LING 473: Day 12

START THE RECORDING
Clustering
Classifiers
Project 5

Machine learning

- Automatically find patterns in the data
- Requirements
  - training data
  - a model
  - tunable parameters
  - objective function: a repeatable quantitative performance measurement procedure

Ockham's ("Occam's") razor

"It is futile to do with more things
that which can be done with
fewer." – William of Ockham (c. 1495)

Ockham's razor

- Of all hypotheses with "parameters" which explain the observations, we seek the simplest hypothesis.

- What's the simplest way to explain the observations?

- In practice, this is not so easy
  - No feature set may model the data with complete accuracy
  - Simplicity/accuracy trade-off
    - Accuracy and complexity are sometimes in direct relation
    - We are often faced with trading reduced accuracy for increased simplicity.
- Machine learning is a tool for exploring this space

Machine Learning

- Supervised
  - methods that require labeled training data
- Unsupervised
  - methods that are trained on raw (unlabeled) training data
- Semi-supervised
  - bootstrap an unsupervised method with a small amount of labeled data
  - "co-training" between two machine learners

Machine Learning Tasks

- Clustering
  - Unsupervised
  - We don't know what the clusters are/will be
  - We don't know the optimal number of clusters
- Classification
  - Supervised
  - We know the classes we seek

Partitioning a corpus

- Training set
  - initial training
- Development set ("dev-test")
  - used for tuning (like a mini-test set)
- Test set ("held-out")
  - reserved for reporting results.
  - Model and parameters cannot be adjusted after touching this data

Train:
build model

Test:
evaluate unseens

Generative *v.* Discriminative: Two sides of the same coin

Discriminative models model a conditional probability between data and output. "Which label is most likely given this data?"

Generative models model a joint probability between the data and output. "Based on my knowledge of the real world, which label is most likely to generate this data?"

Discriminative Models

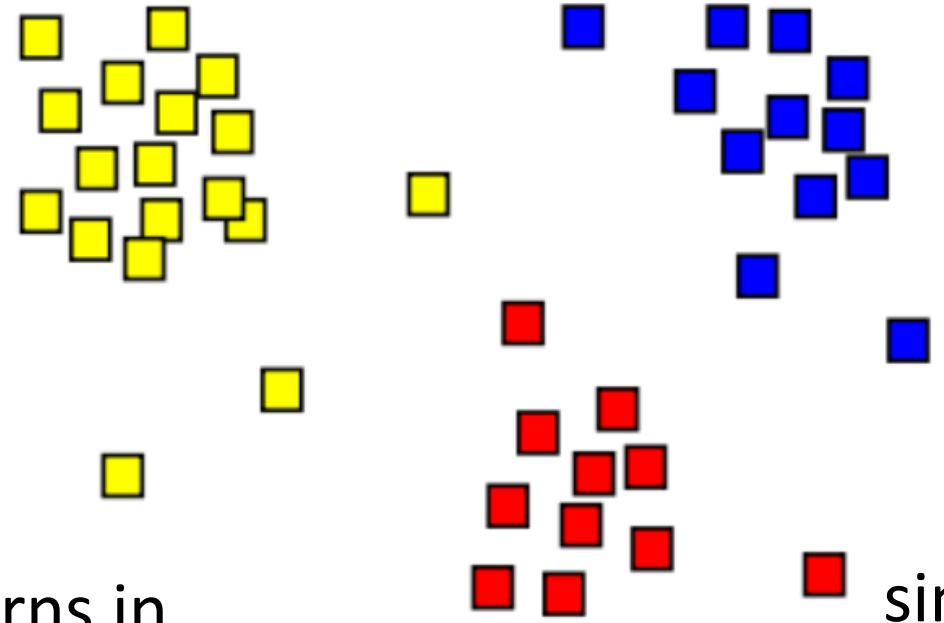
- Linear regression
- Maximum entropy classifiers
- Support Vector Machines (SVM)
- Random Forests
- Multilayer Perceptrons

Generative Models

- Examples:
  - Hidden Markov Model
  - Naïve Bayes
  - Topic Models
- Powerful models - allow for data generation and incorporation of priors
- Generally harder to train than discriminative models, but that is quickly changing.

Clustering

Automated grouping



Learn patterns in unlabeled data

Use a distance metric to find similarity amongst instances

Types of clustering

- Hierarchical vs Non-hierarchical
- Clustering is a form of dimensionality reduction

Uses of clustering

- Exploratory data analysis
  - Find groups in an initial analysis
- Generalization
  - Find equivalence classes in the data

Classification

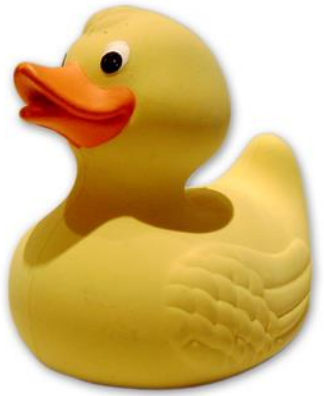
- Classification is automated labeling
- Example: Observe a bird. Its features are:

  walks_like: a duck

  swims_like: a duck

  quacks_like: a duck

  Classification result:



Types of classifiers

- Hidden Markov Model (HMM)
- Decision tree
- Support Vector Machine
- Perceptron
- Naïve Bayes
- k-Nearest Neighbor (kNN)
- Maximum Entropy “MaxEnt”
- Neural networks
- Conditional Random Fields (CRF)
- Many more…

Naïve Bayes Classification

- Intuitive Generative Model
- Simple to implement
- Works well even compared to much more sophisticated techniques
- Does not require large amounts of training data

Why "naïve?"

- It is given this name because of the underlying independence assumption:
  - Features are conditionally independent from one another
- This greatly simplifies calculating the model
  - But may weaken the accuracy

Independence assumption

Your friends have decided to go to a restaurant. It is in Fremont, cheap, an independent store, and open late. What's the likelihood that it is a burger place?

However, do some features influence the others? Does *Fremont* influence *independent store*?

Language Classification

- Given a set of words, which language did it come from?

Naïve Bayes language classifier

- Consider space of n languages **L** = $\{L_1, \ldots, L_n\}$
- Priors: Assume all languages are equally likely, and that probability text is in (some) language is 1.0
- Given data **X =** $\{x_1, \ldots, x_n\}$, we want to model $P(L_i, X)$.
- Using Chain Rule:

$$P(L_i, \mathbf{X}) = P(L_i, x_1, \ldots, x_n)$$
$$= P(x_1 \mid x_2, \ldots, x_n, L_i)\, P(x_2, \ldots, x_n, L_i) \ldots$$
$$= P(x_1 \mid x_2, \ldots, x_n, L_i)\, P(x_2 \mid x_3, \ldots, x_n, L_i)\, P(x_3, \ldots, x_n, L_i)$$
$$= \ldots$$
$$= P(x_1 \mid x_2, \ldots, x_n, L_i)\, P(x_2 \mid x_3, \ldots, x_n, L_i) \ldots P(x_{n-1} \mid x_n, L_i)\, P(x_n \mid L_i)\, P(L\_i)$$

Naïve assumption

All features are independent of all others

For this task, a "feature" is the occurrence of a word

With conditional independence on **X,**

$$P(L_i, X) = P(L_i)\prod_{j=1}^{n} P(x_j|L_i)$$

Last step

To find the **most** probable language:

$$\hat{L} = argmax_i p(L_i) \prod_{j=1}^{n} P(x_j | L_i)$$

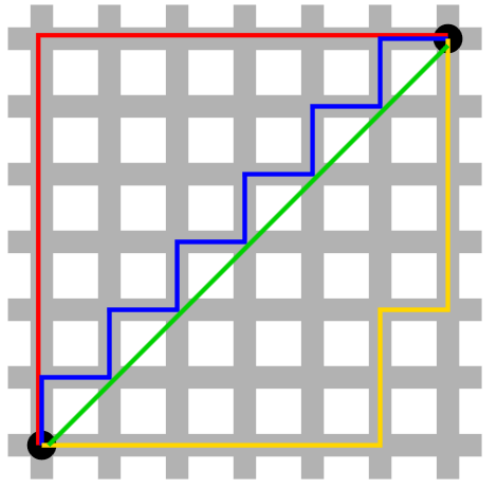
k-Nearest Neighbor Classification

- "Classification by peer pressure"
- Instance-based learning ("lazy learning")
  - No training
- Need a distance metric
- Test instance is given the same label as its closest neighbors
  - Voting schemes resolve conflict
- To test, need to calculate distance to all training instances
  - This can be slow at runtime

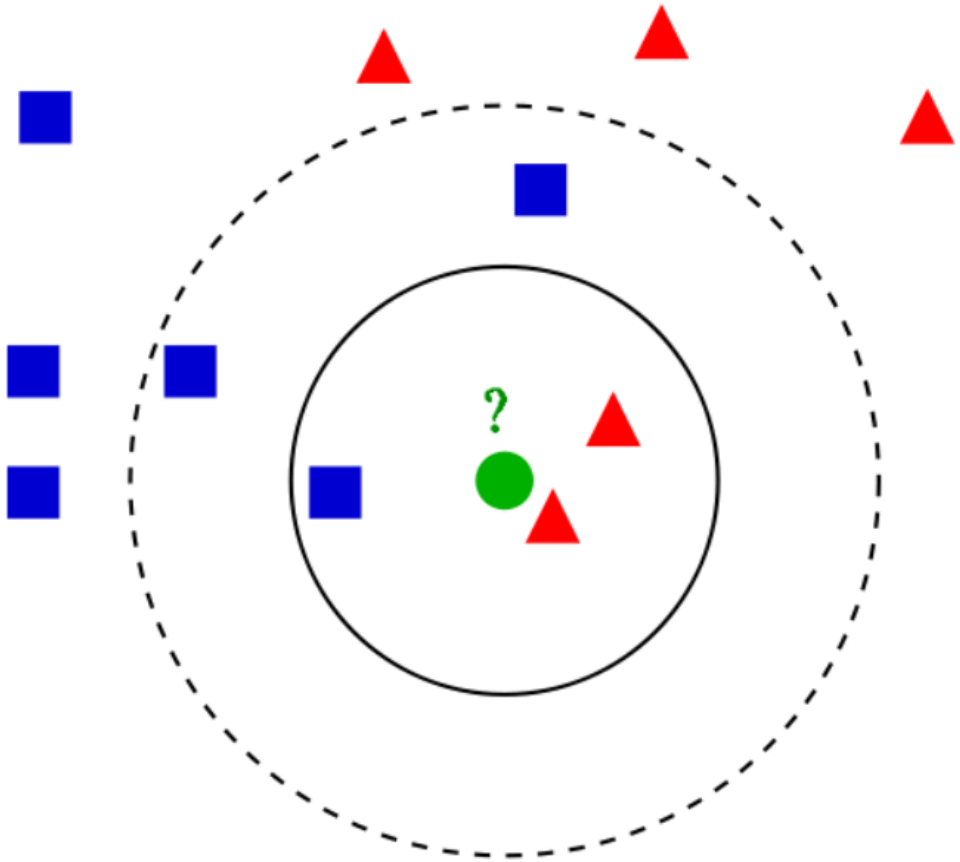
Distance Metrics

- Euclidian
- Cosine
- Hamming distance
- Taxicab distance
- others…

Distance metrics are also used in kNN classification



kNN Classification



http://en.wikipedia.org/wiki/File:KnnClassification.svg

kNN Voting schemes

- Majority voting
  - Choose majority class amongst k closest neighbors
- Weighted voting
  - Weight each of the k neighbors' labels according to the distance to the training instance
  - In principle, this can be applied to an all-neighbors approach

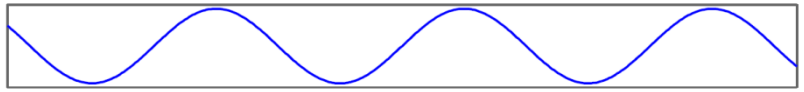
Information Theory

Claude Shannon. 1948. *A mathematical theory of communication*.

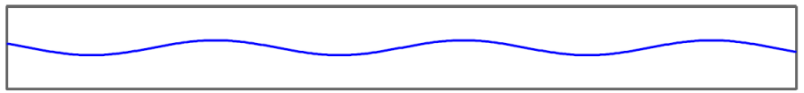
"The fundamental problem of communication is that of **reproducing** at one point… a message **selected** at another point… Semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one **selected from a set** of possible messages."

Information theory

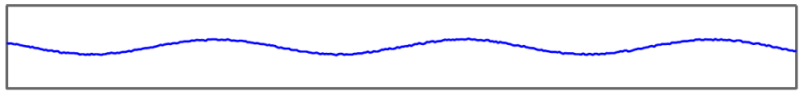
original signal



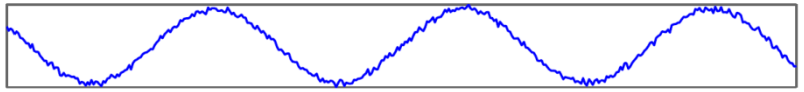
attenuate



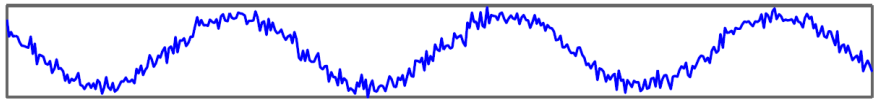
add noise



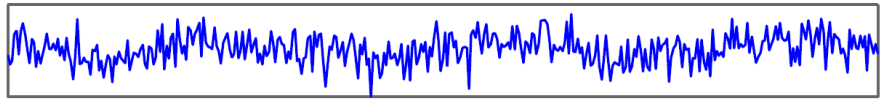
boost



Repeat process 5 times



Repeat process 100 times



Information theory

- Digital communications involves the transfer of symbols
- drawn from a discrete alphabet
  - English letters
  - English words
  - Decimal digits
  - Binary digits
  - DNA sequences
  - Quantized analog signals

Encoding information

- Minimal “piece” of information is one bit
- A bit can take on two values: { 0, 1 }
- There are 2^n ways to arrange n bits
- Therefore the number of bits required to encode n different sequences is: ceiling(log2(n))

Example

- Transmit information about a poker hand

{ straight flush, four of a kind, full house, flush, straight, three of a kind, two pair, pair, high card }

- There are 9 “messages”
- Baseline message length: ceiling(log2(9)) = 4 bits

Binary code for poker hands

straight flush	0000
four of a kind	0001
full house	0010
flush	0100
straight	1000
three of a kind	0011
two pair	0101
pair	1001
high card	0111

Note: Some messages (e.g. 0110, 1010…) are unused; suggesting that there is waste in this encoding

Prefix encoding

- Probabilities can be used to reduce the expected message length

straight flush	0.0000154	000011
four of a kind	0.000240	0000100
full house	0.00144	0000101
flush	0.00196	00000
straight	0.00393	0001
three of a kind	0.0211	010
two pair	0.0475	011
pair	0.422	001
high card	0.501	1

- More likely messages are encoded with smaller bits

Information and probability

- The information encoded is the identity of the poker hand
- The length of the message ought to be related to its information content
- A message that the opponent only has a pair or high card seems less informative than a message that they have four of a kind
  - because it happens more often
- Transmitting rare messages is more informative than transmitting common ones

Entropy

- For the information content a message or a whole system, which is called its entropy, we sum over all possible messages or states

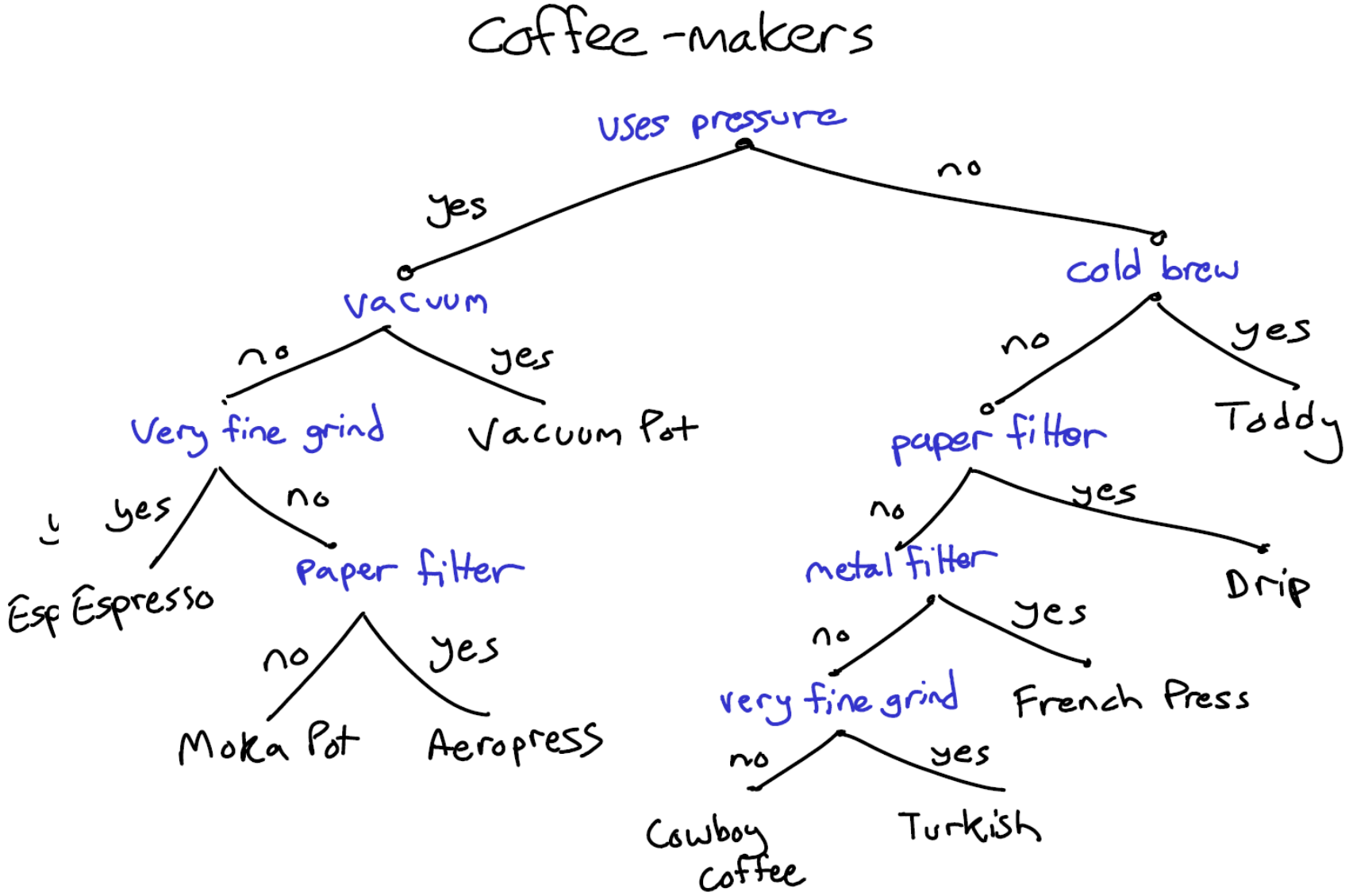
Entropy : H(X)
The measure of uncertainty in a system

Information Theory

- Joint Entropy: H(Y, X)
- Conditional Entropy: H(Y | X)
- Mutual Information: H(X) - H(Y | X)
  - the expected reduction in entropy due to knowing something

Decision Tree Classifier

- Build a tree where each node represents a test
  - Decision tree: leaf nodes assign labels
  - Regression tree: leaf nodes assign real values
- Decide quality measure for choosing branching features
- Building the tree is expensive, but testing is fast


Coffee-makers
Uses pressure
yes
no
Vacuum
cold brew
no
yes
no
yes
Very fine grind
Vacuum Pot
paper filter
Toddy
yes
no
no
yes
Esp
Espresso
Paper filter
metal filter
Drip
no
yes
no
yes
Moka Pot
Aeropress
very fine grind
French Press
no
yes
Cowboy coffee
Turkish

Building the tree

- Choose feature that is most discriminative across the training set
  - Mutual information is commonly used
- Split the training data according to this feature
- Repeat for each subset of data
- Stop at some threshold

Next time

- Dynamic Programming