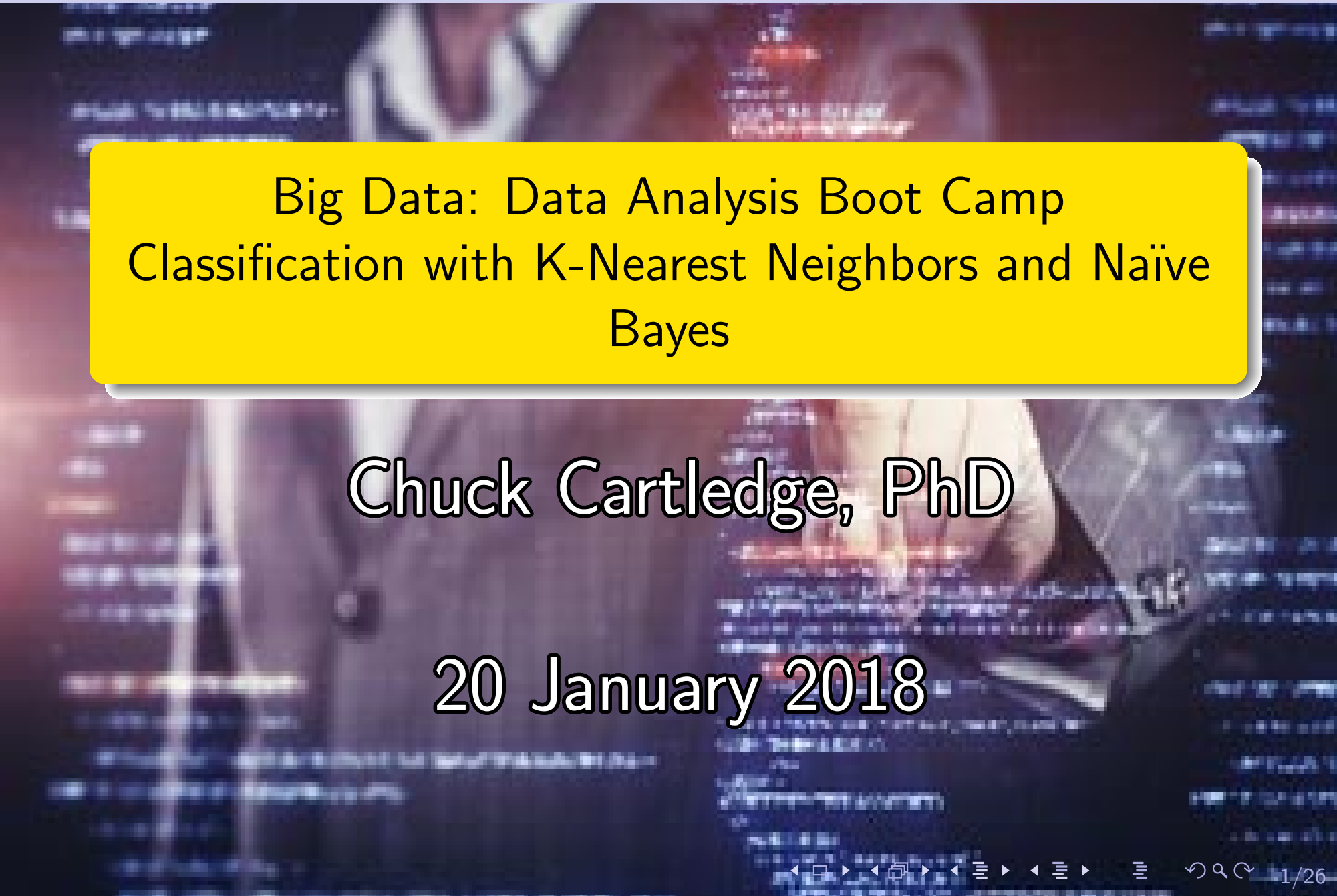


Big Data: Data Analysis Boot Camp
Classification with K-Nearest Neighbors and Naïve Bayes

Chuck Cartledge, PhD

20 January 2018

Table of contents (1 of 1)

What are we going to cover?

We're going to talk about:

- K-nearest neighbors (knn) as a clustering technique.
- Naïve Bayes classifier as a simple way to cluster data.

How does classification work?

Simply:

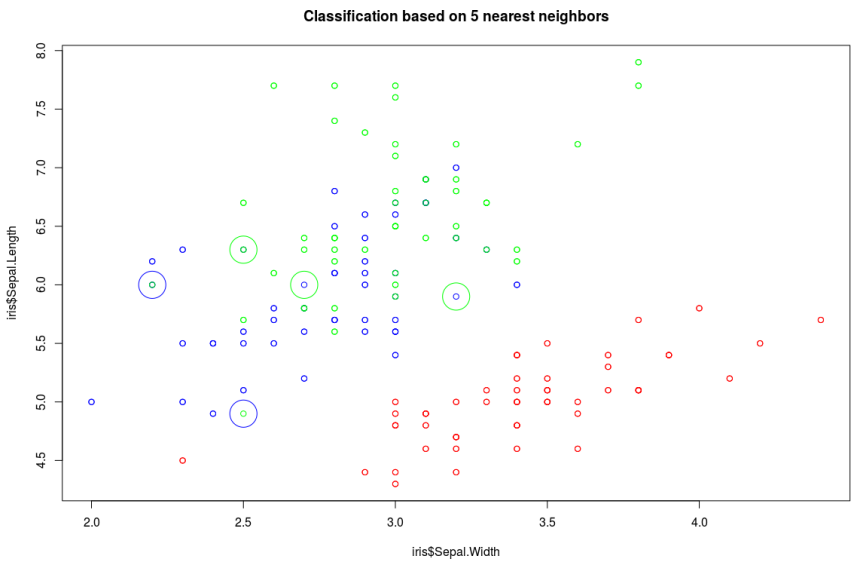
1. A known collection data points have some collection of attributes, including a label
2. The distance between data points can be computed (different distance computations will result in different neighbor sets)[2]
3. A new data point is added to the system
4. The labels of k nearest neighbors are used to assign a label to the new point

(Same approach can be used for regression. Instead of assigning a label, a weight can be assigned.)

Our friend iris data.

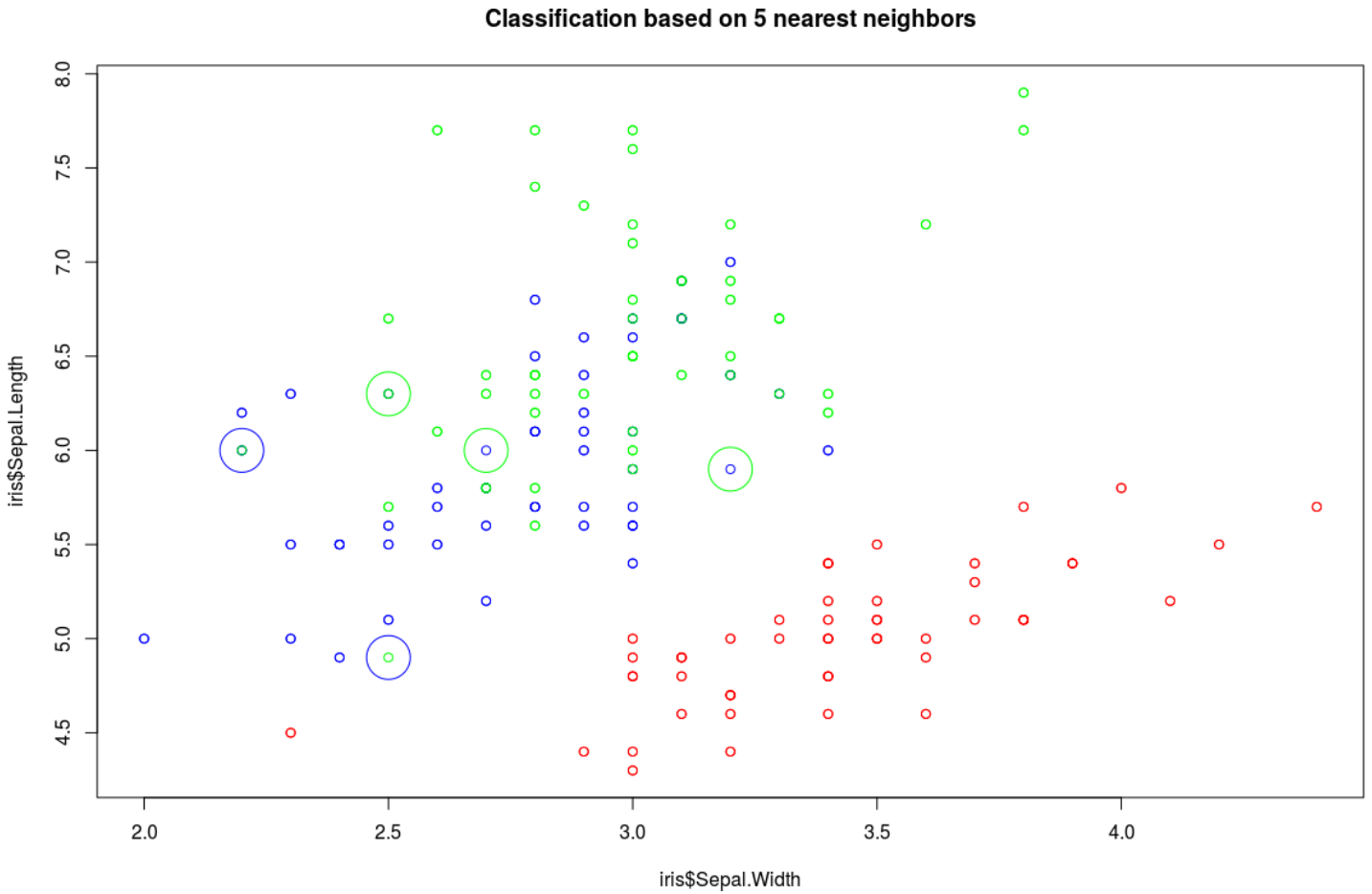
The attached file `chapter-10-classification-revised.R` expands on the code in the text.

- A manual $k - nn$ implementation is compared to a built-in one.
- Different k values result in different classifications. `main(3)`
- Differences between the classification and truth are highlighted.



Attached file.

Same image.

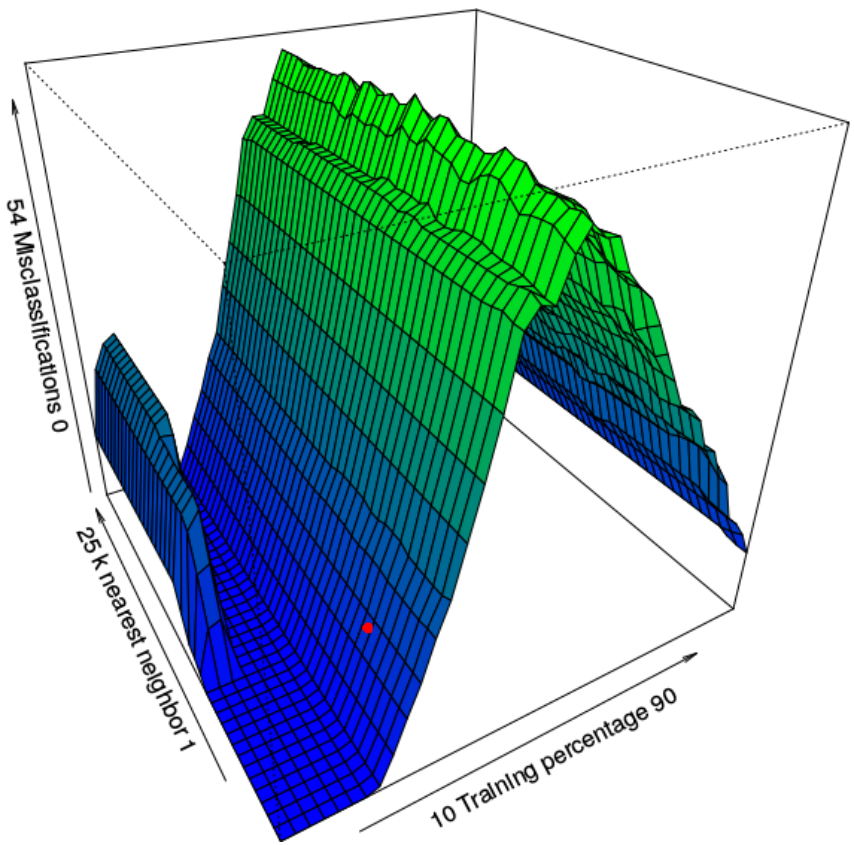


Attached file.

How well does k-nn do?

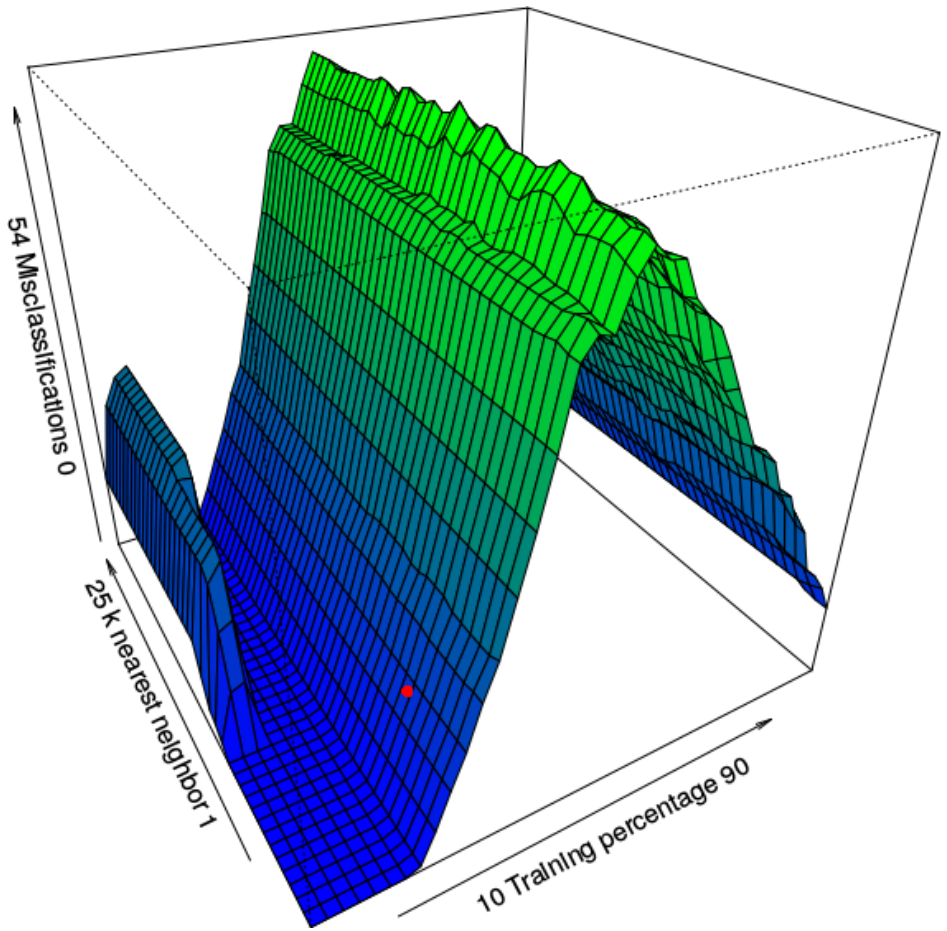
We can vary both training percentage and k to test the k-nn algorithm and evaluate the results.
The red dot represents:

- 30% training
- 5 k-nn processing
- 10 misclassificaitons



Attached file.

Same image.



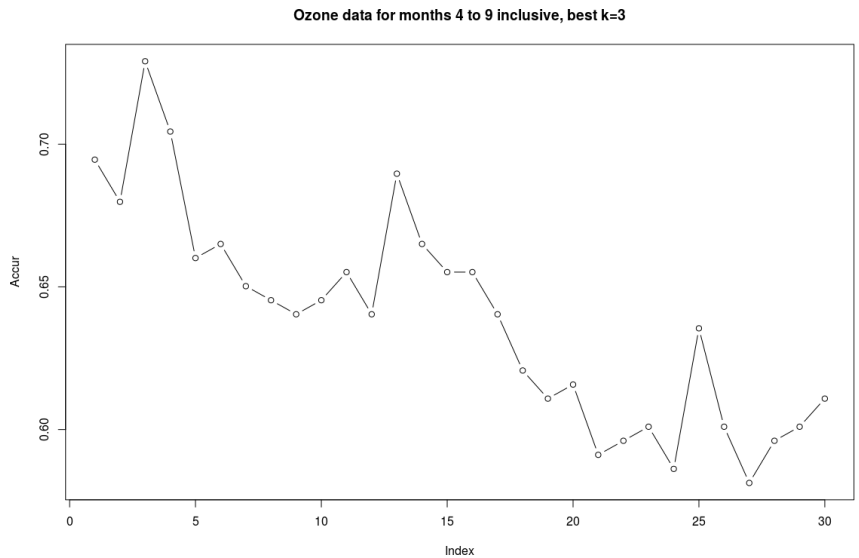
Attached file.

Looking at ozone data.

Help file claims that data comes from [1]. Not able to see data.

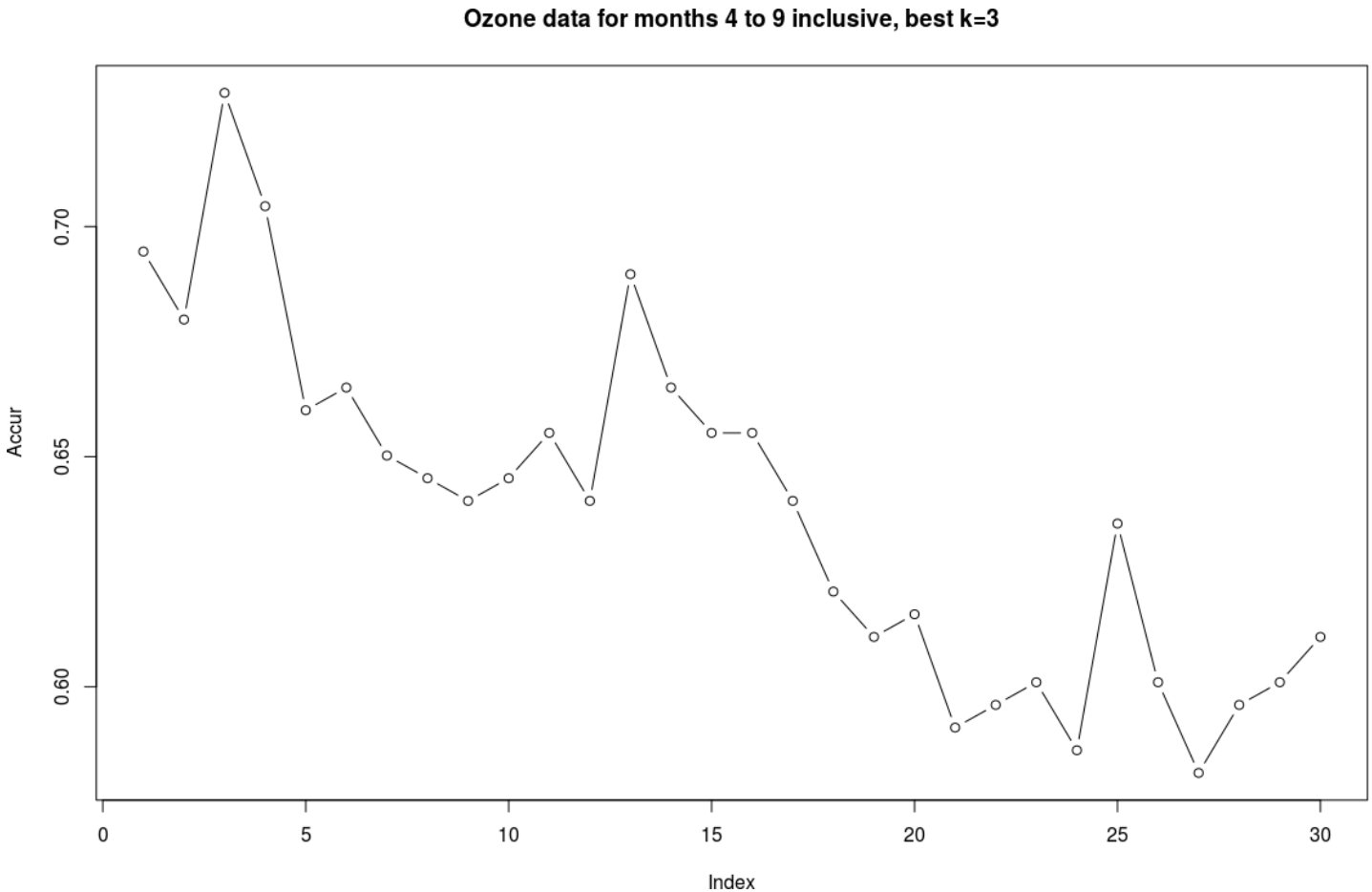
- k was varied from 1 to 30
- Accuracy was computed as percentage of assigned classifications that matched original data
- The highest accuracy was selected to determine the best k

Attached file (`chapter-10-ozone-revised.R`) replicates and expands on text approach.



Attached file.

Same image.



Attached file.

What is naïve Bayes?[3] (1 of 2)

"Bayes theorem named after Rev. Thomas Bayes. It works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge."

R. Saxena [3]

What is naïve Bayes? (2 of 2)

Baye's theorem is:

$$P(H \mid E) = \frac{P(E|H) * P(H)}{P(E)}$$

Where:

- $P(H)$ is the probability of hypothesis H being true. This is known as the prior probability.
- $P(E)$ is the probability of the evidence (regardless of the hypothesis).
- $P(E \mid H)$ is the probability of the evidence given that hypothesis is true.
- $P(H \mid E)$ is the probability of the hypothesis given that the evidence is true.

Ultimately naïve Bayes predicts membership of a class.

> *"Naïve Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class."*
>
> *R. Saxena [3]*

The implication is that you have to compute the class membership probability for all classes to identify the most likely membership.

Load the `chapter-10-bayes.R` file.

Things to notice:

- **Apriori probabilities:** are the probabilities that DiseaseZ column 10, for rows 1 through 10 are NO or YES
- **Conditional probabilities:** is a confusion matrix for the named column and DiseaseZ column 10. The table's diagonal is TN and TP.

Long hand probabilities for person #11 (pg. 184 from the text):

P(DiseaseZ == YES)	=	0.00908
P(DiseaseZ == NO)	=	0.00175

Person #11 is $\frac{0.00908}{0.00175} \approx 5$ time more likely to have DiseaseZ than to not have it.

Looking at the datasets Titanic results.

- Half of the data was used for training, half for testing.
- The Bayes model created from the training data was used to predict survival of the testing data
- The results aren't encouraging.

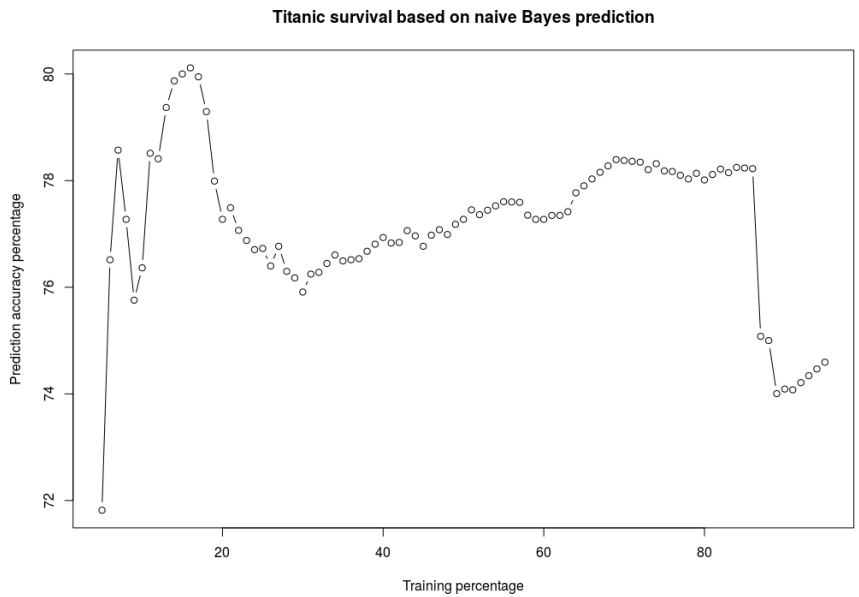
The Titanic predictions are 76% accurate.

How well does naïve Bayes do?

Wrap the text's code in a loop, varying the training percentage from low to high.

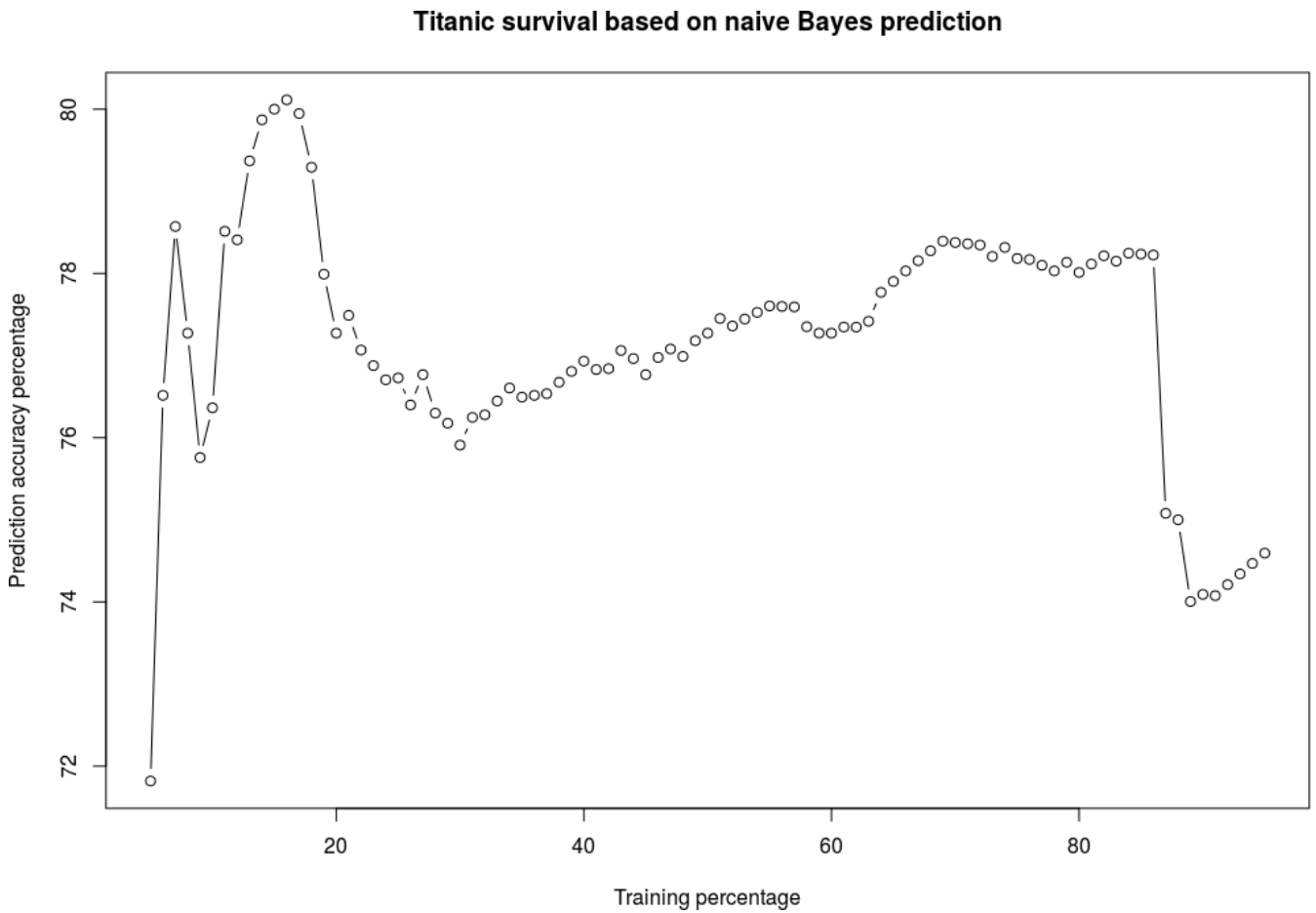
- Percentage < 17 is under fitting.
- Percentage > 80 is over fitting.
- Percentage between 25ish to 80ish is OK.
- Percentage between 18 and 19 is best.

Hard to know best without checking all.



Attached file.

Same image.



Attached file.

Some simple exercises to get familiar with Naïve Bayes

The HouseVotes84 is an R dataset[4] tabulating how members of the US House of Representatives voted on a series of issues. The dataset is in the mlbench library.

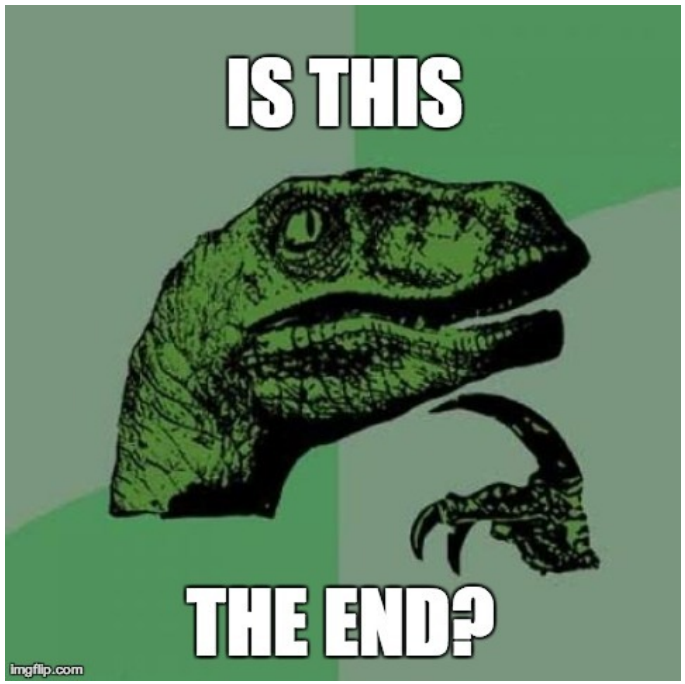
- Create a naïve Bayes model that will predict party membership based on education spending given that they voted for a synthetic fuels cutback.
- Quantify how accurate your model is.

Q & A time.

"Women and cats will do as they please, and men and dogs should relax and get used to the idea."
Robert Heinlein

What have we covered?

- Explored k-nn using the iris and ozone datasets
- Identified strengths and limitations of k-nn
- Worked with naïve Bayes classifier on artificial and Titanic datasets



Next: LPAR Chapter 11, classification trees

References (1 of 2)

[1] Leo Breiman and Jerome H. Friedman,
Estimating Optimal Transformations for Multiple Regression and Corr
Journal of the American Statistical Association **80** (1985),
no. 391, 580–598.

[2] Michel Marie Deza and Elena Deza, Encyclopedia of Distances,
Encyclopedia of Distances, Springer, 2009, pp. 1–583.

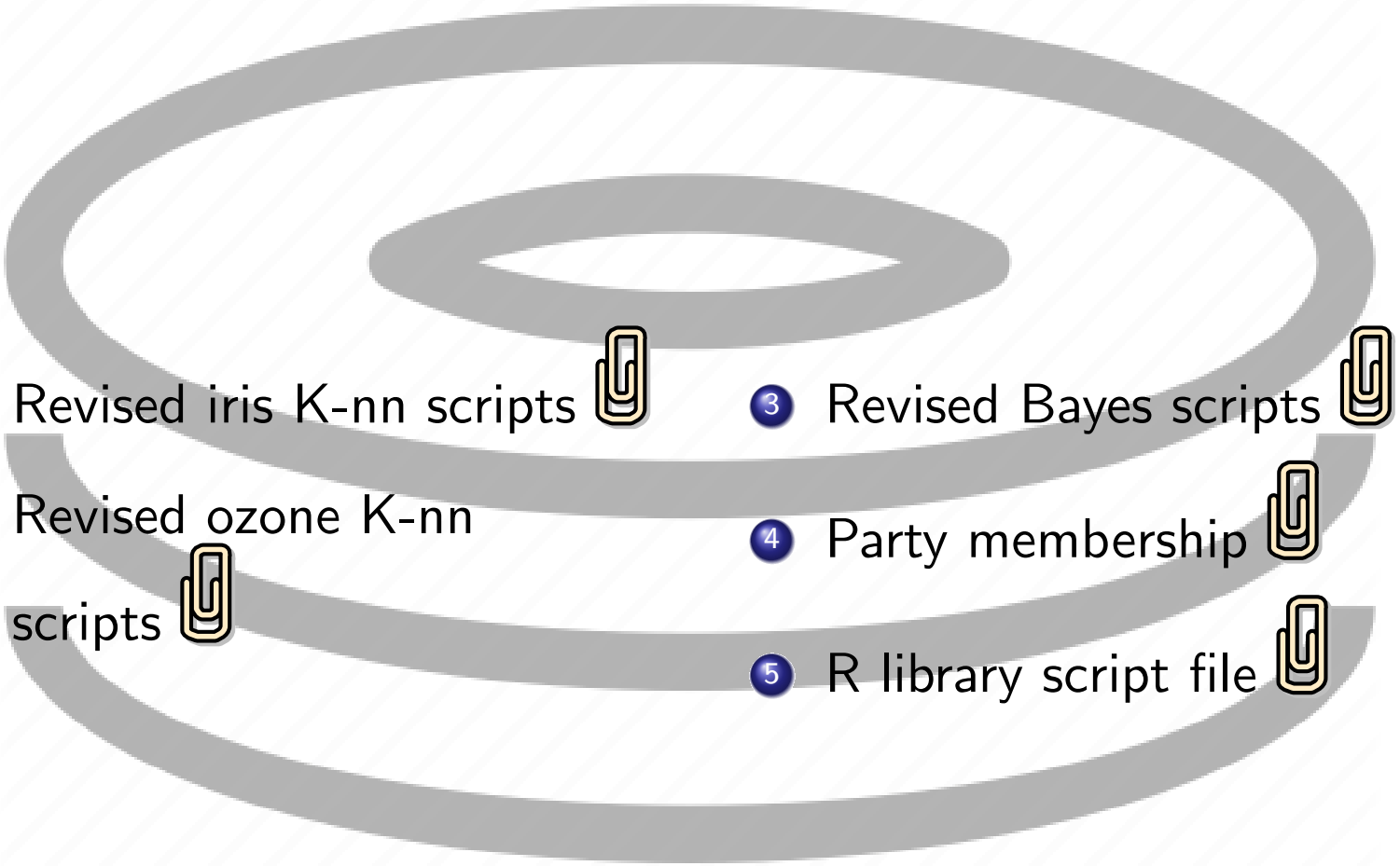
[3] Rahul Saxena,
How the Naive Bayes Classifier works in Machine Learning,
`http://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/`, 2017.

[4] UCI Staff, Congressional Voting Records Data Set,
`https://archive.ics.uci.edu/ml/datasets/congressional+voting+records`, 2017.

References (2 of 2)

[5] Wikipedia Staff, Confusion matrix, https://en.wikipedia.org/wiki/Confusion_matrix, 2017.

Files of interest



1. Revised iris K-nn scripts
2. Revised ozone K-nn scripts
3. Revised Bayes scripts
4. Party membership
5. R library script file

Confusion matrix (1 of 2)[5]

Some definitions:

- TP True Positive – hit
- TN True Negative – correct rejection
- FP False Positive – false alarm (Type I error)
- FN False Negative – miss (Type II error)

$$TPR \text{ (sensitivity, recall, hit rate, or true positive rate)} = \frac{TP}{TP + FN}$$

$$TNR \text{ (specificity or true negative rate)} = \frac{TN}{TN + FP}$$

$$PPV \text{ (precision or positive predictive value)} = \frac{TP}{TP + FP}$$

$$NPV \text{ (negative predictive value)} = \frac{TN}{TN + FN}$$

$$FNR \text{ (miss rate or false negative rate)} = 1 - TPR$$

Confusion matrix (2 of 2)[5]

$$
\begin{aligned}
FPR \text{ (fall-out or false positive rate)} &= 1 - TNR \\
FDR \text{ (false discovery rate)} &= 1 - PPV \\
FOR \text{ (false omission rate)} &= 1 - NPV \\
ACC \text{ (accuracy)} &= \frac{TP + TN}{TP + TN + FP + FN} \\
F_1 \text{ (F1 score)} &= \frac{2TP}{2TP + FP + FN}
\end{aligned}
$$

Confusion matrix

		predicted condition			
	total population	prediction positive	prediction negative	Prevalence $= \frac{\Sigma \text{ condition positive}}{\Sigma \text{ total population}}$	
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection $= \frac{\Sigma \text{ TP}}{\Sigma \text{ condition positive}}$	False Negative Rate (FNR), Miss Rate $= \frac{\Sigma \text{ FN}}{\Sigma \text{ condition positive}}$
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**	False Positive Rate (FPR), Fall-out, Probability of False Alarm $= \frac{\Sigma \text{ FP}}{\Sigma \text{ condition negative}}$	True Negative Rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ condition negative}}$
	Accuracy $= \frac{\Sigma \text{ TP} + \Sigma \text{ TN}}{\Sigma \text{ total population}}$	Positive Predictive Value (PPV), Precision $= \frac{\Sigma \text{ TP}}{\Sigma \text{ prediction positive}}$	False Omission Rate (FOR) $= \frac{\Sigma \text{ FN}}{\Sigma \text{ prediction negative}}$	Positive Likelihood Ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False Discovery Rate (FDR) $= \frac{\Sigma \text{ FP}}{\Sigma \text{ prediction positive}}$	Negative Predictive Value (NPV) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ prediction negative}}$	Negative Likelihood Ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Image from [5].