CSE 4/587

Data Intensive Computing

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

Day 16
Naive Bayes

Announcements and Feedback

- Read Doing Data Science Chapter 4

Classification

- Classification involves taking a set of unlabeled data points and labeling them in some fashion
  - k-NN was one way to use a model to automatically classify a set of points
- Why?
  - To learn from the classification/data
  - To discover patterns
  - Automate some process, ie handwriting recognition

Classification

Classification relies on ***apriori reference structures*** that divide the space of all possible data points into a set of classes that are not overlapping.

- 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

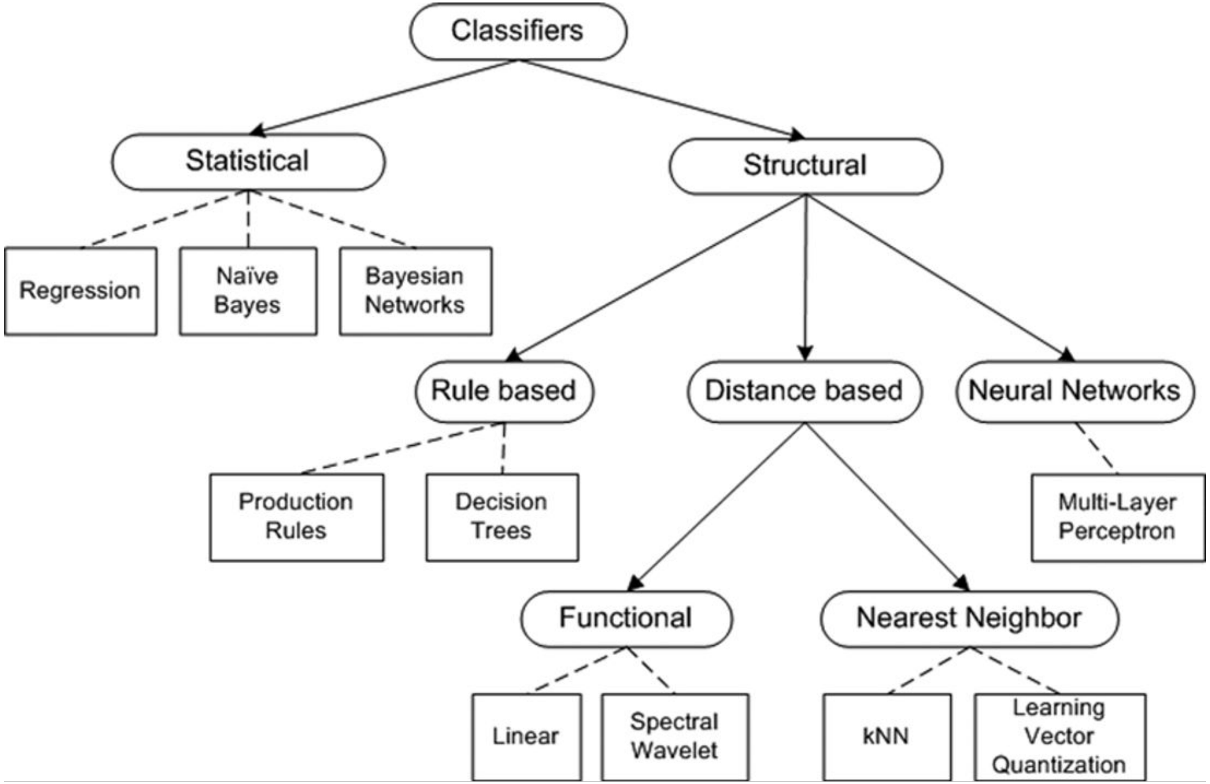
- Restaurant menu: appetizers, salads, soups, entrée, dessert, drinks, …
- Library of congress (LIC) system classifies books according to a standard scheme
- Injury and disease classification in healthcare
- Classification of all living things: eg., Home Sapiens (genus, species)
- Classification across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

Classification of Classification Algorithms

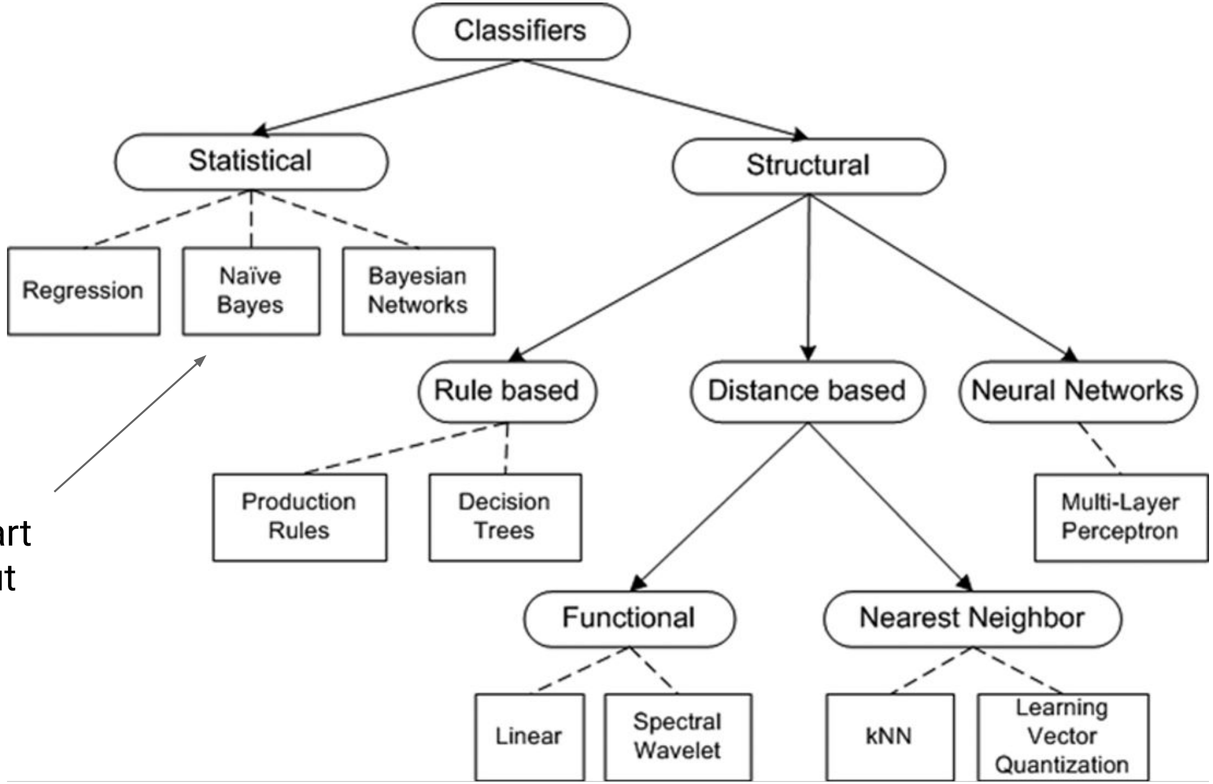
Classification algorithms can be divided into two broad categories:

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

Classification of Classification Algorithms



Classification of Classification Algorithms

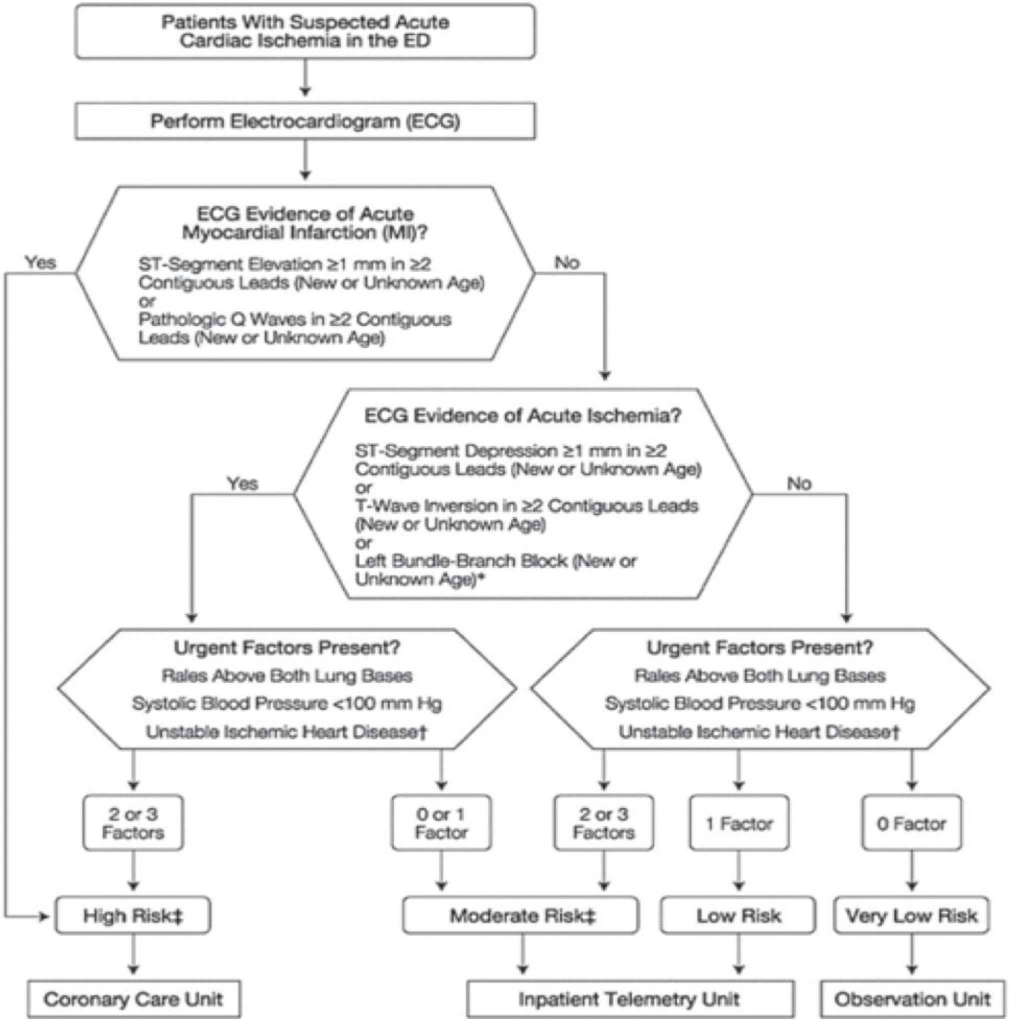


Today we'll start learning about Naive Bayes

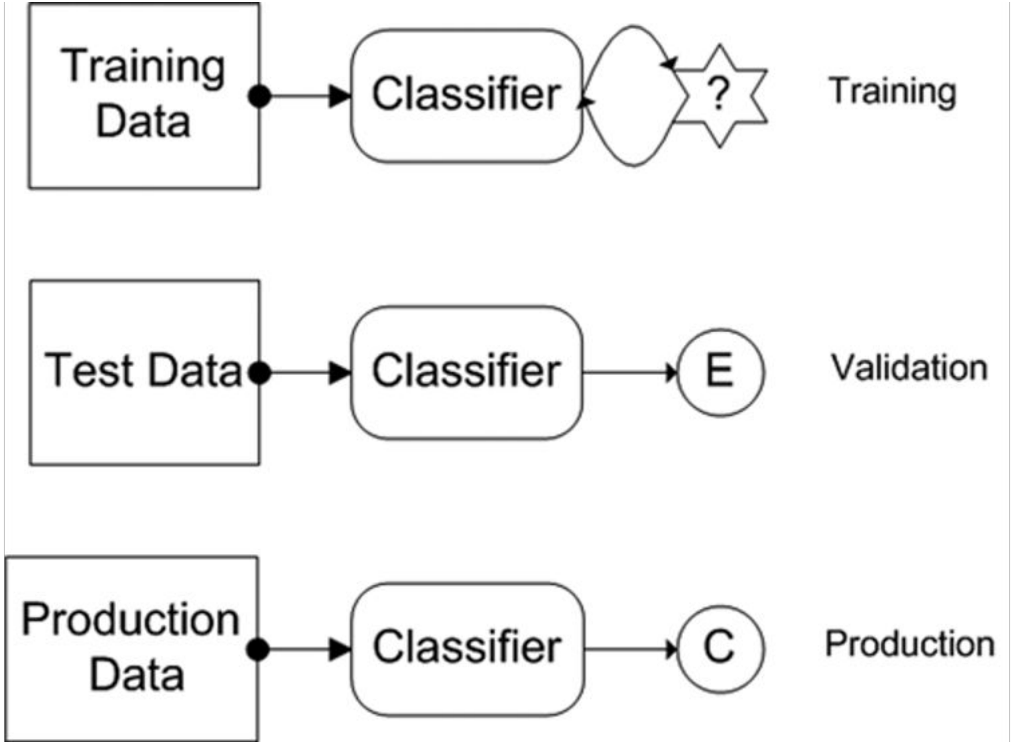
Some Notes on Structural Classifiers

- **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)

Decision tree in the ER of Cooke County hospital, Chicago, IL



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

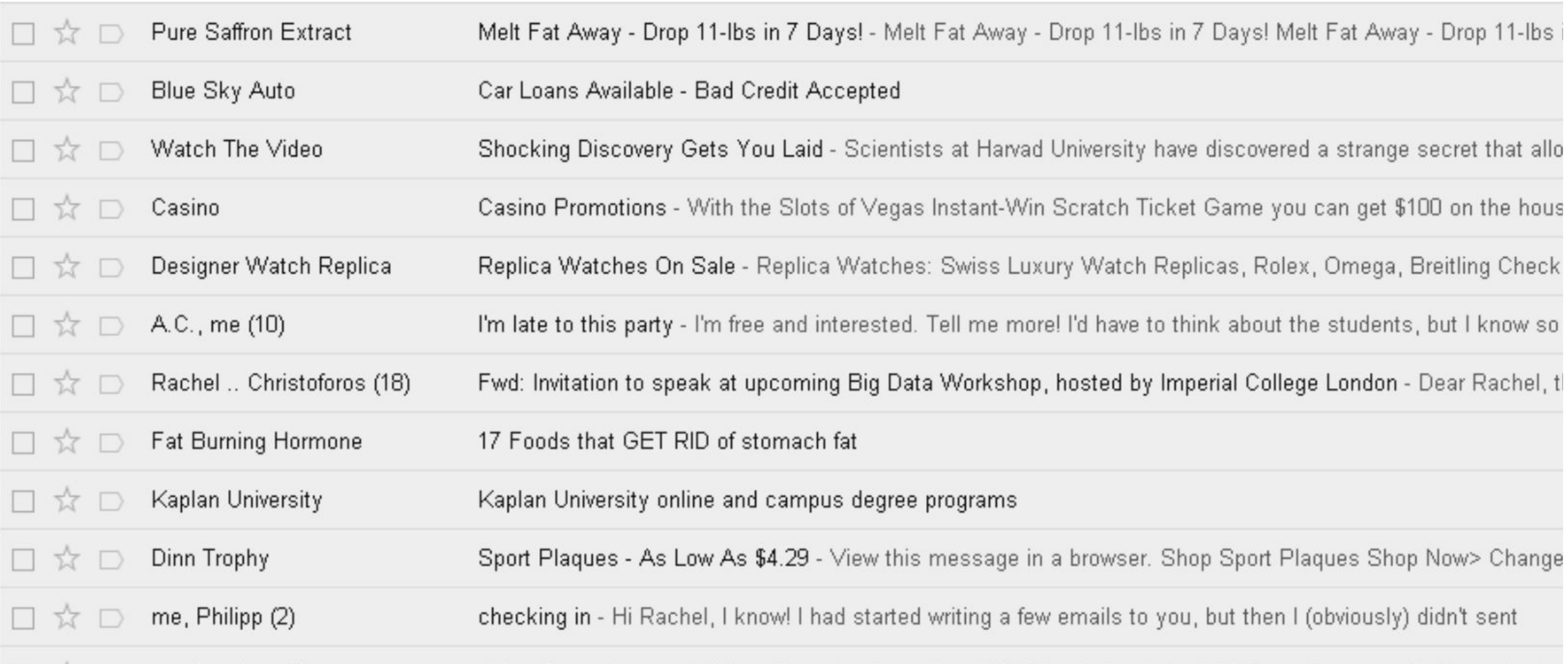
- In the validation stage we validate the classifier to ensure credibility
- Primary goal of this stage is to determine the classification errors
- Quality of the results should be evaluated using various metrics
- Training and testing stages ***may be repeated several times*** before a classifier transitions to the production stage
  - We could evaluate several types of classifiers and pick one or combine all classifiers into a meta-classifier scheme

Production Stage

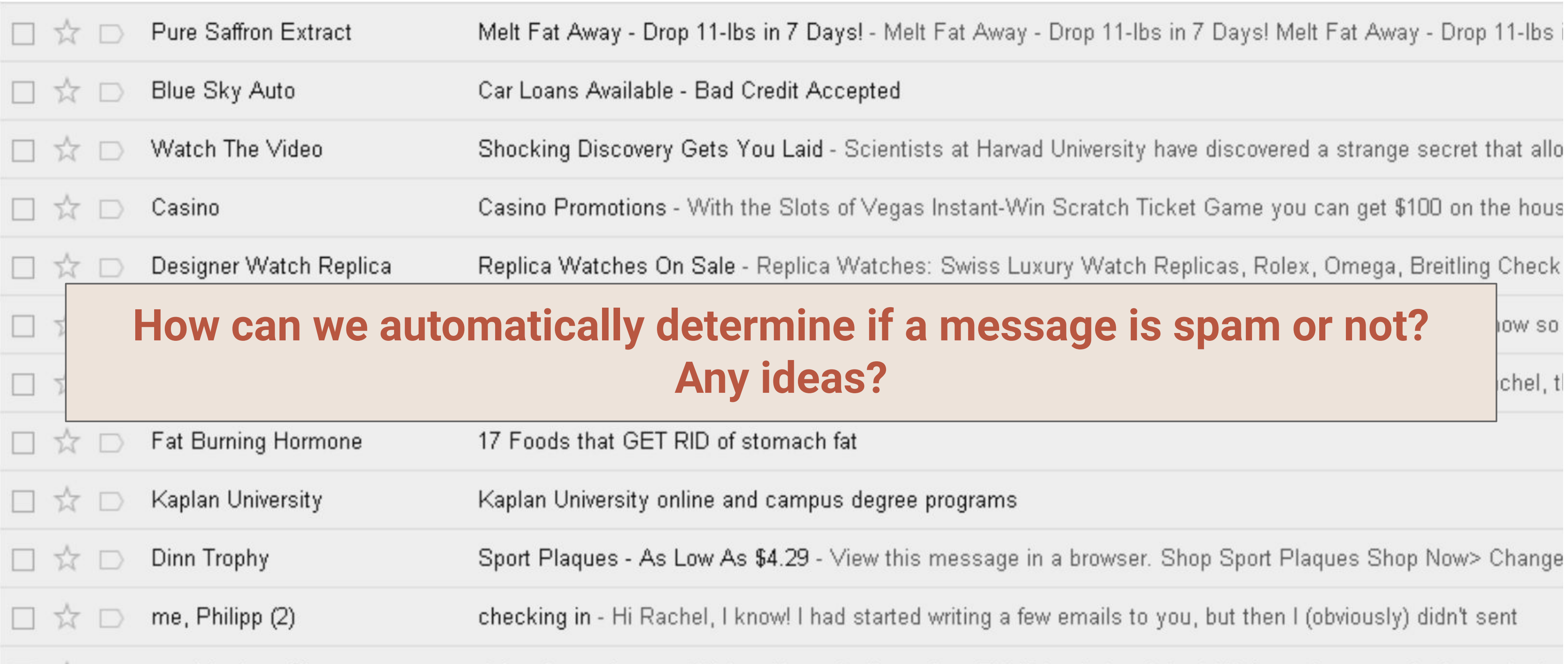
- Now our classifier(s) are ready for use in a live production system
- We can enhance the results by allowing human-in-the-loop feedback

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

Motivating Example: Spam Classification



Motivating Example: Spam Classification



How can we automatically determine if a message is spam or not? Any ideas?

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…

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…

So what do we do?

Naive Bayes

Basic Idea: Make a probabilistic model – have many ***simple rules***, and aggregate those rules together to provide a probability.

Bayes Law and Probability Theory

Basic principle: P(*H* | *E*) = P(*E* | *H*) * P(*H*) / P(*E*)

Posterior probability is proportional to likelihood times prior

- *H* – hypothesis   *E* – evidence
- **Prior** = probability of the *E* given *H*; P(*E* | *H*)
- **Likelihood** = P(*H*) / P(*E*)
- **Posterior** = Probability of *H* given *E*; P(*H* | *E*)

Probability Theory Refresher

Here is the derivation from first principles of probabilities:

Probability Theory Refresher

Here is the derivation from first principles of probabilities:

$$P(A \mid B) = P(A \,\&\, B) \,/\, P(B)$$

Probability Theory Refresher

Here is the derivation from first principles of probabilities:

$$P(A \mid B) = P(A \,\&\, B) \,/\, P(B)$$

$$P(B \mid A) = P(A \,\&\, B) \,/\, P(A)$$

Probability Theory Refresher

Here is the derivation from first principles of probabilities:

$$P(A \mid B) = P(A \,\&\, B) \,/\, P(B)$$

$$P(B \mid A) = P(A \,\&\, B) \,/\, P(A)$$

Multiply both sides by $P(A)$

$$P(B \mid A)\, P(A) = P(A \,\&\, B)$$

Probability Theory Refresher

Here is the derivation from first principles of probabilities:

$$P(A \mid B) = P(A \,\&\, B) \,/\, P(B)$$

$$P(B \mid A) = P(A \,\&\, B) \,/\, P(A)$$

$$P(B \mid A)\, P(A) = P(A \,\&\, B)$$

Sub P(A & B) into first eq

$$P(A|B) = (P(B|A)\, P(A)) \,/\, P(B)$$

Bayes Law - Example

Suppose you know that I work 5 days out of the week.

Also suppose you know that on work days, I never wear flip flops, and on non-work days I wear flip flops 70% of the time.

Given this information, if you see me on a random day of the week wearing shoes, what is the probability that I had work that day?

Bayes Law - Example

Basic principle: $P(H \mid E) = P(E \mid H) * P(H) / P(E)$

Bayes Law - Example

Basic principle: $P(H \mid E) = P(E \mid H) * P(H) / P(E)$

- What is our hypothesis, H?

Bayes Law - Example

Basic principle: $P(H \mid E) = P(E \mid H) * P(H) / P(E)$

- What is our hypothesis, H? **I went to work today**

Bayes Law - Example

Basic principle: P(H | E) = P(E | H) * P(H) / P(E)

- What is our hypothesis, H? **I went to work today**
- What is the evidence, E, that we observed?

Bayes Law - Example

Basic principle: $P(H \mid E) = P(E \mid H) * P(H) / P(E)$

- What is our hypothesis, H? **I went to work today**
- What is the evidence, E, that we observed? **I'm wearing shoes**

Bayes Law - Example

Basic principle: $P(H \mid E) = P(E \mid H) * P(H) / P(E)$

- What is our hypothesis, H? **I went to work today**
- What is the evidence, E, that we observed? **I'm wearing shoes**
- What is $P(H)$?

Bayes Law - Example

Basic principle: P(*H* | *E*) = P(*E* | *H*) * P(*H*) / P(*E*)

- What is our hypothesis, H? **I went to work today**
- What is the evidence, E, that we observed? **I'm wearing shoes**
- What is P(*H*)? **5/7 = 0.71**

Bayes Law - Example

Basic principle: P(*H* | *E*) = P(*E* | *H*) * P(*H*) / P(*E*)

- What is our hypothesis, H? **I went to work today**
- What is the evidence, E, that we observed? **I'm wearing shoes**
- What is P(*H*)? **5/7 = 0.71**
- What is P(*E*)?

Bayes Law - Example

Basic principle: $P(H \mid E) = P(E \mid H) * P(H) / P(E)$

- What is our hypothesis, H? **I went to work today**
- What is the evidence, E, that we observed? **I'm wearing shoes**
- What is $P(H)$? **5/7 = 0.71**
- What is $P(E)$? **5/7 * 1.0 + 2/7 * 0.3 = 0.8**

Bayes Law - Example

Basic principle: $P(H \mid E) = P(E \mid H) * P(H) / P(E)$

- What is our hypothesis, H? **I went to work today**
- What is the evidence, E, that we observed? **I'm wearing shoes**
- What is $P(H)$? **5/7 = 0.71**
- What is $P(E)$? **5/7 * 1.0 + 2/7 * 0.3 = 0.8**
- What is $P(E \mid H)$?

Bayes Law - Example

Basic principle: P(H | E) = P(E | H) * P(H) / P(E)

- What is our hypothesis, H? **I went to work today**
- What is the evidence, E, that we observed? **I'm wearing shoes**
- What is P(H)? **5/7 = 0.71**
- What is P(E)? **5/7 * 1.0 + 2/7 * 0.3 = 0.8**
- What is P(E | H)? **1.0**

Bayes Law - Example

Basic principle: P(*H* | *E*) = P(*E* | *H*) * P(*H*) / P(*E*)

- What is our hypothesis, H? **I went to work today**
- What is the evidence, E, that we observed? **I'm wearing shoes**
- What is P(*H*)? **5/7 = 0.71**
- What is P(*E*)? **5/7 * 1.0 + 2/7 * 0.3 = 0.8**
- What is P(*E* | *H*)? **1.0**

Therefore, if you see me in shoes, there is an 88% I went to work today

Bayes Law - Spam Classification

Given Bayes Law, how can we start classifying emails as spam?

Bayes Law - Spam Classification

Given Bayes Law, how can we start classifying emails as spam?

Let's start one word at a time:

P(*spam*|*word*) = P(*word*|*spam*) * P(*spam*) / P(*word*)

Bayes Law - Spam Classification

Given Bayes Law, how can we start classifying emails as spam?

Let's start one word at a time:

Probability that the given word appears in an email

$$P(spam|word) = P(word|spam) * P(spam) / P(word)$$

Probability that an email is spam if it contains a given word

Probability that the given word appears in an email known to be spam

Probability that an email is spam

Bayes Law - Spam Classification

We've now boiled our classification problem down to a counting problem:

Given a set of emails that have been classified as spam or not spam (ham):

1. Count number of spam vs ham emails to compute **P(*spam*)**
2. Count number of times the given word, ie lottery, appears in emails to compute **P(*word*)**
3. Count number of times the given word appears in spam emails to compute **P(*word|spam*)**

Bayes Law - Spam Classification

We've now boiled our classification problem down to a counting problem:

Given a set of emails that have been classified as spam or not spam (ham):

1. Count number of spam vs ham emails to compute **P(*spam*)**
2. Count number of times the given word, ie lottery, appears in emails to compute **P(*word*)**
3. Count number of times the given word appears in spam emails to compute **P(*word*|*spam*)**

Enron Email Example - DDS Chapter 4

- **Input:** Enron data set containing employee emails
- A small subset chosen for EDA
- 1500 spam, 3672 ham
- Test word is “meeting”
- Running a simple shell script reveals that there are 16 spam emails containing “meeting” and 153 ham emails containing "meeting"
- **Output:** What is the probability that an email containing "meeting" is spam? What is your intuition? Now prove it using Bayes Law…

Enron Email Example - DDS Chapter 4

Enron Email Example - DDS Chapter 4

P(*spam*) = 1500 / (1500+3672) = 0.29

Enron Email Example - DDS Chapter 4

P(*spam*) = 1500 / (1500+3672) = 0.29

P(*ham*) = 1 - **P(*spam*)** = 0.71

Enron Email Example - DDS Chapter 4

P(*spam*) = 1500 / (1500+3672) = 0.29

P(*ham*) = 1 - **P(*spam*)** = 0.71

P(*meeting|spam*) = 16/1500 = 0.0106

Enron Email Example - DDS Chapter 4

P(*spam*) = 1500 / (1500+3672) = 0.29

P(*ham*) = 1 - **P(*spam*)** = 0.71

P(*meeting|spam*) = 16/1500 = 0.0106

P(*meeting|ham*) = 153/3672 = 0.0416

Enron Email Example - DDS Chapter 4

$P(\textit{spam}) = 1500 / (1500+3672) = 0.29$

$P(\textit{ham}) = 1 - P(\textit{spam}) = 0.71$

$P(\textit{meeting}|\textit{spam}) = 16/1500 = 0.0106$

$P(\textit{meeting}|\textit{ham}) = 153/3672 = 0.0416$

$P(\textit{meeting}) = (16+153) / (1500+3672) = 0.0326$

Enron Email Example - DDS Chapter 4

P(*spam*) = 1500 / (1500+3672) = 0.29

P(*ham*) = 1 - **P(*spam*)** = 0.71

P(*meeting|spam*) = 16/1500 = 0.0106

P(*meeting|ham*) = 153/3672 = 0.0416

P(*meeting*) = (16+153) / (1500+3672) = 0.0326

P(*spam|meeting*) = **P(*meeting|spam*)*P(*spam*)/P(*meeting*)** = 0.094 (9.4%)

Naive Bayes

Basic Idea: Make a probabilistic model – have many ***simple rules***, and aggregate those rules together to provide a probability.

Naive Bayes

Bayes law for each word

Basic Idea: Make a probabilistic model – have many ***simple rules***, and aggregate those rules together to provide a probability.

Naive Bayes

Bayes law for each word

Basic Idea: Make a probabilistic model – have many ***simple rules***, and <u>aggregate those rules together to provide a probability</u>.

Putting It All Together - Naive Bayes

So we've counted and computed probabilities for all words in our input

Putting It All Together - Naive Bayes

So we've counted and computed probabilities for all words in our input

Let's say we have $\boldsymbol{i}$ words. Let $\boldsymbol{x}$ be a vector of size $\boldsymbol{i}$,

where $\boldsymbol{x_j}$ **= 1** if the $\boldsymbol{j^{th}}$ word is present in an email, **0** otherwise.

Putting It All Together - Naive Bayes

So we've counted and computed probabilities for all words in our input

Let's say we have $\boldsymbol{i}$ words. Let $\boldsymbol{x}$ be a vector of size $\boldsymbol{i}$,

where $\boldsymbol{x_j} = \mathbf{1}$ if the $\boldsymbol{j^{th}}$ word is present in an email, **0** otherwise.

Now how do we compute P(*x*|*spam*)?

Once we do this, we can apply Bayes Law to find P(*spam*|*x*)

Putting It All Together - Naive Bayes

So we've counted and computed probabilities for all words in our input

Let's say we have $\boldsymbol{i}$ words. Let $\boldsymbol{x}$ be a vector of size $\boldsymbol{i}$,

where $\boldsymbol{x_j}$ **= 1** if the j^{th} word is present in an email, **0** otherwise.

Now how do we compute P(*x*|*spam*)?

Once we do this, we can apply Bayes Law to find P(*spam*|*x*)

This is where we will begin next lecture...