CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Classifiers

Classification

- Classification involves taking a set of unlabeled data points and labeling them in some fashion
- Why?
  - To learn from the classification/data
  - To discover patterns
  - Automate some process, ie handwriting recognition

Classification

- What are the problems it (classification) can solve?
- What are some of the common classification methods?
- Which one is better for a given situation? (meta classifier)

Classification Examples

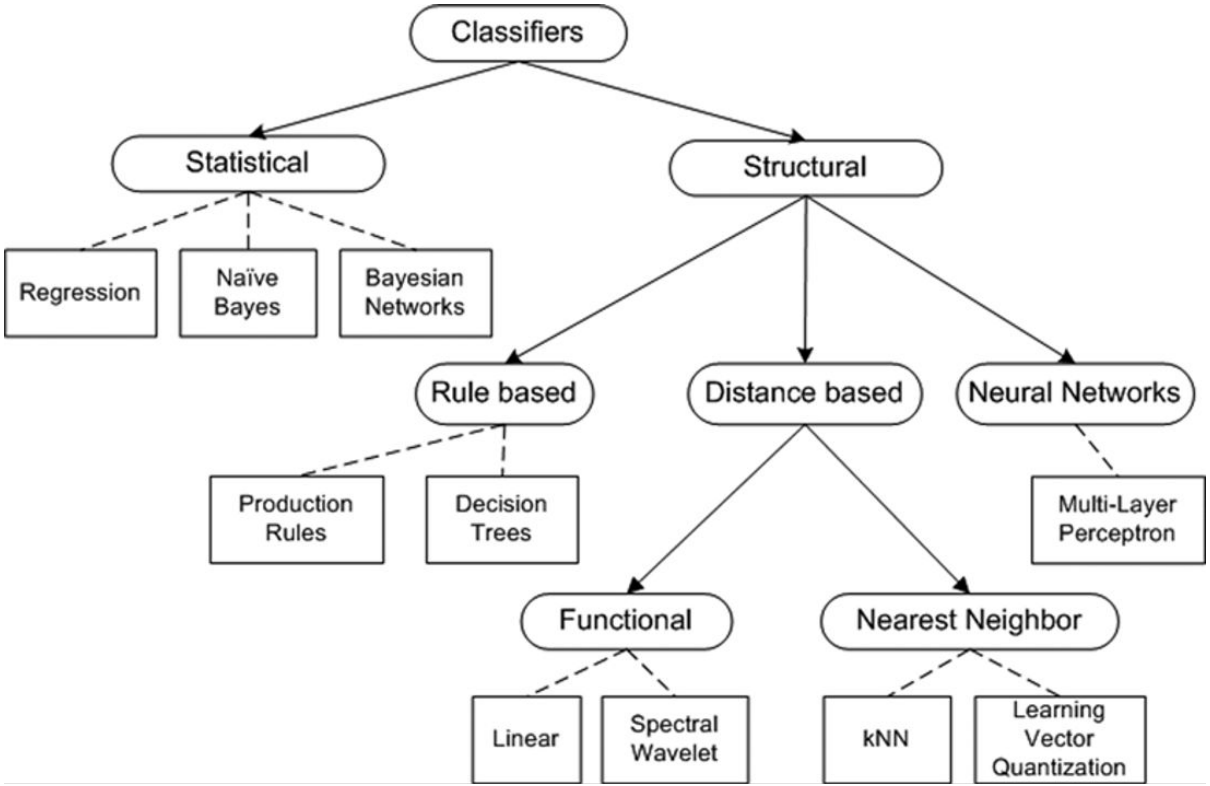
- Restaurant menu: appetizers, salads, soups, entrée, dessert, drinks, …
- Library of congress (LIC) system classifies books according to a standard scheme
- Injury and disease classification in healthcare
- Classification of all living things: eg., Home Sapiens (genus, species)
- Classification across a variety of aspects in the automobile domain from services (classes), parts (classes), incidents (classes) etc.

Classification of Classification Algorithms

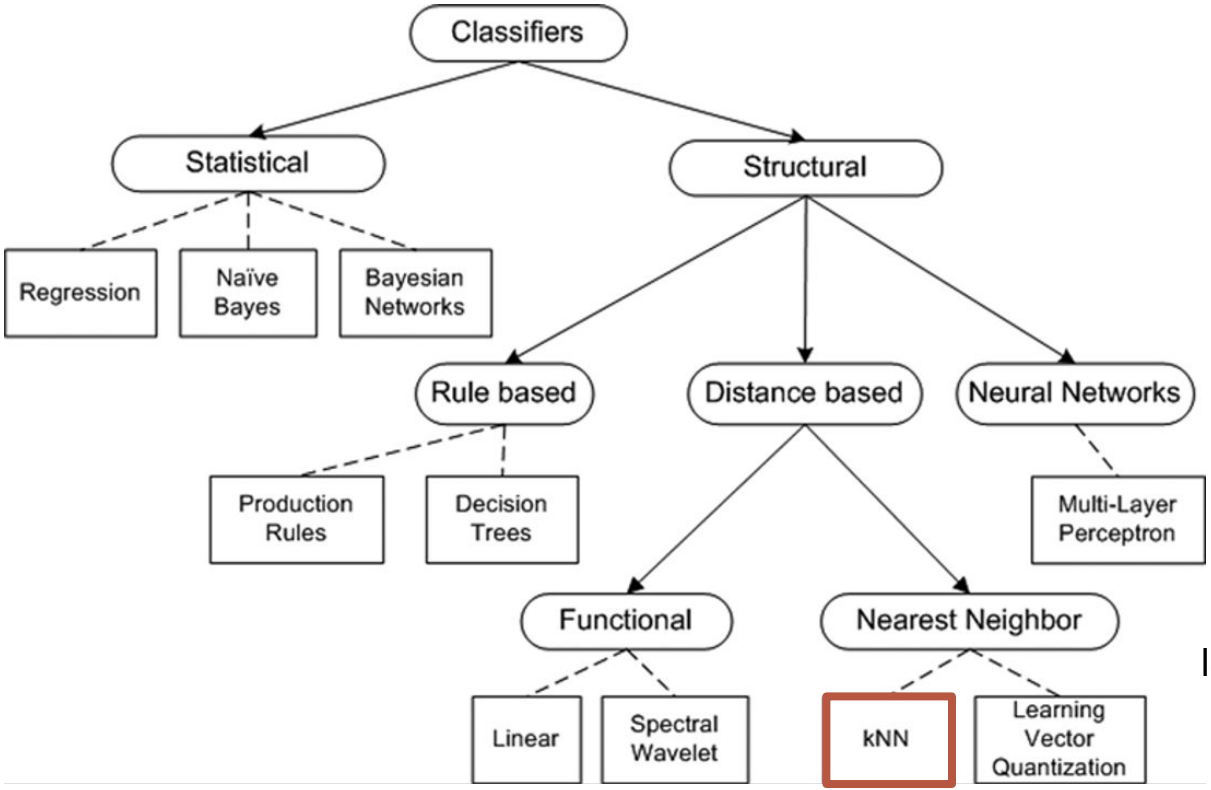
Classification algorithms can be divided into two broad categories:

- Statistical algorithms
  - Regression
  - Probability based classification: Bayes
- Structural algorithms
  - Rule-based algorithms: if-else, decision trees
  - Distance-based algorithm: nearest neighbor
  - Neural networks

Classification of Classification Algorithms

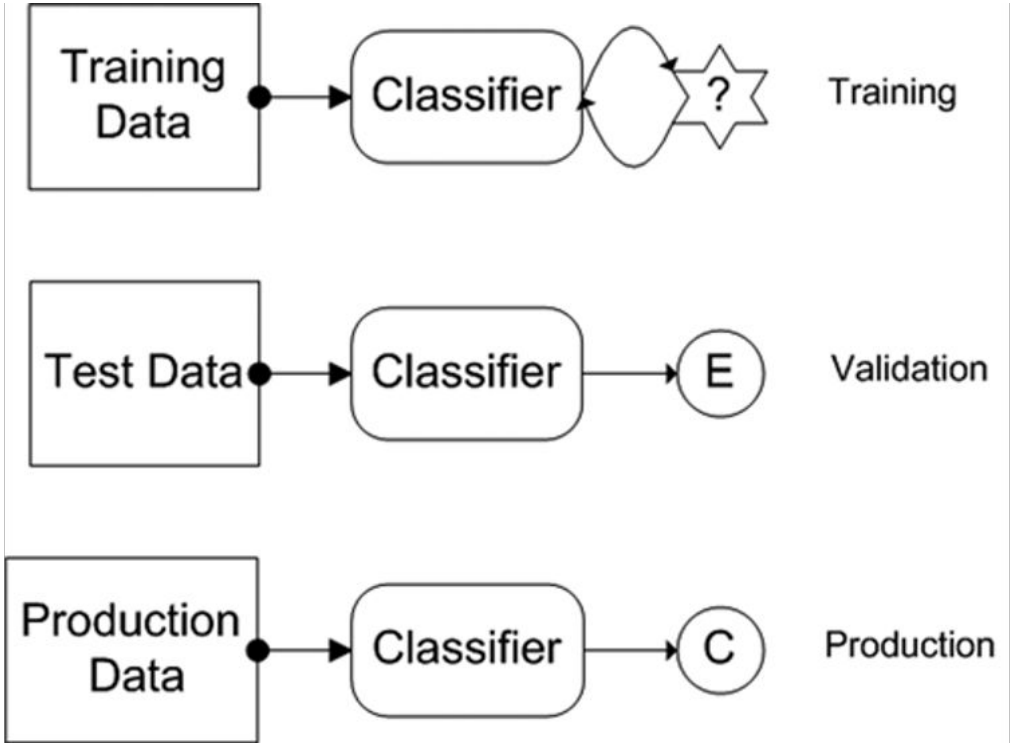


Classification of Classification Algorithms



Today we'll start learning about kNN

Life Cycle of Classifiers



Training Stage

- Provide classifier with data points for which we have already assigned an appropriate class
- Purpose of this stage is to determine the parameters of our model

Validation Stage

- In the validation stage we validate the classifier to ensure credibility
- Primary goal of this stage is to determine the classification errors
- Quality of the results should be evaluated using various metrics
- Training and testing stages ***may be repeated several times*** before a classifier transitions to the production stage

Production Stage

- Now our classifier(s) are ready for use in a live production system
- We can enhance the results by allowing human-in-the-loop feedback

All steps are repeated as we get more data from the production system.

k-Nearest Neighbors (k-NN)

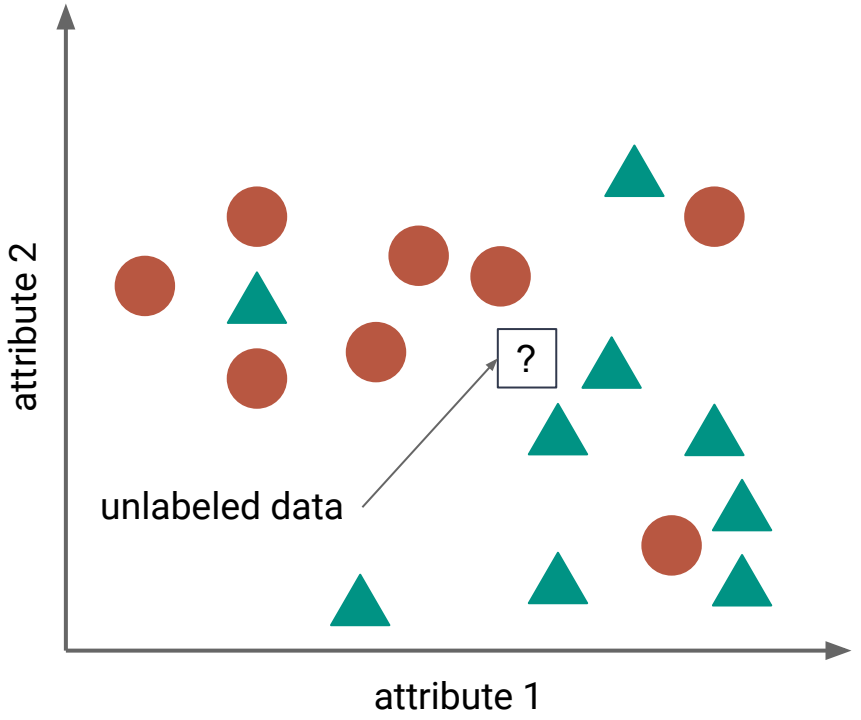
- Algorithm used to ***classify*** or label objects/data points
  - You start with some already labeled data points
  - Uses proximity to make classification.
  - Goal is to be able to automatically label a new set of unlabeled points
- Examples could be: "Good" or "Bad" credit score, political affiliation, star rating of a restaurant, at risk for illness, etc.
- Would linear regression work for this?
  - ...maybe, but it depends on what you are doing
  - Not all data can be easily mapped to continuous scale

k-Nearest Neighbors (k-NN)

- Intuition: For a given unlabeled element, look at just the k *most similar* elements in the labeled dataset based on various attributes, and choose the label that most of those elements have
  - ie: Look at movies with similar runtime, budget, genre, actors, awards to label a movie as good or bad
  - ie: Look at people with similar height, weight, age, gender, to determine if a person is at risk or not for a certain disease

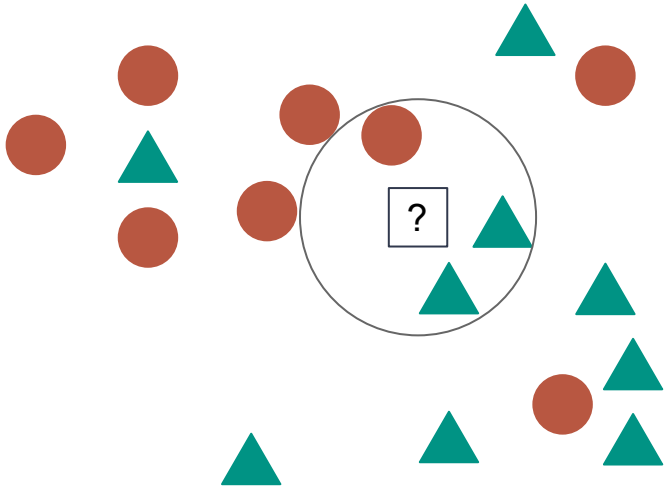
A Simple Example

- For the example to the left, we have a number of data points labeled as either red circles, or green triangles
- How do we label the new unknown data point?
- Depends on the value of k



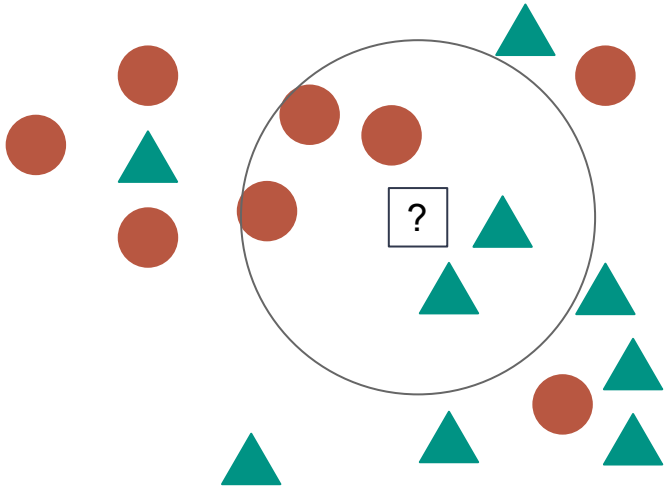
A Simple Example

- If k = 3:
  - Green triangles have 2 votes
  - Red circles have 1 vote
  - The new point will be labeled green triangle



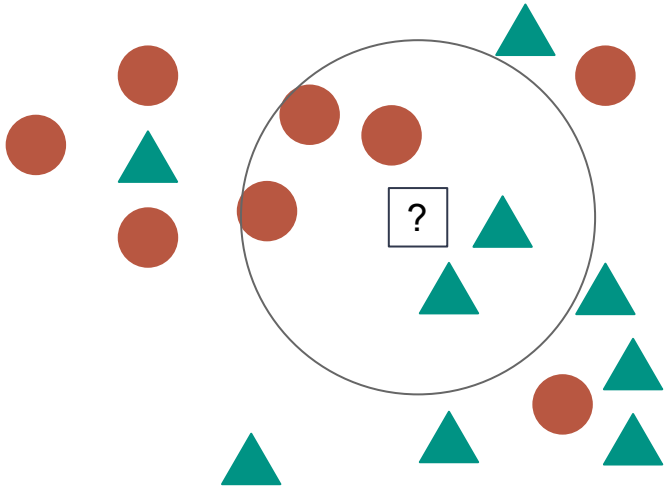
A Simple Example

- If k = 5:
  - Green triangles have 2 votes
  - Red circles have 3 votes
  - The new point will be labeled red circle



A Simple Example

- In order to apply our intuition to real data we need to:
  - Determine how we measure *closeness*
  - Determine a good value for k



The Basic Process

1. Decide on your *similarity* metric –and the scaling of your data!
2. Split the labeled set into training and test data
3. Pick an evaluation metric (similar to R^2 and p-values for linear reg)
4. Run with a few different values of k, check against evaluation metric
5. Select k with the best evaluation metric
6. Run on unlabeled data

Distance Metrics

- This varies a lot based on context
  - Numerical values (ie salary, height, age, etc) are "easy" (sort of)
  - What about more abstract attributes
    - Social networks
    - Text based data
    - Movie genre

Numerical Distance and Scale

- If our data is numerical in nature, there are a number of known ways to define "distance" between two things
  - Euclidian, Cosine, Manhattan, Mahalanobis, etc
- What about scale?
  - Consider clustering people based on salary and SAT scores:
    - The distance between ($30,000, 1400) and ($100,000, 1450) is dominated by the salary difference
    - Rescaling data, ie (30, 1400) and (100, 1450) balances the effect of each parameter…but is that necessarily the goal?

Numerical Distance and Scale

- If our data is numerical in nature, there are a number of known ways to define "distance" between two things
  - Euclidian, Cosine, Manhattan, Mahalanobis, etc
- What about scale?
  - Consider clustering people based on salary and SAT scores:
    - The distance between ($30,000, 1400) and ($100,000, 1450) is dominated by the salary difference
    - Rescaling data, ie (30, 1400) and (100, 1450) balances the effect of each parameter…but is that necessarily the goal?

How you scale your data can have a significant impact on outcome, and therefore is also part of your model!

Non-Numerical Data

- Certain distance metrics can deal with non-numerical data
  - ie Jaccard Distance, Hamming Distance
- Many times, however, you will have to define your own
  - Consider movie genre, how "far" apart are two genres?
    - Could define the same genre as 0 apart and different genres as x apart. x can be chosen based on the scale of other numeric attributes
      - This choice is now also a parameter to your model

Evaluation Metrics

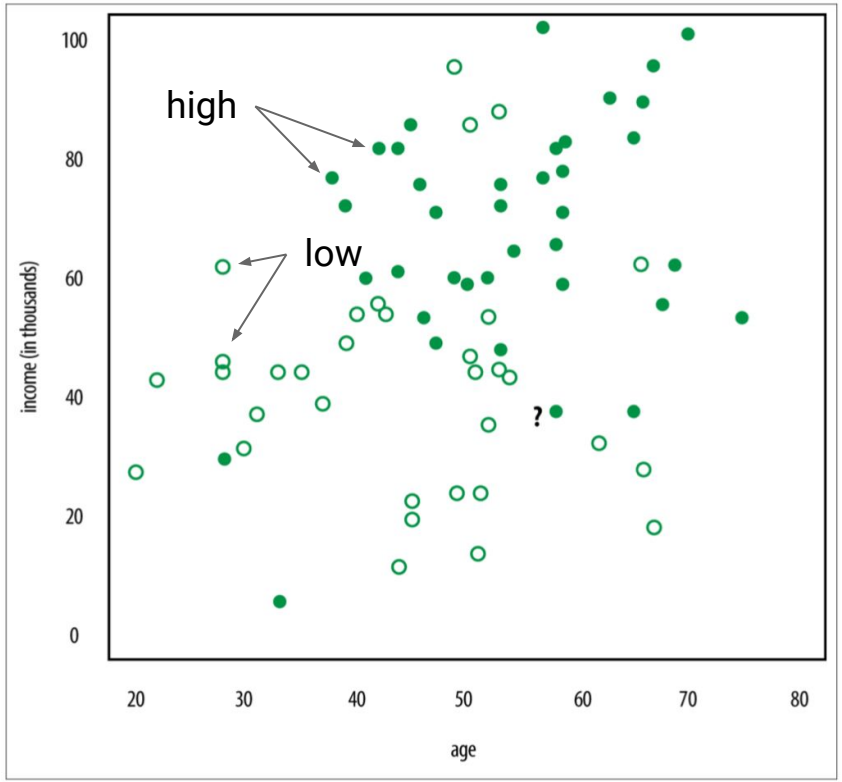
- How do you measure the effectiveness of your model?
  - Accuracy? (number of items correctly categorized)
- Accuracy seems like an obvious choice, but that's not always true…
  - See DSfS Ch 11
- Are all misclassifications created equal? Does a false-positive carry more weight than a false-negative?
  - Precision: how accurate our positive predictions are
  - Recall: what fraction of positive results did our model identify

Finding k

- Now that you have your model setup and know how you will evaluate, you can run the algorithm for different values of k
  - For each item in your *test* set, assume you don't know its label
  - Find its k-nearest neighbors in the training set to determine its label by majority vote
  - After labeling everything in the test set, evaluate effectiveness with your chosen evaluation metric
- Select k which yielded the best results based on your chosen metric

Doing Data Science Ch. 3 Example

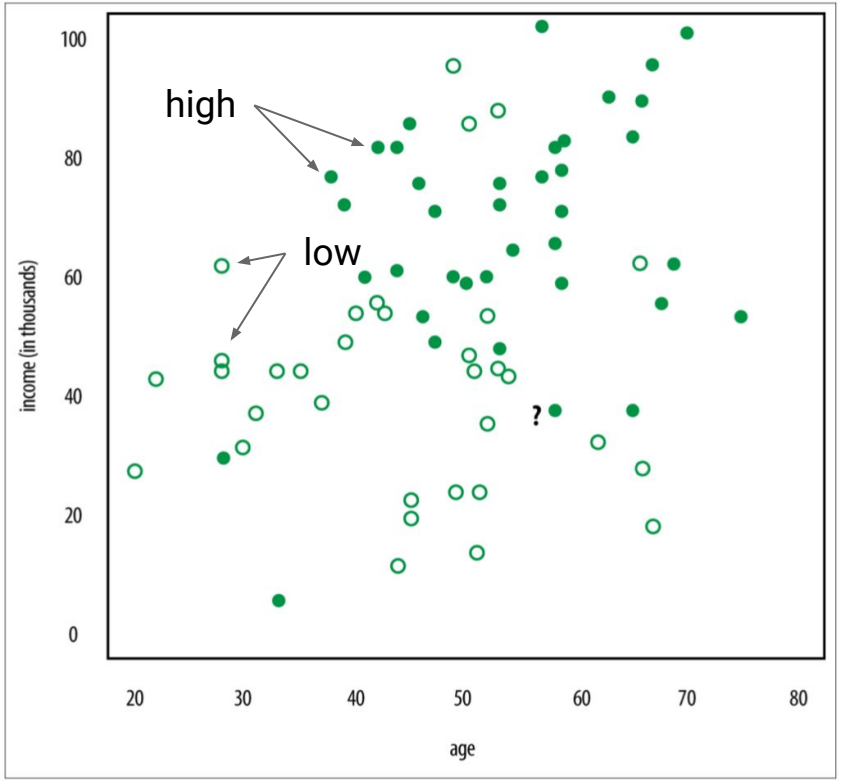
Age	Income	credit
69	3	Low
66	57	Low
49	79	Low
49	17	Low
58	26	high
44	71	high



Doing Data Science Ch. 3 Example

- Dataset tracking age, income (in thousands), and "high" or "low" credit
- What if a new guy comes in who is a 57 year old with $37k income?
- k=5 had the lowest misclassification rate

- Model predicts "low" credit



Applications of k-NN

- **Data preprocessing**: Datasets frequently have missing values, but the KNN algorithm can estimate for those values in a process known as missing data imputation.
- **Recommendation Engines**: Using clickstream data from websites, the KNN algorithm has been used to provide automatic recommendations to users on additional content.
- **Healthcare**: KNN has also had application within the healthcare industry, making predictions on the risk of heart attacks and prostate cancer.

Advantages and disadvantages of the KNN algorithm

- Advantages
- **- Easy to implement**: Given the algorithm's simplicity and accuracy, it is one of the first classifiers that a new data scientist will learn.

- **- Adapts easily**: As new training samples are added, the algorithm adjusts to account for any new data since all training data is stored into memory.
- **- Few hyperparameters**: KNN only requires a k value and a distance metric, which is low when compared to other machine learning algorithms.

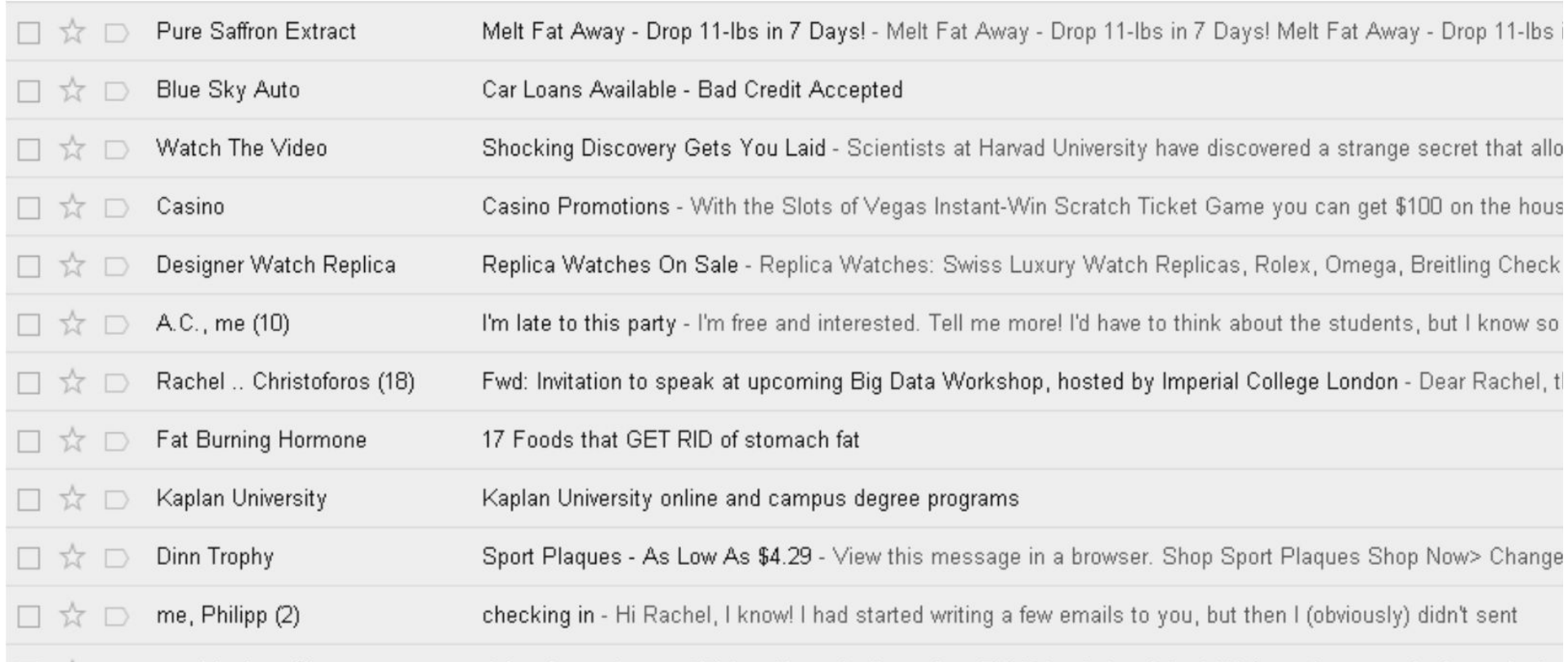
Advantages and disadvantages of the KNN algorithm

- **Does not scale well**: Since KNN is a lazy algorithm, it takes up more memory and data storage compared to other classifiers. This can be costly from both a time and money perspective.
- **Curse of dimensionality**: which means that it doesn’t perform well with high-dimensional data inputs.

Some Notes on Structural Classifiers

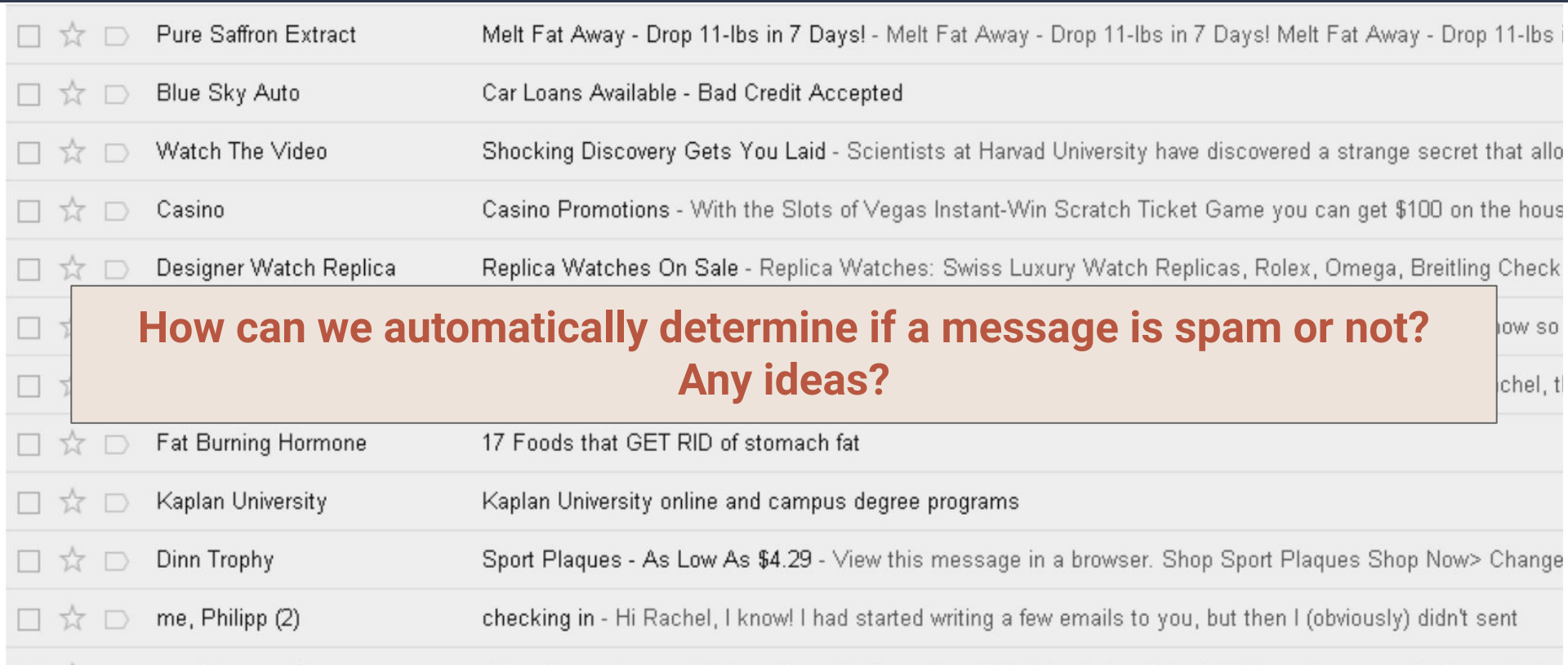
- **Decision trees:** simple and powerful; work well for discrete (0,1/yes,no) rules
- **Neural nets:** a black box approach; can be hard to interpret results
- **Distance-based (ie k-NN):** work well for low-dimensionality spaces)

Motivating Example: Spam Classification



Pure Saffron Extract	Melt Fat Away - Drop 11-lbs in 7 Days! - Melt Fat Away - Drop 11-lbs in 7 Days! Melt Fat Away - Drop 11-lbs
Blue Sky Auto	Car Loans Available - Bad Credit Accepted
Watch The Video	Shocking Discovery Gets You Laid - Scientists at Harvad University have discovered a strange secret that allo
Casino	Casino Promotions - With the Slots of Vegas Instant-Win Scratch Ticket Game you can get $100 on the hous
Designer Watch Replica	Replica Watches On Sale - Replica Watches: Swiss Luxury Watch Replicas, Rolex, Omega, Breitling Check
A.C., me (10)	I'm late to this party - I'm free and interested. Tell me more! I'd have to think about the students, but I know so
Rachel .. Christoforos (18)	Fwd: Invitation to speak at upcoming Big Data Workshop, hosted by Imperial College London - Dear Rachel, t
Fat Burning Hormone	17 Foods that GET RID of stomach fat
Kaplan University	Kaplan University online and campus degree programs
Dinn Trophy	Sport Plaques - As Low As $4.29 - View this message in a browser. Shop Sport Plaques Shop Now> Change
me, Philipp (2)	checking in - Hi Rachel, I know! I had started writing a few emails to you, but then I (obviously) didn't sent

Motivating Example: Spam Classification



Motivating Example: Spam Classification

Goal: Classify email into spam and not spam (binary classification)

Motivating Example: Spam Classification

Goal: Classify email into spam and not spam (binary classification)

Let's say you get an email saying "You've won the lottery!"

How do we know right away that this email is spam?

Motivating Example: Spam Classification

Goal: Classify email into spam and not spam (binary classification)

Let's say you get an email saying "You've won the lottery!"

How do we know right away that this email is spam?

Idea: The use of certain words, ie lottery, can indicate an email is spam.

What about previous techniques?

So, our features in this problem are individual words…

Can we use linear regression or k-NN to detect spam?

What about previous techniques?

So, our features in this problem are individual words…

Can we use linear regression or k-NN to detect spam?

- Linear regression deals with continuous variables
  - We could use a heuristic to convert a continuous range into a binary range…but we are dealing with a huge number of features

What about previous techniques?

So, our features in this problem are individual words…

Can we use linear regression or k-NN to detect spam?

- Linear regression deals with continuous variables
  - We could use a heuristic to convert a continuous range into a binary range…but we are dealing with a huge number of features
- k-NN works well for low dimensionality…but again, we have a huge number of features (potentially thousands of words).
  - Curse of Dimensionality…