CSE 150 Summer 2016 Zhen Zhai

Discussion Wed / Fri 11am CSE 2154
Tutoring Hour Thu 2:30pm Mon 11am CSE 4262
Office Hour Mon / Fri 9am CSE 3214

Homework : Due Tuesday in class
Due Friday 10 am

6-28 section I

Course Overview

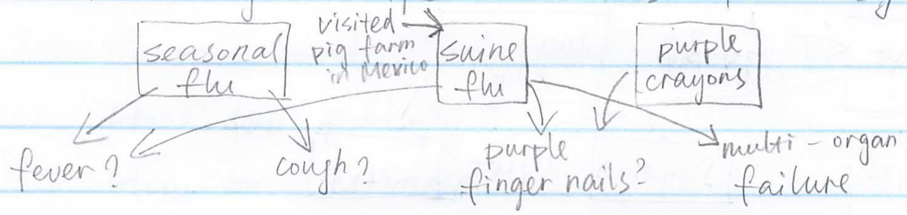
1. Probabilistic Reasoning

Ex: medical diagnosis

Knowledge representation : diseases cause symptoms

Modeling uncertainty : same diseases, same symtoms more likely than others.

Reasoning : infer diseases from symptoms.



Probability : quantitative, self-consistent framework that captures commonsense patterns at reasoning

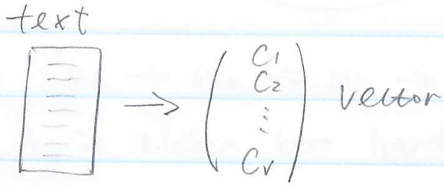
Graphical Model : How do graphs represent correlation causation, statistical dependence ?
Marriage of probability and graph theory.

2. Classification

Ex: spam filtering

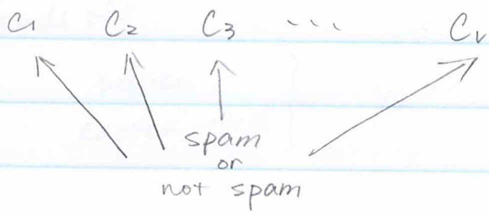
* input: email message
* output: {spam, not spam}
* How to represent input?

Convert text into fixed length vector of word counts.



V = vocabulary size (# of words in dictionary)

c_i = # times that i^{th} word appears.



Certain words are more likely in spam.

How to quantify?

How to estimate?

3. Sequential Modeling

How to model systems where "state" changes over time (or has some other extended representation)?

Ex: text (written language)

"states" — words

Which sentense is more likely?

① Mary had a little lamb.

② Colorless green ideas sleep furiously.

→ Markov Models for statistical language processing.

Let W_l = word at l^{th} position in sentence.

Graphical Model

Model A: $W_1 \to W_2 \to W_3 \to W_4 \to \cdots\ W_{L-1} \to W_L$

Model B: $W_1 \to W_2 \to W_3 \to W_4 \to \cdots \to W_{L-1} \to W_L$

Model A is richer but harder to estimate

Model B is impoverished but easier to estimate.
(obviously too simple)

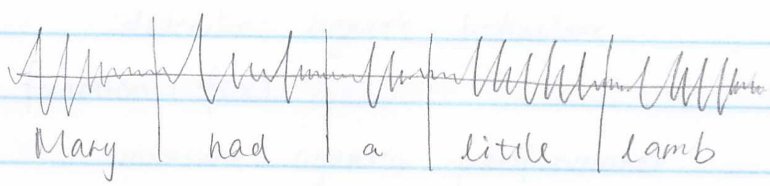
Trade off:

power, expressiveness vs. tractability, ease to estimate

Ex: speech (spoken language)

states = words (or syllables or smaller units of speech)

observations = sounds or waveforms.

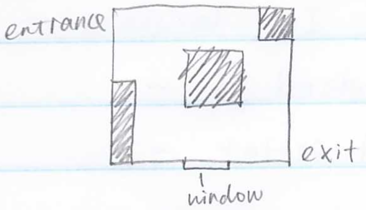


How to infer words from waveforms?

→ Hidden Markov Models for automatic speech recognition.

4. Planning & Decision-making

Ex: Robot navigation



* 2D grid world
* "states" = cells of 2D grid
* actions = west, north, south, east
* noisy dynamics
* rewards = feedback from environment
  - delayed vs. immediate
  - evaluated vs. instructive

More generally:

How can autonomous agents learn from experience?

→ Markov decision processes for reinforcement learning.

Other "embodied" agents: elevator, helicopters, ...

Other "embedded" agents: enemy AI, telephone operators, ...

5. Core ideas of modern AI

(1) Probablistic Modeling of uncertainty

(2) Learning as optimization

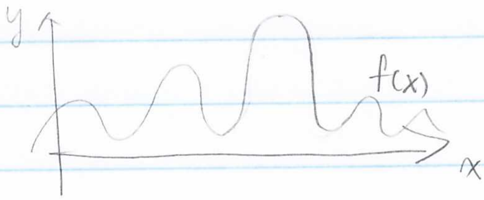
- parameter(s): x

describe agent's behaviors

- function: $f(x)$

measure agent's performance

How to optimize $f(x)$?



(3) Knowledge as predictions (dynamic)
not facts (static)

- classical AI
  - a book is on the table
  - tables have edges
  - ...
- modern agent-centric AI
  prediction: if action a, then consequence c
  observed with probability p.

Themes (broader) of class

(1) power vs. tractability
how to develop compact representations of complex worlds?

(2) principles vs. heuristic

optimization
probability vs. rules-of-thumb
calculus

6.28 section II

Motivation

* Modeling uncertainty
  (1) inherent randomness in world
  (2) gross statistical description of complex (deterministic) world
  (3) probability: guardian of commonsense reasoning

Review of probability

* Discrete random variable X (capitalized)
  Domain of possible values $\{x_1, x_2, \cdots, x_n\}$ (lowercase)
  Ex: month M, $\{m_1 = \text{Jan}, m_2 = \text{Feb}, \cdots m_{12} = \text{Dec}\}$
* "Unconditional" or "prior" probabilities $P(X = x)$
* Basic axioms
  (1) $P(X = x) \geq 0$ probability that event $X = x$ is true.
  (2) $\sum_{i=1}^{n} P(X = x_i) = 1$
  (3) $P(X = x_i \text{ or } X = x_j) = P(X = x_i) + P(X = x_j)$ if $x_i \neq x_j$
  Probabilities add for union of mutually exclusive events.
* "Conditional" or "Posterior" probabilities.
  $P(X = x_i \mid Y = y_j)$ prob that $X = x_i$ given $Y = y_i$
  In general: $P(X = x_i \mid Y = y_i) \neq P(X = x_i)$
* Dependent random variables
  Ex: W = weather $\{w_1 = \text{sunny}, w_2 = \text{rainy}\}$
  $P(W = \text{sunny}) = 0.9$
  $P(W = \text{sunny} \mid M = \text{Jan}\} = 0.83$
  $P(W = \text{sunny} \mid M = \text{Aug}\} = 0.97$

* Independent variables

Ex: D = day of week $\{d_1 = \text{Sunday}, d_2 = \text{Monday}, \cdots, d_7 = \text{Saturday}\}$

$P(W = \text{sunny} \mid D = \text{sunday}) = P(W = \text{sunny}) = 0.f$

$P(W = \text{sunny} \mid D = d) = P(W = \text{sunny})$

d: any day of week

* Conditionally independent variables

Ex: binary random variables

R = did Robert ace the test?

S = did Samantha ace the test?

T = was the test very easy?

$P(R=1) < P(R=1 \mid S=1)$

$P(R=1 \mid T=1) = P(R=1 \mid T=1, S=1)$

R and S are not independent, but they are conditionally independent given T.

* Conditionally dependent variables

Ex: binary variables

B = burglary?

E = earthquake?

A = alarm goes off?

$P(B=1) = P(B=1 \mid E=1) = P(B=1 \mid E=0)$

B and E are independent.

$P(B=1 \mid A=1) > P(B=1 \mid A=1, E=1)$

B and E are conditionally dependent given A.

* Same axioms hold for conditional probabilities

(1) $P(X = x_i \mid Y = y_i) \geq 0$

(2) $\sum_i P(X = x_i \mid Y = y_j) = 1$ ☆ sum over x not y

— sum over x, not y.

(2) $\sum_i P(X = x_i \mid Y = y_j) = 1$

Note $\sum_j P(X = x_i \mid Y = y_j) \neq 1$ in general, this sum can be anything from zero to # possible values of y.

Note: $P(X = x_i \mid \underline{Y = y_1, Y = y_2, \cdots})$

all these are conditions, no specific order

* "Joint" probabilities

$P(X = x_i, Y = y_j)$ = probability that $X = x_i$ AND $Y = y_j$

* Product rule — from conditional probabilities and prior probabilities to joint prob.

For all i, j: $P(X = x_i, Y = y_j) = P(X = x_i \mid Y = y_j) P(Y = y_j)$

Also: $P(X = x_i, Y = y_j) = P(Y = y_j \mid X = x_i) P(X = x_i)$

* Generalization:

$P(A = a_i, B = b_j, C = c_k, D = d_l, \cdots)$

$= P(A = a_i) P(B = b_j \mid A = a_i) P(C = c_k \mid A = a_i, B = b_j)$

$P(D = d_l \mid A = a_i, B = b_j, C = c_k) \cdots$

* It's easier to access conditional probs (RHS) than joint probs (LHS)

Ex: A = wake up on time

B = eat breakfast

C = hit traffic

D = arrive at UCSD on time

* Marginalization: from joint distribution to marginal distribution

$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$

$P(X = x_i, Y = y_j) = \sum_k P(X = x_i, Y = y_j, Z = z_k)$

Probabilities on LHS are called "marginal" probs over some subset of variables.

Also true:

$$P(X=x_i \mid Z=z_k) = \sum_j P(X=x_i, Y=y_j \mid Z=z_k)$$

* Shorthand Notation

(i) implied universality

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

implies that equality holds for all possible consistent assignments of $X=x_i$, $Y=y_j$

(ii) implied assignment

$$P(x,y,z) = P(X=x, Y=y, Z=z)$$

omit assignment when unambiguous.

$$P(a,b,c,d,\cdots) = P(a)P(b|a)P(c|a,b)P(d|a,b,c)\cdots$$

* Bayes Rule — relates conditional probs to other conditional probs.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

if you observe an effect Y, you can infer the likely cause X.

Ex: cancer diagnosis

Given: 1% population has cancer

Test has 10% false negative rate.

Test has 20% false positive rate.

Patient tests positive. Does patient has cancer?

* Random Variables

Diagnosis $\in$ {cancer, healthy}

Test $\in$ {pos, neg}

* Probabilities

P(cancer) = 0.01   P(healthy) = 1 − 0.01 = 0.99

P(pos | cancer) = 0.9   P(neg | cancer) = 0.1

P(pos | healthy) = 0.2   P(neg | healthy) = 0.8

Want to find:

$$P(\text{cancer} \mid \text{pos}) = \frac{\overset{0.9}{P(\text{pos} \mid \text{cancer})}\ \overset{0.01}{P(\text{cancer})}}{\underset{?}{P(\text{pos})}}$$

use marginalization to compute this

Marginalization:

$$P(\text{pos}) = P(\text{Test} = \text{positive}) = \sum_{d \in \{\text{cancer}, \text{healthy}\}} P(\text{Test} = \text{positive}, \text{Diagnosis} = d)$$

$$= \sum_d P(\text{Diagnosis} = d)\, P(\text{Test} = \text{pos} \mid \text{Diagnosis} = d)$$ Product Rule

$$= P(\text{cancer})\, P(\text{pos} \mid \text{cancer}) + P(\text{healthy})\, P(\text{pos} \mid \text{healthy})$$

$$= 0.01 \times 0.9 + 0.99 \times 0.2$$

$$= 0.207$$

$\therefore P(\text{pos}) = 0.207$

Bayes Rule: $P(\text{cancer} \mid \text{pos}) = \frac{0.9 \times 0.01}{0.207} = 0.043$

Before test: $P(\text{cancer}) = 1\%$

After test: $P(\text{cancer} \mid \text{pos}) = 4.3\%$

* Conditioning on background evidence

Often useful to reason in context of background knowledge

Consider events X and Y, and background evidence E

(i) conditionalized version of product rule

$$P(X, Y \mid E) = \frac{P(X, Y, E)}{P(E)}$$ product rule in reverse

$$= \frac{P(X, Y, E)}{P(Y, E)} \cdot \frac{P(Y, E)}{P(E)}$$

$$= P(X \mid Y, E) \cdot P(Y \mid E)$$ product rule in reverse.

Conditional: $P(X, Y \mid E) = P(Y \mid E)\, P(X \mid Y, E)$

Original: $P(X, Y) = P(Y)\, P(X \mid Y)$.

(ii) conditionalized version of Baye's rule

Ordinary: $P(X \mid Y) = \frac{P(Y \mid X)\, P(X)}{P(Y)}$

Conditional: $P(X \mid Y, E) = \frac{P(Y \mid X, E)\, P(X \mid E)}{P(Y, E)}$

* Conditional independence statements

The following three statements are equivalent.

(1) $P(x, y \mid E) = P(x \mid E)\, P(y \mid E)$

(2) $P(x \mid y, E) = P(x \mid E)$

(3) $P(y \mid x, E) = P(y \mid E)$

* How to measure the difference between two distributions over the same random variables?

Let $p_i = P(X = x_i \mid E)$ conditioned on different

$q_i = P(X = x_i \mid E')$ sets of evidence $E \neq E'$

Look at $\sum_i p_i \log(p_i / q_i) \leftarrow$ show that this measures "distance" between distribution