More probability

CS151
David Kauchak
Fall 2010

Some material borrowed from:
Sara Owsley Sood and others

Admin

- Assignment 4 out
  - Work in partners
  - Part 1 due by Thursday at the beginning of class
- Midterm exam time
  - Review next Tuesday

Joint distributions

For an expression with *n* boolean variables e.g. $P(X_1, X_2, \ldots, X_n)$ how many entries will be in the probability table?

- 2^n

Does this always have to be the case?

Independence

Two variables are independent if one has nothing whatever to do with the other

For two independent variables, knowing the value of one does not change the probability distribution of the other variable (or the probability of any individual event)

- the result of the toss of a coin is independent of a roll of a dice
- price of tea in England is independent of the whether or not you pass AI

Independent or Dependent?

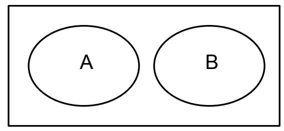
Catching a cold and having cat-allergy

Miles per gallon and driving habits

Height and longevity of life

Independent variables

How does independence affect our probability equations/properties?



If A and B are independent (written …)

- P(A,B) = P(A)P(B)
- P(A|B) = P(A)
- P(B|A) = P(B)

Independent variables

If A and B are independent

- P(A,B) = P(A)P(B)
- P(A|B) = P(A)
- P(B|A) = P(B)

Reduces the storage requirement for the distributions

Conditional Independence

Dependent events can become independent given certain other events

Examples,

- height and length of life
- “correlation” studies
  - size of your lawn and length of life

If A, B are conditionally independent of C

- P(A,B|C) = P(A|C)P(B|C)
- P(A|B,C) = P(A|C)
- P(B|A,C) = P(B|C)
- but P(A,B) ≠ P(A)P(B)

Cavities

$$P(W,CY,T,CH) = P(W)P(CY)P(T \mid CY)P(CH \mid CY)$$

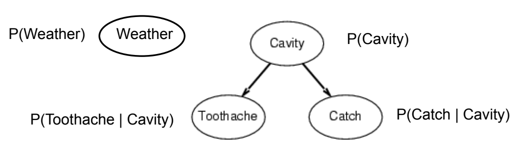
What independences are encoded (both unconditional and conditional)?

Bayes nets

Bayes nets are a way of representing joint distributions

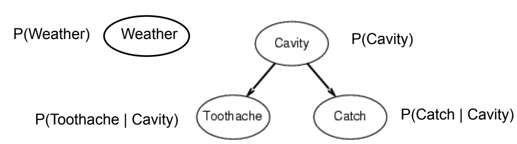
- Directed, acyclic graphs
- Nodes represent random variables
- Directed edges represent dependence
- Associated with each node is a conditional probability distribution
  - P(X | parents(X))
- They encode dependences/independences

Cavities



What independences are encoded?

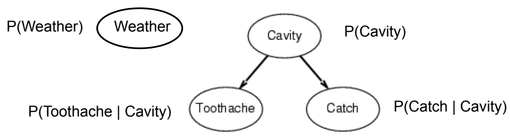
Cavities



Weather is independent of all variables
Toothache and Catch are conditionally independent GIVEN Cavity

Does this help us in storing the distribution?

Why all the fuss about independences?



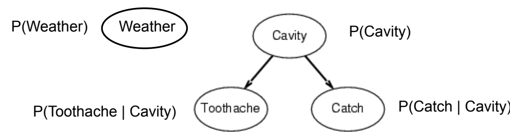
Basic joint distribution

- $2^4 = 16$ entries

With independences?

- 2 + 2 + 4 + 4 = 12
- If we're sneaky: 1 + 1 + 2 + 2 = 6
- Can be much more significant as number of variables increases!

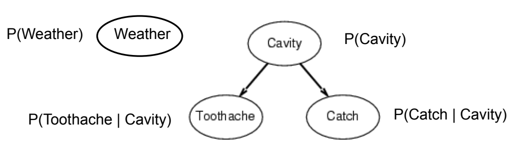
Cavities



$$P(W,T,CY,CH) = P(W)P(T,CY,CH \mid W)$$
$$= P(W)P(CY \mid W)P(T,CH \mid CY,W)$$
$$= P(W)P(CY \mid W)P(CH \mid CY,W)P(T \mid CH,CY,W)$$

Independences?

Cavities



$$= P(W)P(CY)P(CH \mid CY)P(T \mid CY)$$

Graph allows us to figure out dependencies, and encodes the same information as the joint distribution.

Another Example

Question: Is the family next door out?

Variables that give information about this question:

- DO: is the dog outside?
- FO: is the family out (away from home)?
- LO: are the lights on?
- BP: does the dog have a bowel problem?
- HB: can you hear the dog bark?

Exploit Conditional Independence

Which variables are directly dependent?

Variables that give information about this question:
- DO: is the dog outside?
- FO: is the family out (away from home)?
- LO: are the lights on?
- BP: does the dog have a bowel problem?
- HB: can you hear the dog bark?

Are LO and DO independent?
What if you know that the family is away?

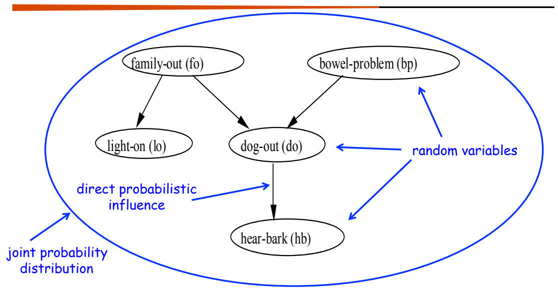
Are HB and FO independent?
What if you know that the dog is outside?

Some options

- lights (LO) depends on family out (FO)
- dog out (DO) depends on family out (FO)
- barking (HB) depends on dog out (DO)
- dog out (DO) depends on bowels (BP)

What would the network look like?

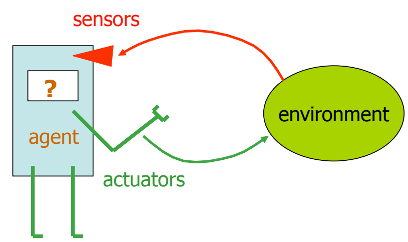
Bayesian Network Example



Graph structure represents direct influences between variables
(Can think of it as causality—but it doesn't have to be)

Learning from Data

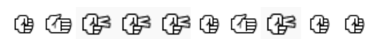
Learning



As an agent interacts with the world, it should learn about it's environment

We've already seen one example…

$$\frac{\text{Number of times an event occurs in the data}}{\text{Total number of times experiment was run (total number of data collected)}}$$

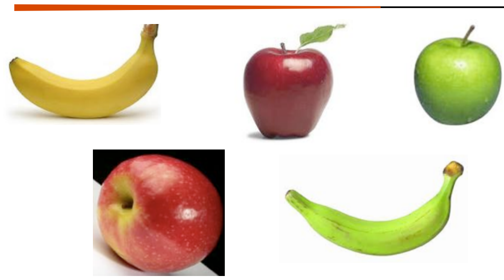


P(rock) = 4/10 = 0.4

P(rock | scissors) = 2/4 = 0.5

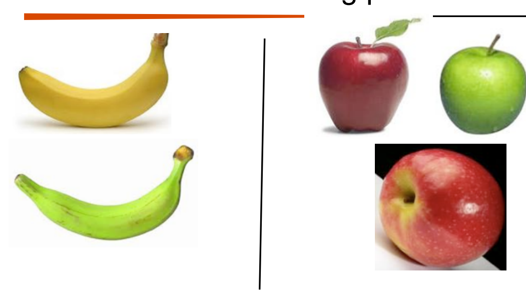
P(rock | scissors, scissors, scissors) = 1/1 = 1.0

Lots of different learning problems



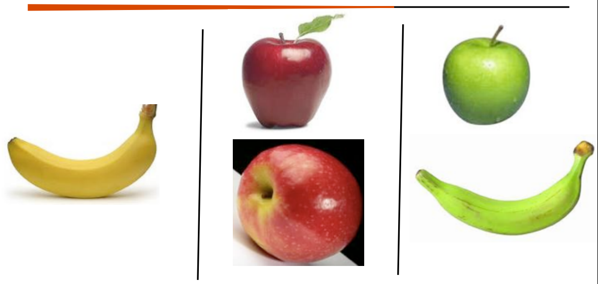
Unsupervised learning: put these into groups

Lots of different learning problems



Unsupervised learning: put these into groups

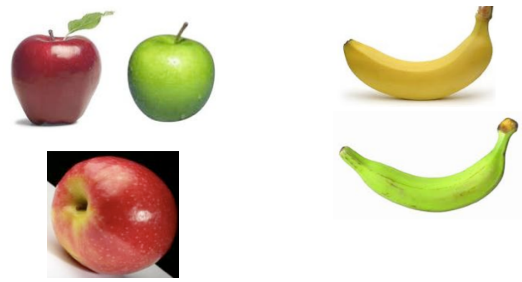
Lots of different learning problems



No explicit labels/categories specified

Unsupervised learning: put these into groups

Lots of learning problems



APPLES **BANANAS**

Supervised learning: given labeled data

Lots of learning problems

Given labeled examples, learn to label unlabeled examples

APPLE or BANANA?

Supervised learning: learn to classify unlabeled

Lots of learning problems

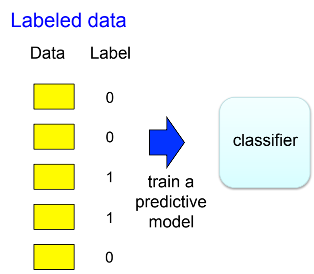
Many others

- semi-supervised learning: some labeled data and some unlabeled data
- active learning: unlabeled data, but we can pick some examples to be labeled
- reinforcement learning: maximize a *cumulative* reward. Learn to drive a car, reward = not crashing

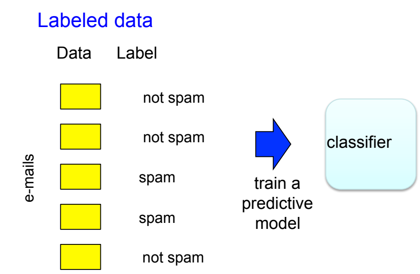
and variations

- online vs. offline learning: do we have access to all of the data or do we have to learn as we go
- classification vs. regression: are we predicting between a finite set or are we predicting a score/value

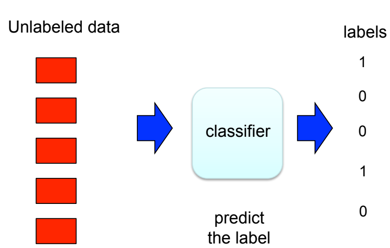
Supervised learning: training



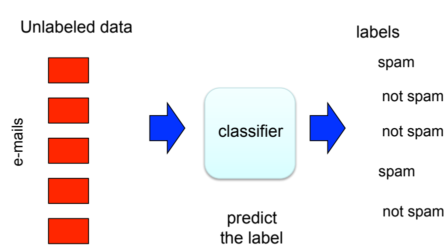
Training



Supervised learning: testing/classifying



testing/classifying



Some examples

image classification
- does the image contain a person? apple? banana?

text classification
- is this a good/bad review?
- is this article about sports or politics?
- is this e-mail spam?

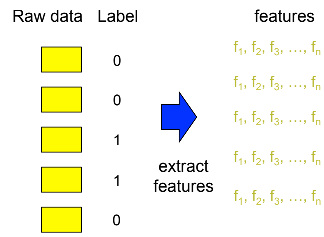
character recognition
- is this set of scribbles an 'a', 'b', 'c', …

credit card transactions
- fraud or not?

audio classification
- hit or not?
- jazz, pop, blues, rap, …

Tons of problems!!!

Features

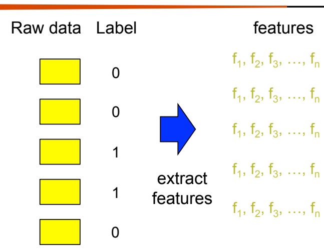


We're given "raw data", e.g. text documents, images, audio, …

Need to extract "features" from these (or to think of it another way, we somehow need to represent these things)

What might be features for: text, images, audio?

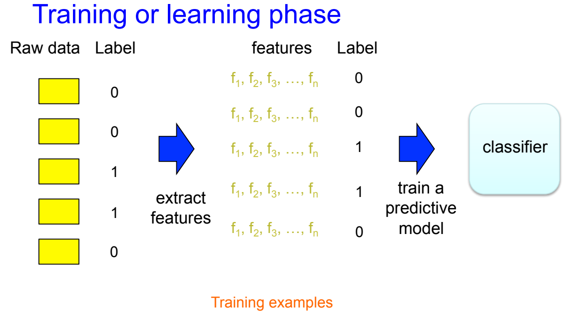
Examples



Terminology: An ***example*** is a particular instantiation of the features (generally derived from the raw data). A **labeled example**, has an associated label while an **unlabeled example** does not.

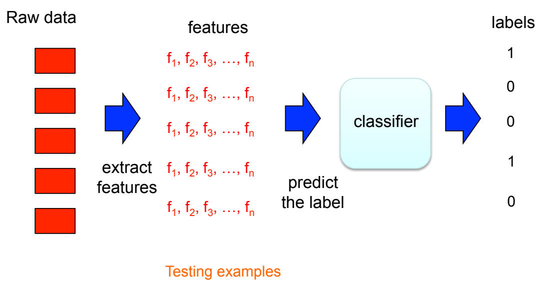
Feature based classification

Training or learning phase



Feature based classification

Testing or classification phase



Bayesian Classification

We represent a data item based on the features:

$$D = \langle f_1, f_2, \ldots, f_n \rangle$$

Training

a: $p(a \mid D) = p(a \mid f_1, f_2, \ldots, f_n)$

b: $p(b \mid D) = p(b \mid f_1, f_2, \ldots, f_n)$

$$p(Label \mid f_1, f_2, \ldots, f_n)$$

For each label/class, **learn** a probability distribution based on the features

Bayesian Classification

We represent a data item based on the features:

$$D = \langle f_1, f_2, \ldots, f_n \rangle$$

Classifying

$$p(Label \mid f_1, f_2, \ldots, f_n)$$

How do we use this to classify a new example?

For each label/class, **learn** a probability distribution based on the features

Bayesian Classification

We represent a data item based on the features:

$$D = \langle f_1, f_2, \ldots, f_n \rangle$$

Classifying

$$label = \underset{l \in Labels}{\operatorname{argmax}} P(l \mid f_1, f_2, \ldots, f_n)$$

Given an *new* example, classify it as the label with the largest conditional probability

Bayesian Classification

We represent a data item based on the features:

$$D = \langle f_1, f_2, \ldots, f_n \rangle$$

Classifying

$$p(Label \mid f_1, f_2, \ldots, f_n)$$

How do we use this to classify a new example?

For each label/class, **learn** a probability distribution based on the features

Training a Bayesian Classifier

features	Label
$f_1, f_2, f_3, \ldots, f_n$	0
$f_1, f_2, f_3, \ldots, f_n$	0
$f_1, f_2, f_3, \ldots, f_n$	1
$f_1, f_2, f_3, \ldots, f_n$	1
$f_1, f_2, f_3, \ldots, f_n$	0

train a predictive model

classifier

$$p(Label \mid f_1, f_2, \ldots, f_n)$$

How are we going to learn these?

Bayes rule for classification

$$P(Label \mid Features) = \frac{\text{number with features with label}}{\text{total number of items with features}}$$

Is this ok?

Bayes rule for classification

$$P(Label \mid Features) = \frac{\text{number with features with label}}{\text{total number of items with features}}$$

Very sparse! Likely won't have many with particular set of features.

Training a Bayesian Classifier

$p(Label \mid f_1, f_2, \ldots, f_n)$

Bayes rule for classification

conditional (posterior) probability

prior probability

$$P(Label \mid Features) = \frac{P(F \mid L)P(L)}{P(F)}$$

Bayes rule for classification

$$P(Label \mid Features) = \frac{P(F \mid L)P(L)}{P(F)}$$

$p(f_1, f_2, \ldots, f_n \mid Label)$

$p(Label)$

How are we going to learn these?

Bayes rule for classification

$$p(f_1, f_2, \ldots, f_n \mid Label) = \frac{\text{number with features with label}}{\text{total number of items with label}}$$

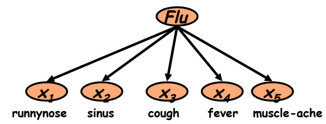
Is this ok?

Bayes rule for classification

$$p(f_1, f_2, \ldots, f_n \mid Label) = \frac{\text{number with features with label}}{\text{total number of items with label}}$$

Better (at least denominator won't be sparse), but still unlikely to see any given feature combination.

The Naive Bayes Classifier



Conditional Independence Assumption: features are independent of eachother given the class:

$$P(f_1, \ldots, f_n \mid l) = P(f_1 \mid l) P(f_2 \mid l) \cdots P(f_n \mid l)$$

$$label = \underset{l \in Labels}{\operatorname{argmax}} P(f_1 \mid l) P(f_2 \mid l) \ldots p(f_n \mid l) P(l)$$

Estimating parameters

$$label = \underset{l \in Labels}{\operatorname{argmax}} P(f_1 \mid l) P(f_2 \mid l) \ldots p(f_n \mid l) P(l)$$

How do we estimate these?

Maximum likelihood estimates

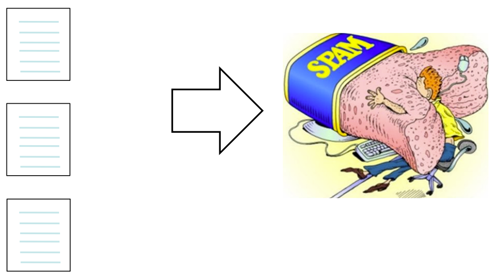
$$\hat{P}(l) = \frac{N(l)}{N} \qquad \frac{\text{number of items with label}}{\text{total number of items}}$$

$$\hat{P}(f_i \mid l) = \frac{N(f_i, l)}{N(l)} \qquad \frac{\text{number of items with the label with feature}}{\text{number of items with label}}$$

Any problems with this approach?

Naïve Bayes Text Classification

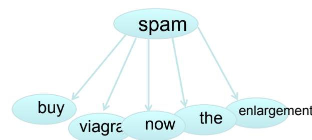
How can we classify text using a Naïve Bayes classifier?



Naïve Bayes Text Classification

Features: word occurring in a document (though others could be used…)

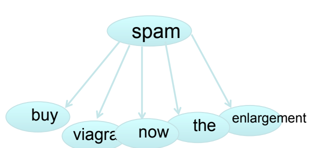
$$label = \underset{l \in Labels}{\operatorname{argmax}} P(word_1 \mid l) P(word_2 \mid l) \ldots p(word_n \mid l) P(l)$$



Naïve Bayes Text Classification

Does the Naïve Bayes assumption hold?

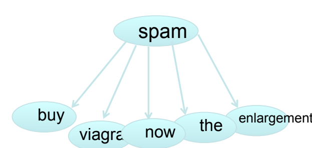
Are word occurrences independent given the label?



Naïve Bayes Text Classification

We'll look at a few application for this homework

- sentiment analysis: positive vs. negative reviews
- category classifiction



Text classification: training?

$$label = \underset{l \in Labels}{\operatorname{argmax}} P(word_1 \mid l) P(word_2 \mid l) \ldots p(word_n \mid l) P(l)$$

$$\hat{P}(f_i \mid l) = \frac{N(f_i, l)}{N(l)}$$

number of times the word occurred in documents in that class

number of items in text class