


Machine Learning

ML


Artificial Intelligence → ML


### Machine Learning


DEER




HORSE









DEERI




How does a human learn?


### What is Machine Learning?

- A system that can improve on task  $T$ , with respect to performance measure  $P$ , after observing experience  $E$ .
- Task: Distinguish
 



- Experience: Labeled instances of Deer & Horses
- Performance: Accuracy

### Machine learning applications









Task? Performance? Experience?

### Central challenge in machine learning

How can we build computer systems that automatically improve with experience, and what laws govern learning in general?

- Statistics
  - What can be inferred from a set of data, with what reliability?
- Computer science
  - How can we build computers to solve problems, and which problems are tractable/intractable?
- Human learning
  - What mechanisms explain learning in humans, and what teaching strategies are most effective?

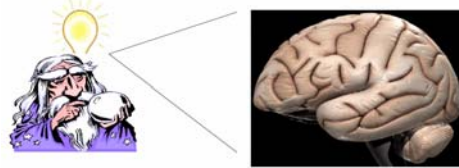
Eric Xing 2006

### Why use machine learning?

- Data comes in too fast for humans to process
  - Every credit card transaction
  - Every e-mail message
- Data set is just too large for humans to process
  - Protein folding
  - Sloan Digital Sky Survey III
- Machines can make decisions faster
  - Once trained, many models predict almost instantly
- Personalization / adaptation
  - Speech recognition
  - ...



### What's going on in the wizard's head?



- What are the **concepts** or **models** being learned?
- We'll talk about three kinds
  - Rules
  - Linear models
  - Memory-based

### Rule-based models

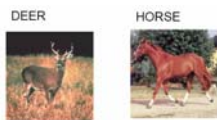
- A sequence of if-then
  - Just like Matlab's **if-else!**
- Need to express rules explicitly
- Example: grammar/spelling checker

### Linear models

- Represent a problem as a set of **features**; each feature gets a number of points
- Example: is this document about soccer?
  - Contains "soccer": 50 points
  - Contains "basketball": -50 points
  - Contains "Beckham": 100 points
  - Contains "Posh Spice": -100 points
  - ...
  - If total number of points > 0, say "yes"

### Memory-based models

- The model is the training data!



Which is closer?

Very, very hard to write an algorithm but data is easy to collect

### How do we get these models?

- Labeled training data
  - Humans have to
    - Collect data (emails, pictures of animals, ...)
    - Label the data (spam vs. not-spam; deer vs. Horse, ...)



- There are many algorithms
  - We'll discuss one: Naïve Bayes

### Experimentation

- Need to train and evaluate
- Split data into a **training** set and a **test** set
  - Train the wizard on the training data
  - Evaluate the wizard on the test data
- Lots of data is needed to get a wise wizard, so why not use the whole data set for training?
  - If you evaluate the wizard on (any part of the) training data, it's like letting the wizard "cheat" on a test

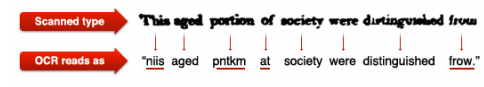
### Measuring the wizard's skill

- Simplest measure is **accuracy** – on what fraction of the test cases does the wizard predict correctly?
- Accuracy is not a good measure in some cases
  - E.g., credit card fraud
  - Very rare event → always negative = 99.9% accurate
  - Better measure: **false positive rate**, **false negative rate**
- **Precision** and **recall** (remember them?)

### Other machine learning topics

- Active learning
  - Labels are expensive!
  - Remember CAPTCHA?
  - Image labeling game

### reCAPTCHA



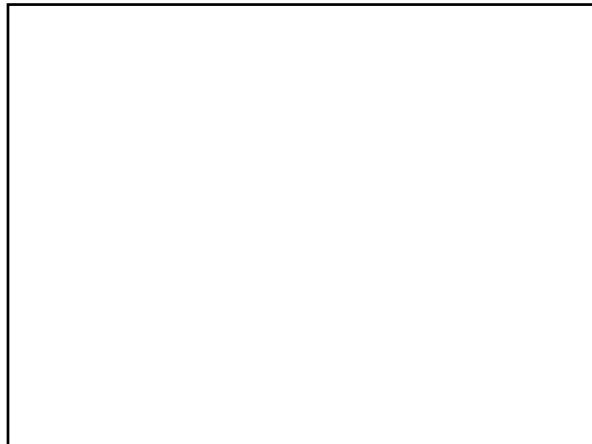
Is OCR "learning" when you "teach" the system how to read/recognize the characters in the word?

### Image labeling game



### Other machine learning topics

- Active learning
  - Labels are expensive!
  - Remember CAPTCHA?
  - Image labeling game
- Unsupervised learning
  - Learning without the correct answers
- Theory
  - How much data do you need to learn something?
  - What kinds of concepts can you learn?



- Also known as “junk mail” or “unsolicited bulk” e-mail”
  - Unsolicited – you didn't ask for it
  - Bulk – sent to lots and lots of people, not just you
- Typical legal definition: **unsolicited commercial** email from someone without a **pre-existing** business relationship
- Huge problem
  - 50% of all e-mail sent is spam

Much of this material is from Joshua Goodman's SPAM tutorial

What are the recent spam topics

Why is there spam?

- **Money!**
- Almost free advertisement
- Cost of sending spam ~0.01 cent per message.
- If 1 in 100,000 people buy, and I earn \$11 for that purchase, then I make a profit!

Spamming techniques to defeat filters

- Content in image
- Chaff
  - Text chaff
  - Content chaff
- Obscuring words



Weather Report Guy

- Content in Image
- Good Word Chaff

Weather, Sunny, High 82, Low 81, Favorite...

100's of Lenders Compete for your Loan to get you the Lowest Rate!

- Refinancing
- New Home Loans
- Debt Consolidation
- Debt Consultation
- Auto Loans
- Credit Cards
- Student Loans
- Second Mortgage
- Home Equity

Interest Rates are at their lowest point in 40 years! We help you find the best rate for your situation by matching your needs with hundreds of lenders!

100% Free Service!

[Click Here To Begin](#)

Good Credit - Bad Credit  
Bankruptcy - Foreclosure

Weather  
MA, VA - Sunny  
High: 82, Low: 81 degrees  
Favorites

### The Hitchhiker Chaffer

- Content Chaff
  - Random passages from the *Hitchhiker's Guide to the Galaxy*
  - Footers from valid mail

"This must be Thursday," said Arthur to himself, sinking low over his beer, "I never could get the hang of Thursdays."

Express yourself with MSN Messenger 6.0...

We are offering a 7 day / 6 night vacation at a huge discount and we will also add a 7 night vacation to any of our 17 destinations.

Many hotels pay for most of your accommodations in exchange for you to view their resorts, hoping you will spend money on their services.

Click below to claim your trip:  
<http://www.acktoype.com/8/>

To update your list preference: [acktoype.com/8/](http://www.acktoype.com/8/)

"Does your health insurance cover pets?"; "This must be Thursday," said Arthur to himself, sinking low over his beer, "I never could get the hang of Thursdays."

Express yourself with MSN Messenger 6.0 -- download now!  
[http://www.msnmessenger-download.com/tracking/each\\_general](http://www.msnmessenger-download.com/tracking/each_general)

### Diploma Guy

- Word Obscuring
  - Diplmoia Pragorm
  - Caerte a mroe prosoepeprus

**Diplmoia Pragorm**

Caerte a mroe prosoepeprus fruse for yormself

Receive a fdl diploma from non accredited universities based upon your real life emmeritpece

You will not be tested, or interarround  
Receve a Master's, Bachelor's or Doctorooe

Clid 24 hours a day 7 days a week

1 - 270 - 817 - 8247


### Besides email, where else do you get spam?

- "old media"
  - Physical junk mail
  - Phone calls
- Instant messenger
- Chat rooms
- Popups
- Link spam

### Naïve Bayes spam filtering


- Most common kind of spam filter
  - Although many filters include rules and other features
- Who is Bayes?
  - 18<sup>th</sup> century mathematician
  - Let's learn some probability

### Probability



- If I pick one sock out of these 8, what are the chances that it is/has red?
- $P(\text{sock is red}) = 5/8$

### Conditional Probability



- $P(\text{sock is red}) = 5/8$
- What if I decide I'm going to pick a solid sock?
- $P(\text{sock is red} \mid \text{solid pattern}) = 3/4$

### Bayes' Theorem

- Thomas Bayes



- $P(\text{red} | \text{solid}) = P(\text{solid} | \text{red}) * P(\text{red}) / P(\text{solid})$
- $\frac{3}{4} = \frac{3}{5} * \frac{5}{8} / \frac{4}{8}$



### Naïve Bayes machine learning for spam classification

- Use vector space model of messages
- Decide which is bigger:
  - $P(\text{message is spam} | \text{words in message})$
  - or
  - $P(\text{message is not spam} | \text{words in message})$
- Use Bayes' rule

$$P(\text{spam} | \text{words}) = P(\text{words} | \text{spam}) \cdot \frac{P(\text{spam})}{P(\text{words})}$$

$$P(\text{good} | \text{words}) = P(\text{words} | \text{good}) \cdot \frac{P(\text{good})}{P(\text{words})}$$

If  $P(\text{spam} | \text{words}) > P(\text{good} | \text{words})$   
then predict the message is spam.  
Can simplify the comparison. Check if

$$\underbrace{P(\text{words} | \text{spam}) \cdot \frac{P(\text{spam})}{P(\text{words})}}_{\text{score}(\text{spam})} > \underbrace{P(\text{words} | \text{good}) \cdot \frac{P(\text{good})}{P(\text{words})}}_{\text{score}(\text{good})}$$

$$\text{score}(\text{spam}) = P(\text{words} | \text{spam}) \cdot \underbrace{P(\text{spam})}_{\text{just a value}}$$

$$P(\text{word}_1 | \text{spam}) \cdot P(\text{word}_2 | \text{spam}) \cdot \dots \cdot P(\text{word}_n | \text{spam})$$

A simpler calculation is

$$\log P(\text{word}_1 | \text{spam}) + \log P(\text{word}_2 | \text{spam}) + \dots + \log P(\text{word}_n | \text{spam})$$

So instead of checking whether

$$\text{score}(\text{spam}) > \text{score}(\text{good})$$

we check whether

$$\log \text{score}(\text{spam}) > \log \text{score}(\text{good})$$

### Computation of the "probabilities"

- Simplify the computation by
  - computing a score instead of the probability
  - computing  $\log(P)$  instead of  $P$
- Predict **spam** if
  - score(spam) > score(not-spam)
 Otherwise predict **not-spam**
- Why is it "naïve"?
  - Assume the probability of the words in the message to be independent

### What are some solutions for ending spam?

- Filtering
  - Machine learning
  - Blackhole lists (IP filtering)
  - Whitelisting
- Postage
  - Money
  - Turing tests
  - Other computation