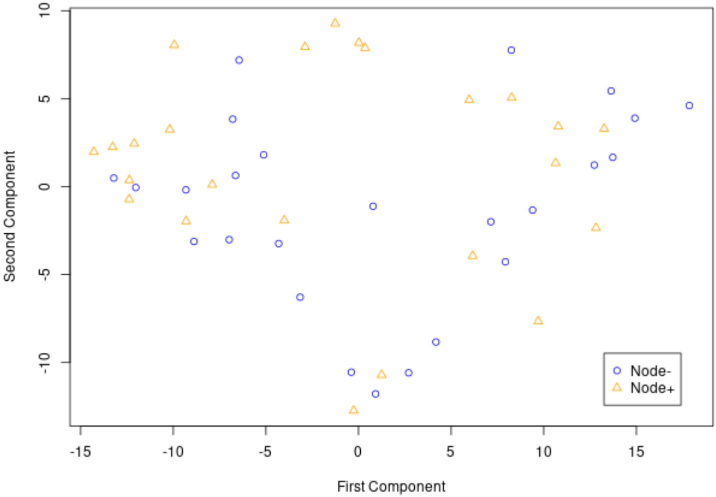
$$\hat{\mathcal{G}} = \begin{cases} \text{Node+} & \hat{Y} > 0.5 \\ \text{Node-} & \hat{Y} \leq 0.5 \end{cases} .$$

KNN, $k = 5$

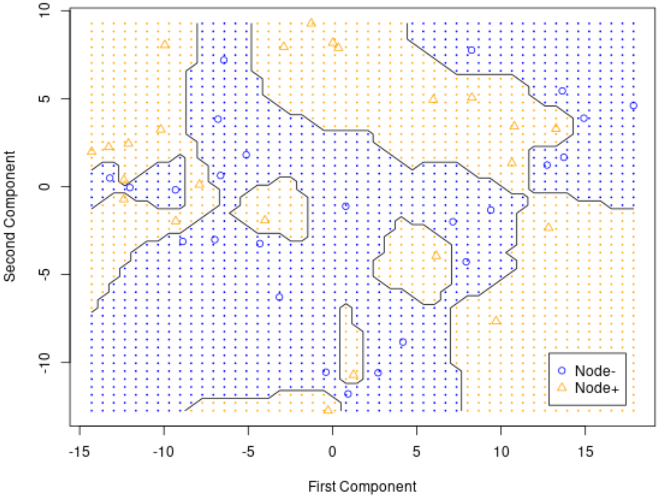


10
5
0
-5
-10
Second Component
-15
-10
-5
0
5
10
15
First Component
○ Node-
△ Node+

KNN, $k = 1$

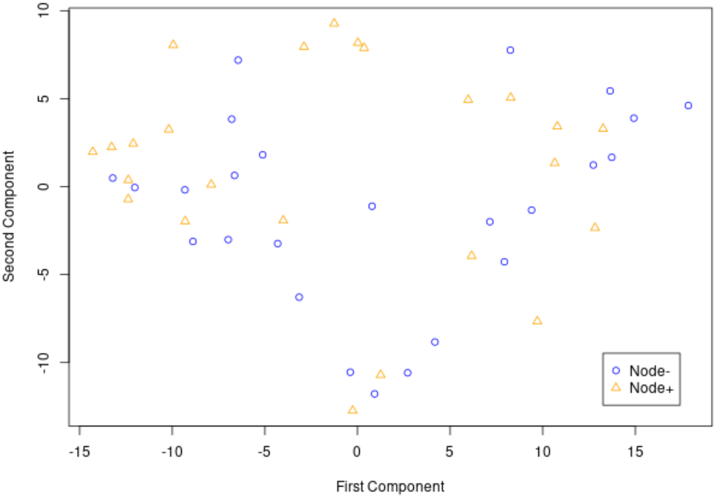

10
5
0
-5
-10
Second Component
-15
-10
-5
0
5
10
15
First Component
○ Node-
△ Node+

KNN, $k = 1$



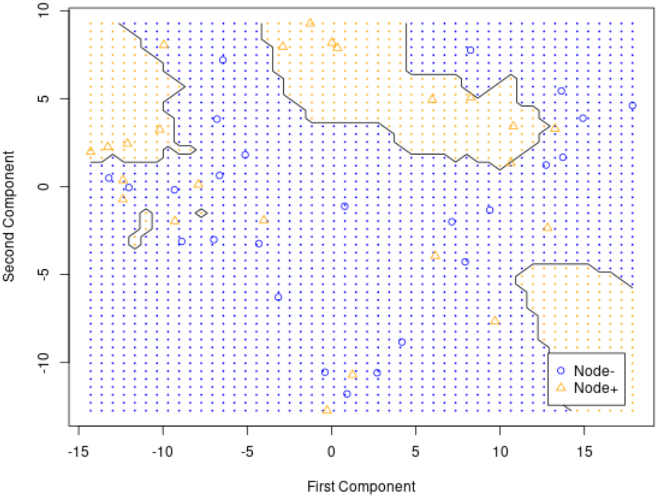
Second Component
10
5
0
-5
-10
-15
-10
-5
0
5
10
15
First Component
○ Node-
△ Node+

KNN, $k = 2$

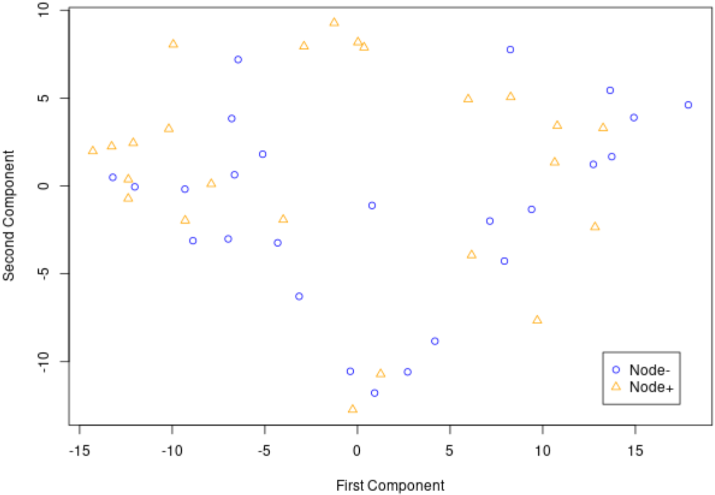


Second Component
10
5
0
-5
-10
-15
-10
-5
0
5
10
15
First Component
○ Node-
△ Node+

KNN, $k = 2$

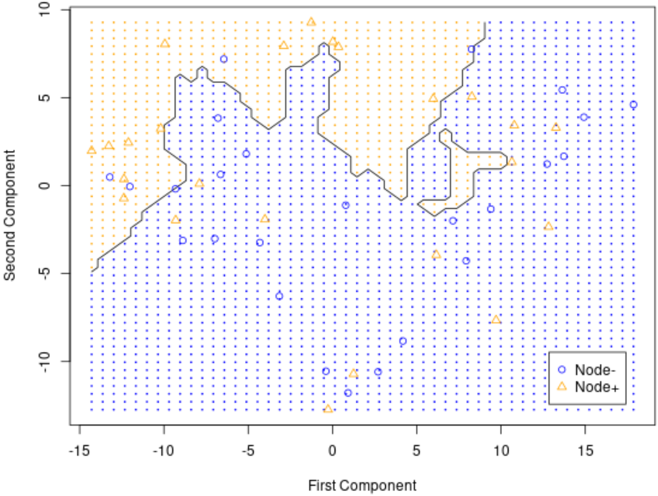

10
5
0
-5
-10
Second Component
-15
-10
-5
0
5
10
15
First Component
○ Node-
△ Node+

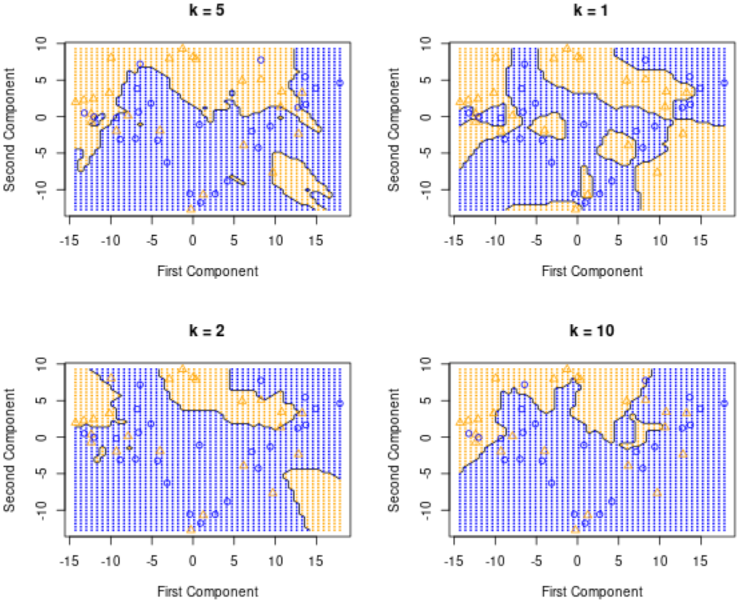
KNN, $k = 10$



10
5
0
-5
-10
Second Component
-15
-10
-5
0
5
10
15
First Component
○ Node-
△ Node+

KNN, $k = 10$


Second Component
10
5
0
-5
-10
-15
-10
-5
0
5
10
15
First Component
○ Node-
△ Node+


k = 5
k = 1
k = 2
k = 10
Second Component
First Component
10
5
0
-5
-10
-15
-10
-5
0
5
10
15

Where are we so far?

Great for prediction, poor for inference

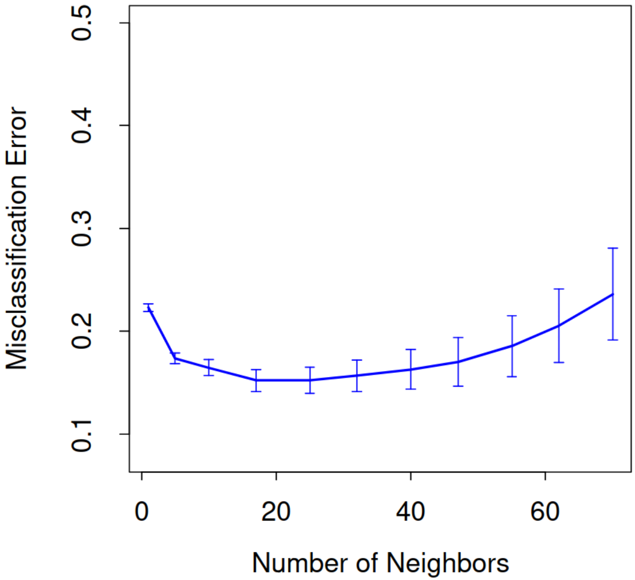
Model-less:

- No preprocessing
- No data reduction either
- Simple to add new observations

What else do we need to know?

- How many neighbors
- Feature selection

How many neighbors?



The bias-variance tradeoff

The predictive function minimizing squared error loss of estimating outcome Y given X is

$$f(x) = E(Y|X = x)$$

If multiple x_i observed at a given point, conditional expectation is average,

$$\hat{f}(x) = \text{Ave}(y_i|x_i = x)$$

More typically, there are *no* other observations at a given x, and we are forced to use points nearby:

$$\hat{f}(x) = \text{Ave}(y_i|x_i \in N_k(x)).$$

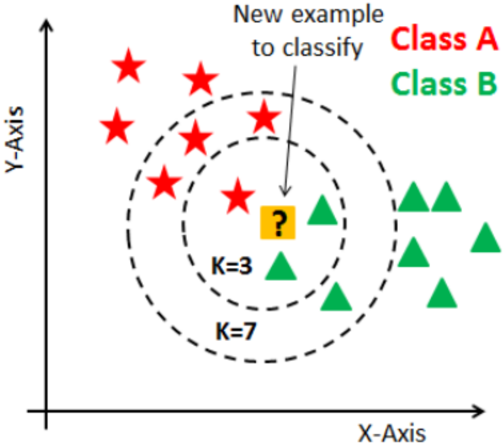
The bias-variance tradeoff

There are two things here affecting the accuracy and precision:

1. The expectation is being approximated by a sample average
2. This average is being conditioned over a region rather than a target point

Consequently, as k increases, the predictor becomes more precise (less variance)

At the same time, as the region moves further from the target x, the predictor becomes more biased



source: https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn

Feature selection

Feature selection for KNN is not dissimilar from other ML algorithms:

- Univariate filtering (variability, correlation)
- Cross-validation
- Principal components
- Clinically relecant

There are a few things to watch out for, though

- Computational issues
- Noisy or irrelevant features
- Curse of dimensionality

Questions

1. What is KNN, and what types of problems is it used for?
2. How does KNN differ from traditional predictive or classification models like logistic regression?
3. How is KNN impacted by samples, neighbors, and features?

60 second survey

- What is one takeaway from this class today?
- What is one thing that is still not very clear?

Sources

Hastie, T., Tibshirani, R., Friedman, J. (2009) *The Elements of Statistical Learning* New York: Springer

Brian Smith BIOS:6720 course notes, Section 5, Spring 2018

Brian Smith BIOS:6720 course notes, Section 11, Spring 2018